Chapter 35

Concept Inventories: Design, Application, Uses, Limitations & Next Steps

Michael W. Klymkowsky & Kathy Garvin-Doxas

Abstract: The idea of a concept inventory or concept test, as opposed to a conventional test, emerged during a period when education researchers and educators began to focus on identifying students' prior conceptions that can act as barriers to learning science. The typical concept test is multiple choice in format and focused on widely applicable concepts rather than facts. The incorrect choices, known as distractors, represent scientifically problematic ideas that students actually hold, presented in students' own (recognizable) language. Often multiple questions are used to target common or linked concepts. The validity of a concept test arises primarily from its distractor. In fact, distractors that do not represent or resonate with what students' actually think undermine the meaning of student scores. While improvements on concept test scores are often used as a proxy for overall "learning gains" within a course or a curriculum, this is problematic, particularly when the same test is used in a pre-/post model. A well-designed concept test can reveal the presence and/or persistence of conceptual confusions, but more accurate measures of learning are necessary to reveal what students can do with their knowledge. That said, evidence of conceptual confusion at the start (or the end) of a course can be critical when one considers the content presented and the types of formative learning activities that might help students develop an accurate working understanding of core disciplinary concepts.

There is little argument that effective teaching in the sciences is important as well as challenging. The development of modern scientific understanding is based on hundreds, and sometimes thousands of years of observation and experiment, leading to rigorous explanatory models underlying mechanisms involve. These models are then tested based on their explicit, often quantitative, predictions. In the experimental sciences, such tests involve the response of a system to various controlled perturbations. The results of this process has led to increasingly accurate, but often highly abstract and counter-intuitive scientific models, ranging from the bending of space and the atomic nature of matter to the physicochemical bases of life and consciousness and the underlying relatedness of all known organisms. While there are untold numbers of facts about the world, our understanding of how the systems that make up the world "works" is based on a relatively small number of general explanatory concepts.

In the context of science education, a key goal is to introduce students to the various observations upon which scientific concepts are based and how these concepts are applied to understand and explain various phenomena. As an example, the observed movements of physical objects (apples and planets) serve as the context that led to Newton's laws of gravitation and mechanics in general, and later, Einstein's theory of general relativity. Similarly, the diversity and structural similarities between organisms, together with cell theory, serve as the context within which modern molecular evolutionary theory provides mechanistic explanations.

A basic premise underlying instruction in the sciences has been that students who score well on summative evaluations understand the concepts on which the test questions are based. But there is a vast literature characterizing problematic ideas that persist even in "good" students

who have scored well on traditional tests. As noted by Jorion et al (2015), "Courses that cover topics broadly and focus on algorithmic understanding result in students who are able to pass exams yet still lack deep disciplinary understanding. Without such understanding, students have difficulty generalizing their knowledge beyond familiar contexts of textbook problems." Tests specifically targeting conceptual understanding have their origins in the "diagnostic tests" pioneered by David Treagust (1988; 1986). Such diagnostic tests are designed to determine whether students know the reasons why the correct answer is correct. This is a process with historic roots in Socratic interrogations and the methods used to established mastery through apprenticeship. Determining unambiguously whether students' understand and can appropriately apply core concepts within a discipline is difficult, but there are strategies available. While often considered impractical in the context of large classes, there are sampling strategies that could be applied if evaluating student learning outcomes is deemed sufficiently important. Concept tests are best at informing the instructor whether students retain or have developed non-scientific or incorrect preconceptions or working models that interfere with the accurate application of established concepts within a discipline.

The significance and usefulness of concept inventories was established further through observations made using the *Force Concept Inventory* (the "*FCI*")(Hestenes et al., 1992). The *FCI* was developed by David Hestenes and colleagues to reveal whether students retain their non-scientific understanding of how objects move and how forces act on them. While there have been questions raised as to what the *FCI* actually measures (Huffman and Heller, 1995; Hestenes and Halloun,1995), its impact is hard to minimize. The surprising result, based on *FCI* scores, was that many students with high scores on summative exams in physics classes retain their original non-scientific notions. In a particularly influential paper, cited over 5600 times, Hake (1998) used pre-/post-instruction results from the *FCI* to compare the outcomes of lecture and "interactive-engagement" forms of instruction. An interesting aspect of the Hake study was that the nature of "interactive engagement," was based on self-report and not unambiguously specified. While many presume that "interactive engagement" had positive effects on learning, a simpler alternative hypothesis is that lecturing, at least in the contexts studied, was toxic to learning, or at least learning as measured by the *FCI*. That said, how "lecturing" is characterized is itself generally imprecise and ambiguous (see Hora, 2014). Suffice it to say, such results stimulated many in the science education community to rethink their teaching strategies (Crouch and Mazur, 2001; Freeman et al., 2014; Klymkowsky et al., 2003; Mazur, 2009; Prince, 2004), as well as the strategies used to determine what students have learned (Cooper et al., 2010; Trujillo et al., 2012; Underwood et al., 2016).

Our focus here is on concept tests, including defining concepts versus facts, constructing a concept test, and their usefulness and limitations. Because our expertise is in the biological sciences, we will refer to concept tests available in other disciplines only in passing. We will argue that concept tests, at least as defined here, are not particularly useful in summative assessment, but there is strong evidence that results from such tests can help focus course and curriculum designers and instructors on problematic ideas and lead, eventually, to the development of more effective learning materials and improved course designs.

What is a concept and why is conceptual understanding important? Different disciplines are characterized by different types of concepts, but as a starting point, we might agree that a concept represents a generally applicable idea, rather than an idiosyncratic detail. In physics, it is

commonly presumed that there will ultimately be a "Theory of Everything," a single model that explains all physical phenomena (Hawking and Mlodinow, 2010). Chemistry is based on the behavior of atoms and molecules and their interactions with one another; interactions governed by electrostatic forces and associated energy changes. All electrons, protons, and neutrons, all carbon and hydrogen atoms and methane molecules are essentially identical to one another, even if they differ in terms of their environment and energy. Momentum, both linear and angular, is always conserved; chemical bonds always require energy to break and the formation of a bond always releases energy (Cooper and Klymkowsky, 2013; Einstein and Infeld, 1938). The result is that there are core concepts that explain the behaviors of subatomic particles, atoms, molecules, planets, stars, solar systems, and galaxies. In contrast, each biological organism is unique, an object shaped by its history, a history impacted by random processes and events, including the occurrence of mutations and various environmental factors (Mayr 1985; 1994). Biology deals with populations of individual cells and organisms, often interacting in ways shaped by adaptive and non-adaptive evolutionary mechanisms.

What concepts are relevant to biology? Biological systems are physical-chemical systems governed by physical and chemical principles. In that light, a number of physical and chemical processes are biologically relevant: concepts such as the behaviors of non-equilibrium systems, the laws of thermodynamics, the coupling of chemical reaction systems, and the factors that influence the strength and specificity of molecular interactions. At the molecular level, biological processes are based on stochastic (diffusive) processes and collisions between relatively small numbers of specific molecules and molecular targets. There are also overarching theories, such as the cell theory and the various selective processes that have shaped the evolutionary history and current state of an organism. The "central dogma" of molecular biology (Crick, 1970) suggests that information flows into the DNA of cells through the various processes of mutation and selection and flows out again through the regulated expression of gene transcription. The evolutionary source of genetic information implies that the phenotypic variation between organisms has its origins in mutations, that molecular processes are involved in the expression of genetic information, and that environmental factors can influence the effects of genetic variation. Because genes are linked together in various ways, effects arising from variation at one genetic locus can influence the effects and inheritability of others (Klymkowsky, 2010). As part of the process of our work on concept assessment, we assembled a Biological Concepts Library (link: http://bioliteracy.colorado.edu/conceptlists/Statements/Navigator.html) housed at the bioliteracy web site (http://bioliteracy.colorado.edu).

So what other biological concepts can we identify? A look at the literature reveals a number of attempts to articulate the concepts applicable to various aspects of biology, including Khodor et al's (2004) "A hierarchical biology concept framework," Scheiner & Willig's (2008) "A general theory of ecology," Zamer & Scheiner's (2014) "A conceptual framework for organismal biology," Michael et al's (2017) "The core concepts of physiology," the GSA website (www.genetics-gsa.org/education/GSAPREP_CoreConcepts_CoreCompetencies.shtml) in Genetics, and the National Research Council's (2012) "A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas" (Dimension 3: Life Science).

Can we be more specific about the concepts that apply to biological systems? Let us consider the details of Khodor et al's "hierarchical biology concept framework" and Zamer &

Scheiner's "conceptual framework for organismal biology" (Figure 35.1) to help clarify the distinction between a concept and a fact. There is no question that key facts (e.g., polynucleotide polymerization is directional) are useful, but they are not concepts; they are products of observation. Moreover, no student has a "common sense" understanding of such a process. What is important is the linearity of information coding, so that the directionality determines that the information can be "read" in an unambiguous manner.

Figure 35.1 Khodor et al (2004) and Zamer & Scheiner (2014) Conceptual Frameworks

Khodor et al. list eighteen "top-level" concepts and a great many sub-concepts. Twenty-four are listed under top-level concept 7: "DNA is the source of heritable information in a cell". In fact, many of the "concepts" listed by both Khodor et al and Zamar & Scheiner (Figure 35.1) are facts about biological systems. Similarly the Genetics Society site lists, "How is DNA organized" as a concept header, while many of the associated "learning objectives" are specific facts not generalizable concepts. What appears to be missing is the concept that information can be stored in the sequence of a DNA molecule; that this information is generated by evolutionary processes; that it is accessed through one or more sequence "addresses"; that these addresses are recognized and form semi-stable interactions with proteins through intermolecular interactions. While we could go on, the main point of this type of analysis is that universal concepts in biological systems can be buried under an avalanche of detail. The result is that the unifying concepts can go unappreciated and so fail to become integrated into a student's mental toolkit.

To construct a concept test we must first decide which concepts we intend to test. We must then determine what it is that students think about those concepts and how and when they are applied. To be useful, we want to focus on concepts for which a high percentage of students hold alternative understandings or do not recognize their relevance to a particular situation. How do we find those concepts? There are two approaches commonly used; one is based on instructors' or disciplinary experts' intuitions, while the other is based on student thinking. Adams and Wieman (2010) note that "there is a large body of literature on the use of student interviews for the purpose of understanding student thinking (Berardi-Coletta et al., 1995; Ericsson and Simon, 1998), but that student interviews are rarely used when developing educational tests" and that in their review of sixteen such instruments developed "in the past 20 years only Redish, Steinberg, and Saul (1998) and Singh & Rosengrant (2003) used interviews during both the development and validation process." More often, the instrument developer builds questions and their distractors based on their own views or the views of disciplinary experts, coupled with their intuitions about what students are thinking. It appears that was the approach used to generate the *FCI* (Hestenes et al., 1992), as well as a number of other concept tests. However, this approach may not capture what students actually think and understand about a particular concept.

Indeed, when Rebello & Zollman (2004) examined students' open-ended responses to *FCI* questions, their analyses revealed aspects of student thinking not found in the original *FCI* distractors. Based on student responses, these authors generated new distractors to replace the original ones, resulting in a decrease in the frequency of correct answers, suggesting that the original *FCI* distractors did not fully capture significant aspects of students' understanding. There are two lessons to be taken from their observations: 1) it can be difficult for disciplinary practitioners and instructors to accurately predict what students are thinking and 2) it is possible

that different populations of students may hold different ideas and apply their ideas differently in different contexts, which will influence overall concept inventory scores. Both can influence the accuracy of concept test results as measures of learning.

How do we build a concept inventory/test? As noted by Adams & Wieman, a more rigorous approach is to directly ask students what they think about a range of key disciplinary concepts. Each response is then analyzed for preconceptions, which may include misconceptions, incorrect or irrelevant assumptions, missing or mistaken facts, or inappropriately applied concepts. To facilitate our analysis of student thinking leading up to the construction (and testing) of the *Biology Concepts Instrument* (the "*BCI*"), we developed an on-line system, EdsTools (Garvin-Doxas et al., 2006; Klymkowsky et al., 2006); a computer-based system that enabled us to capture and encode student responses. As an example, consider students' responses to the question, "What is diffusion and why does it occur?" Almost universally, students described diffusion exclusively in terms of directed movements associated with concentration gradients and membranes; few noted the underlying role of thermal (random / Brownian) motion in driving molecular movements. Similarly questions focused on mutations revealed that students failed to consider the random nature of mutations and often presumed that mutations were driven to satisfy an organism's needs and/or adaptation to the environment.

Once a general map of student thinking has been developed, think-aloud interviews should then be done to clarify specific presumptions and shape the questions for the concept test. Such interviews ask students to articulate their entire process: what they are thinking as they read a question and consider the possible responses; what each potential response means to them and why the response they choose is correct or most closely captures what they understand to be the case. The questions generated through this process are then validated through student and expert interviews to determine whether students interpret them as intended by the instrument designers. The result is an iterative loop approach (Figure 35.2); this was our strategy for the construction of the *BCI* (Klymkowsky et al., 2010). The responses to such open ended questions enabled us to build a number of attractive distractors and revealed that students do not understand a key feature of molecular systems and the ubiquity of stochastic or random processes (Garvin-Doxas and Klymkowsky, 2008), among other things (Champagne-Queloz et al., 2016; Champagne-Queloz et al., 2017).

Figure 35.2 A flow chart for concept test construction, adapted from Klymkowsky & Garvin-Doxas, 2008

In sum, the process of building a concept test rests on the identification of concepts about which students hold non-scientific preconceptions, whether anchored in personal experiences with the world or arising from previous instruction. A relevant caveat is that the ideas that students apply when answering a particular question may not reflect the presence of a coherent explanatory model. The prior conceptions students bring with them are not necessarily coherent (Shapiro et al., 1987) and may be contradictory, a situation reflected in diSessa's "knowledge in pieces" model of student thinking and learning (diSessa, 1985; 2018). Furthermore, instruction that addresses a particular set of preconceptions in a particular context may suppress one set of preconceptions while leaving it and others intact in a different context. Moreover, choosing a

correct answer does not, in itself, guarantee that the student has abandoned the non-scientific preconception, the presence of which may influence their response in a different context.

Available Concept Tests Since the introduction of concept diagnostic tests, there has been a proliferation of such instruments (TABLE 35.1). One should note that few such tests have been subjected to the same level of scrutiny as the *Force Concept Inventory*.

Table 35.1 Available Biology Concept Tests

Testing Concept Tests. Part of the challenge in using concept tests, as with any scientific assay, is to determine what exactly the instrument measures and whether it is relevant to our specific research questions. When a student selects a distractor, as opposed to the correct choice, we can be fairly confident that either the choice was made at random (they guessed) or that they found the distractor more in concert with what they believe than the correct response. When we consider student responses, there is a third possibility, which is that their choice may reflect what they believe the instructor wants them to say, rather than what they actually understand or think. What we cannot say with any degree of certainty is whether they know why the correct response is correct. Recognizing this situation limits the use of concept tests to revealing the presence of persistent, widely-held but scientifically inaccurate ideas that are at least as compelling as the correct response. The fragmented nature of students' preconceptions plagues all concept tests and makes conclusions about what students know problematic. This is why an analysis of student responses to questions that address the same underlying concept often generate different responses. The building of a conceptually coherent understanding of a topic is difficult, and various preconceptions may well be left intact following instruction. Such question-context effects may also be responsible for gender bias in some *Force Concept Inventory* questions (Traxler et al., 2018). It should be noted that detailed analyses of concept test questions are rare and, to our knowledge, largely non-existent outside of physics. Although a small study, we think that Rebello & Zollman's results suggest that every concept test should be examined to determine whether the distractors used for a question reflect what students in a particular class express when they answer the question in a open response setting or whether other distractors might be more appropriate.

Validity, and to a lesser extent reliability, are often a source of controversy in the design of any instrument. Validity means that an instrument measures what it says it measures and nothing else. A thermometer should measure only temperature, not some combination of temperature and blood pressure. Reliability means that the instrument consistently measures the same thing. Validity implies reliability; what is less appreciated, and is often a source of confusion and controversy, is that reliability does not imply validity.

The controversy arises primarily because validity, unlike reliability cannot be demonstrated using standard statistical tests. A valid thermometer has the same statistical properties as a valid voltmeter, so we cannot tell what an instrument measures from its statistical properties alone; another input is needed. That input is always theory. In fact, it is typically a succession of theories, each theory independently tested within its own context that allows us to say that a particular instrument has validity.

For concept tests, the process of calibration is performed by experts who can consistently associate the verbal cues of students with the concepts that each student is using when answering

questions on a given subject. Crucially for concept test construction, the same process of calibration is used to develop both the distractors and the correct answers. Concept test results are therefore multidimensional, providing not only an estimate of a student's use of the correct conceptual framework but also an estimate of the degree to which a student holds and uses known alternative conceptions.

The degree to which the concepts and misconceptions covered by a concept inventory/instrument/test are correlated can be used to investigate whether the calibration of the instrument is consistent (e.g. Jorion et al., 2015), but the fundamental question of validity can only be answered by following a robust calibration process. For example, to develop the *Biology Concepts Instrument*, the process included: (a) initial text analyses of responses to open-ended questions looking for specific responses and patterns across questions; (b) student interviews and focus groups exploring results of text analyses followed by (c) think-aloud interviews; and finally, (d) associating the language captured in student responses to the conceptual frameworks they use in answering the questions (e.g. Garvin-Doxas et al., 2014; Garvin-Doxas & Klymkowsky, 2008; Klymkowsky & Garvin-Doxas, 2008). This iterative process sought to reconcile student think-aloud interviews with standard language for questions as well as the wording and content of each distractor and each correct response. By the final stages of the process, we had narrowed the wording of questions so they are most likely to elicit the existing understanding that each student holds for each fundamental idea; in the same manner, the wording of each distractor is narrowed such that they reflect the most commonly-held student misconceptions for each concept question. There are also secondary tests which, though they do not constitute proof, can provide additional evidence that a particular instrument includes effective distractors that represent widespread persistent misconceptions. In the *MOSART-LS* Project (see: https://www.cfa.harvard.edu/smgphp/mosart/), Sadler et al (per. comm.) undertook an Item Response Theory (IRT) analysis of a number of concept tests, which included statistics on the distractors.

Looking at the average ability of students who choose each option (correct or distractor), Sadler et al., (unpublished data) found that out of the thirty items in the *Biology Concepts Instrument*, 11 have the highest ability average corresponding to a distractor. This is the very essence of a misconception: a widely held model that even high ability students hold. Furthermore, of the remaining nineteen items, eleven have the second-highest ability corresponding to a misconception. So for twenty two out of thirty items, a known misconception corresponds to either the highest or second highest ability students. Students of middle ability are more likely to choose a response that represents a misconception than students of low ability. Far from being a problem exposing the deeply held misconceptions of higher ability students is the fundamental property of concept tests.

Although these results do not constitute proof, they support the claim that concept tests are different from conventional tests and that additional statistical metrics need to be applied. A readily accessible one, for example, consists of treating each concept (correct or not) as the "correct" answer and then re-interpreting, for each concept, student ability to mean the degree to which that student holds that particular concept (see Garvin-Doxas et al., 2014). While this type of analysis reintroduces "the curse of dimensionality," it is nevertheless important to face up to the fact that concept tests are indeed multi-dimensional instruments, and they need to be analyzed with different tools, as they obey different statistical rules, than tests.

Conclusions:

The value of any concept test relies on whether the distractors actually address presumptions made by students. We firmly believe in the value of concept tests to identify whether students hold and apply disciplinarily unreasonable ideas before or after instruction. This is information that can guide the design of course materials, particularly formative assessment activities that require students to work with ideas and help them build an appreciation for why non-scientific ideas do not "work" and need to be replaced by scientifically correct ideas. Because multiple-choice format tests are easy to use, there is a danger that concept test scores will be used as a proxy for a working understanding of underlying concepts. We disagree with other authors (Sands et al., 2018), including those who have used concept tests as evidence for improved learning (Freeman et al., 2014). The assumption that students deeply held presumptions can be easily "remediated" by simply overwriting a wrong idea with targeted instruction ignores the observation that common-sense, experience-based ideas are often remarkably resistant to instruction. Moreover, students can learn what answers are considered correct without understanding why they are correct. In this light, different types of assessments are necessary to evaluate students' working knowledge. In particular, there is a need to use more Socratic (interactive) assessments in which students are called upon to identify and justify their assumptions and to explain and articulate why a particular idea is relevant to the development of the solution to a specific problem.

Acknowledgements: We thank Melanie Cooper, Stacey Bretz, Carl Wieman, and Isidoros Doxas for critical insights, information on concept tests in chemistry and physics, and for pointing out the need for clarifications in an earlier draft. Work represented in this chapter was supported by a grant from the National Science Foundation and a Chancellor's award from the University of Colorado Boulder.

Literature cited.

Abraham, J. K., Perez, K. E. and Price, R. M. (2014). The Dominance Concept Inventory: a tool for assessing undergraduate student alternative conceptions about dominance in Mendelian and population genetics. CBE—Life Sciences Education 13, 349-358.

Adams, W. K. and Wieman, C. E. (2010). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. International Journal of Science Education iFirst, 1-24.

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

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L. and Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 205.

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., Moskalik, C. L. and Huether, C. A. (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. Genetics 178, 15-22.

Bretz, S. L. and Linenberger, K. J. (2012). Development of the enzyme–substrate interactions concept inventory. Biochemistry and Molecular Biology Education 40, 229-233.

Champagne-Queloz, A., Klymkowsky, M. W., Stern, E., Hafen, E. and Köhler, K. (2017). Diagnostic of students' misconceptions using the Biological Concepts Instrument (BCI): A method for conducting an educational needs assessment. PloS one 12, e0176906.

Champagne-Queloz, A., Köhler, K., Stern, E., Klymkowsky, M. W. and Hafen, E. (2016). Debunking Key and Lock Biology: Exploring the prevalence and persistence of students' misconceptions on the nature and flexibility of molecular interactions. Science Matters.

Cooper, M. M., Grove, N., Underwood, S. and Klymkowsky, M. W. (2010). Lost in Lewis Structures: an investigation of student difficulties in developing representational competence. J. Chem. Educ. 87, 869–874.

Cooper, M. M. and Klymkowsky, M. W. (2013). The trouble with chemical energy: why understanding bond energies requires an interdisciplinary systems approach. CBE Life Sci Educ 12, 306-312.

Costa, M. J., Howitt, S., Anderson, T., Hamilton, S. and Wright, T. (2008). A concept inventory for molecular life sciences: how will it help your teaching practice?

Couch, B. A., Wood, W. B. and Knight, J. K. (2015). The Molecular Biology Capstone Assessment: a concept assessment for upper-division molecular biology students. CBE-Life Sciences Education 14, ar10.

Crick, F. (1970). Central dogma of molecular biology. Nature 227, 561-563.

Crouch, C. H. and Mazur, E. (2001). Peer instruction: ten year of experience and results. Am. J. Phys. 69, 970-977.

Deane, T., Nomme, K., Jeffery, E., Pollock, C. and Birol, G. (2014). Development of the biological experimental design concept inventory (BEDCI). CBE—Life Sciences Education 13, 540-551.

Deng, Q., Ramsköld, D., Reinius, B. and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193-196.

diSessa, A. A. (1985). Knowledge in pieces. Berkeley: University of California.

---- (2018). A Friendly Introduction to "Knowledge in Pieces": Modeling Types of Knowledge and Their Roles in Learning. In Invited Lectures from the 13th International Congress on Mathematical Education, pp. 65-84: Springer.

Einstein, A. and Infeld, L. (1938). The evolution of physics. New York: Norton.

Embretson, S. E. and Reise, S. P. (2013). Item response theory: Psychology Press.

Ericsson, K. A. and Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. Mind, Culture, and Activity 5, 178-186.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H. and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 201319030.

Furtak, E. M., Morrison, D., Iverson, H., Ross, M. and Heredia, S. (2011). A conceptual analysis of the Conceptual Inventory of Natural Selection" Improving diagnostic utility through item analysis. In National Association of Research in Science Teaching. Orlando, FL.

Garvin-Doxas, K., Doxas, I. and Klymkowsky, M. W. (2006). Ed's Tools: A web-based software tool set for accelerated concept inventory construction. In Proceedings of the National STEM Assessment of Student Achievement conference. (ed D. Deeds). Washington DC, October 19-21, 2006.

Garvin-Doxas, K., Klymkowsky, M., Doxas, I. and Kintsch, W. (2014). Using Technology to Accelerate the Construction of Concept Inventories: Latent Semantic Analysis and the Biology Concept Inventory. In CSEDU 2014 - 6th International Conference on Computer Supported Education.

Garvin-Doxas, K. and Klymkowsky, M. W. (2008). Understanding Randomness and its impact on Student Learning: Lessons from the Biology Concept Inventory (BCI). Life Science Education 7, 227-233.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am. J. Physics 66, 64-74.

Hambleton, R. K. and Swaminathan, H. (2013). Item response theory: Principles and applications: Springer Science & Business Media.

Haslam, F. and Treagust, D. F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. Journal of Biological Education 21, 203-211.

Hawking, S. and Mlodinow, L. (2010). The (elusive) theory of everything. Scientific American 303, 68-71.

Hestenes, D. and Halloun, I. (1995). Interpreting the FCI. The Physics Teacher 33, 502-506.

Hestenes, D., Wells, M. and Swackhamer, G. (1992). Force concept inventory. The Physics Teacher 30, 141-166.

Hora, M. T. (2014). Limitations in experimental design mean that the jury is still out on lecturing. Proceedings of the National Academy of Sciences 111, E3024-E3024.

Huffman, D. and Heller, P. (1995). What does the force concept inventory actually measure? The Physics Teacher 33, 138-143.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V. and Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. Journal of Engineering Education 104, 454-496.

Kalas, P., O'Neill, A., Pollock, C. and Birol, G. (2013). Development of a meiosis concept inventory. CBE—Life Sciences Education 12, 655-664.

Khodor, J., Halme, D. G. and Walker, G. C. (2004). A hierarchical biology concept framework: a tool for course design. Cell Biol. Educ. 3, 111-121.

Klymkowsky, M. W. (2010). Thinking about the conceptual foundations of the biological sciences. CBE Life Science Education 9, 405-407.

Klymkowsky, M. W. and Garvin-Doxas, K. (2008). Recognizing Student Misconceptions through Ed's Tool and the Biology Concept Inventory. PLoS Biol 6, e3.

Klymkowsky, M. W., Garvin-Doxas, K. and Zeilik, M. (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. Cell Biol Educ 2, 155-161.

Klymkowsky, M. W., Gheen, R., Doxas, I. and Garvin-Doxas, K. (2006). Mapping student misconceptions using Ed's Tools, an on-line analysis system. Dev Biol 295, 349-350.

Klymkowsky, M. W., Underwood, S. M. and Garvin-Doxas, K. (2010). Biological Concepts Instrument (BCI): A diagnostic tool for revealing student thinking. In arXiv: Cornell University Library.

Knudson, D. (2006). Biomechanics concept inventory. Perceptual and motor skills 103, 81-82.

Knudson, D., Noffal, G., Bauer, J., McGinnis, P., Bird, M., Chow, J., Bahamonde, R., Blackwell, J., Strohmeyer, S. and Abendroth-Smith, J. (2003). Development and evaluation of a biomechanics concept inventory. Sports Biomech 2, 267-277.

Marbach-Ad, G., Briken, V., El-Sayed, N. M., Frauwirth, K., Fredericksen, B., Hutcheson, S., Gao, L.-Y., Joseph, S., Lee, V. T. and McIver, K. S. (2009). Assessing student understanding of host pathogen interactions using a concept inventory. Journal of Microbiology & Biology Education: JMBE 10, 43.

Mayr, E. (1985). The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Cambridge, MA: Belknap Press of Harvard University Press.

Mayr, E. (1994). Typological versus population thinking. Conceptual issues in evolutionary biology, 157-160.

Mazur, E. (2009). Farewell, Lecturer? Science 323, 50-51.

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., Modell, H. and Wright, A. (2017). Development and validation of the homeostasis concept inventory. CBE—Life Sciences Education 16, ar35.

Michael, J., Cliff, W., McFarland, J., Modell, H. and Wright, A. (2017). The core concepts of physiology: A new paradigm for teaching physiology. New York: Springer.

Newman, D. L., Snyder, C. W., Fisk, J. N. and Wright, L. K. (2016). Development of the central dogma concept inventory (CDCI) assessment tool. CBE—Life Sciences Education 15, ar9.

NRC (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC:: The National Academies Press,.

Odom, A. L. and Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. Journal of Research In Science Teaching 32, 45-61.

Paustian, T. D., Briggs, A. G., Brennan, R. E., Boury, N., Buchner, J., Harris, S., Horak, R. E., Hughes, L. E., Katz-Amburn, D. S. and Massimelli, M. J. (2017). Development, validation, and application of the microbiology concept inventory. Journal of microbiology & biology education 18.

Perez, K. E., Hiatt, A., Davis, G. K., Trujillo, C., French, D. P., Terry, M. and Price, R. M. (2013). The EvoDevoCI: a concept inventory for gauging students' understanding of evolutionary developmental biology. CBE—Life Sciences Education 12, 665-675.

Price, R. M., Andrews, T. C., McElhinny, T. L., Mead, L. S., Abraham, J. K., Thanukos, A. and Perez, K. E. (2014). The Genetic Drift Inventory: a tool for measuring what advanced undergraduates have mastered about genetic drift. CBE-Life Sciences Education 13, 65-75.

Prince, M. (2004). Does active learning work? A review of the research. Journal of engineering education 93, 223-231.

Rebello, N. S. and Zollman, D. A. (2004). The effect of distracters on student performance on the force concept inventory. American Journal of Physics 72, 116-125.

Redish, E. F., Saul, J. M. and Steinberg, R. N. (1998). Student expectations in introductory physics. American Journal of Physics 66, 212-224.

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

Sands, D., Parker, M., Hedgeland, H., Jordan, S. and Galloway, R. (2018). Using concept inventories to measure understanding. Higher Education Pedagogies 3, 60-69.

Scheiner, S. M. and Willig, M. R. (2008). A general theory of ecology. Theoretical Ecology 1, 21-28.

Seitz, H. M., Horak, R. E., Howard, M. W., Jones, L. W. K., Muth, T., Parker, C., Rediske, A. P. and Whitehurst, M. M. (2017). Development and validation of the microbiology for health sciences concept inventory. Journal of microbiology & biology education 18.

Shapiro, I., Whitney, C., Sadler, P. and Schneps, M. (1987). A Private Universe. Harvard-Smithsonian Center for Astrophysics, Science Education Department, Science Media Group.

Shi, J., Power, J. and Klymkowsky, M. W. (2011). Revealing student thinking about experimental design and the roles of control experiments. Int. J. Sci. Ed. 5, http://hdl.handle.net/10518/13647.

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q. and Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. CBE-Life Sciences Education 9, 453-461.

Singh, C. and Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. American Journal of Physics 71, 607-617.

Smith, M. K., Wood, W. B. and Knight, J. K. (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ 7, 422-430.

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., Zieffler, A., Brodfuehrer, P., Preest, M. and M. Belitsky, J. (2017). Development of a Biological Science Quantitative Reasoning Exam (BioSQuaRE). CBE—Life Sciences Education 16, ar66.

Traxler, A., Henderson, R., Stewart, J., Stewart, G., Papak, A. and Lindell, R. (2018). Gender fairness within the Force Concept Inventory. Physical Review Physics Education Research 14, 010103.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. Int. J. Sci. Educ. 10, 159-169.

Treagust, D. F., Smith, C. L. (1986). Secondary students understanding of the solar system: implication for curriculum revision. In GIREP conference 1986: Cosmos - an educational challenge. Proceedings of a conference held in Copenhagen, Denmark (ed. J. J. Hunt), pp. 363-368. Noordwijk, Netherlands: European Space Agency Publications Division.

Trujillo, C., Cooper, M. M. and Klymkowsky, M. W. (2012). Using graph-based assessments within socratic tutorials to reveal and refine students' analytical thinking about molecular networks. Biochem Mol Biol Educ 40, 100-107.

Underwood, S. M., Reyes-Gastelum, D. and Cooper, M. M. (2016). When do students recognize relationships between molecular structure and properties? A longitudinal comparison of the impact of traditional and transformed curricula. Chemistry Education Research and Practice 17, 365-380.

Wright, T. and Hamilton, S. (2008). Assessing student understanding in the molecular life sciences using a concept inventory. ATN Assessment 1.

Zamer, W. E. and Scheiner, S. M. (2014). A conceptual framework for organismal biology: Linking theories, models, and data. American Zoologist 54, 736-756.

from "A Hierarchical Biology Concept Framework" by Khodor et al., 2004.

Table 2. Concepts are organized into a nested hierarchy based on relative importance

7. DNA is the source of heritable information in a cell.
  7-1. The amino acid sequence of proteins is encoded in DNA.
    7-1-1. Sets of three letters in the nucleic acid alphabet (sets of four letters) specify one letter in the protein alphabet (sets of 20 letters).
      7-1-1-1. 64 triplet codons: ATG initiating methionine, three Stop codons, 60 other codons for the remaining 19 amino acids.
  7-2. Information is encoded in DNA, using different languages that are recognized by different machinery.
    7-2-1. DNA encodes when a gene will be expressed or not.
      7-2-1-1. DNA sequence: promoter, operator, enhancer.
      7-2-1-2. Protein machinery: activator, repressor, transcription factors.
    7-2-2. DNA encodes the point at which replication begins.
      7-2-2-1. DNA sequence: origin of replication (ori).
      7-2-2-2. Protein machinery: origin recognition complexes.
    7-2-3. t-RNA acts an adaptor to translate the nucleotide sequence into an amino acid sequence.
      7-2-3-1. The anticodon of a t-RNA is complementary and antiparallel to the codon it binds.
      7-2-3-2. Ribosomes are responsible for bringing the mRNA and t-RNA together and catalyzing the formation of peptide bonds.
    7-2-4. DNA encodes the information to properly segregate chromosomes during cell division.
      7-2-4-1. DNA sequence: centromere.
    7-2-5. DNA encodes the cellular address of each protein.
      7-2-5-1. DNA sequence encodes: nuclear localization signal, mitochondrial uptake sequence, signal sequence, and transmembrane domain.
      7-2-5-2. Protein machinery: receptors bind these amino acid sequences and localize proteins accordingly.
    7-2-6. DNA encodes: restriction endonucleases recognition sites.
  7-3. When DNA is mutated, the information it contains may be changed.
    7-3-1. Because DNA can encode amino acid sequences, the structure and therefore the function of proteins may be changed.
  7-4. Segments of DNA that contain all of the information to encode the sequence of a product and regulate its expression are called genes.
    7-4-1. The DNA that comprises an organism's genome is organized into chromosomes.

Attempts are made to use as little technical terminology and as much natural language as possible in the upper levels of the structure.

from "A Conceptual Framework for Organismal Biology" by Zamer & Scheiner, 2014

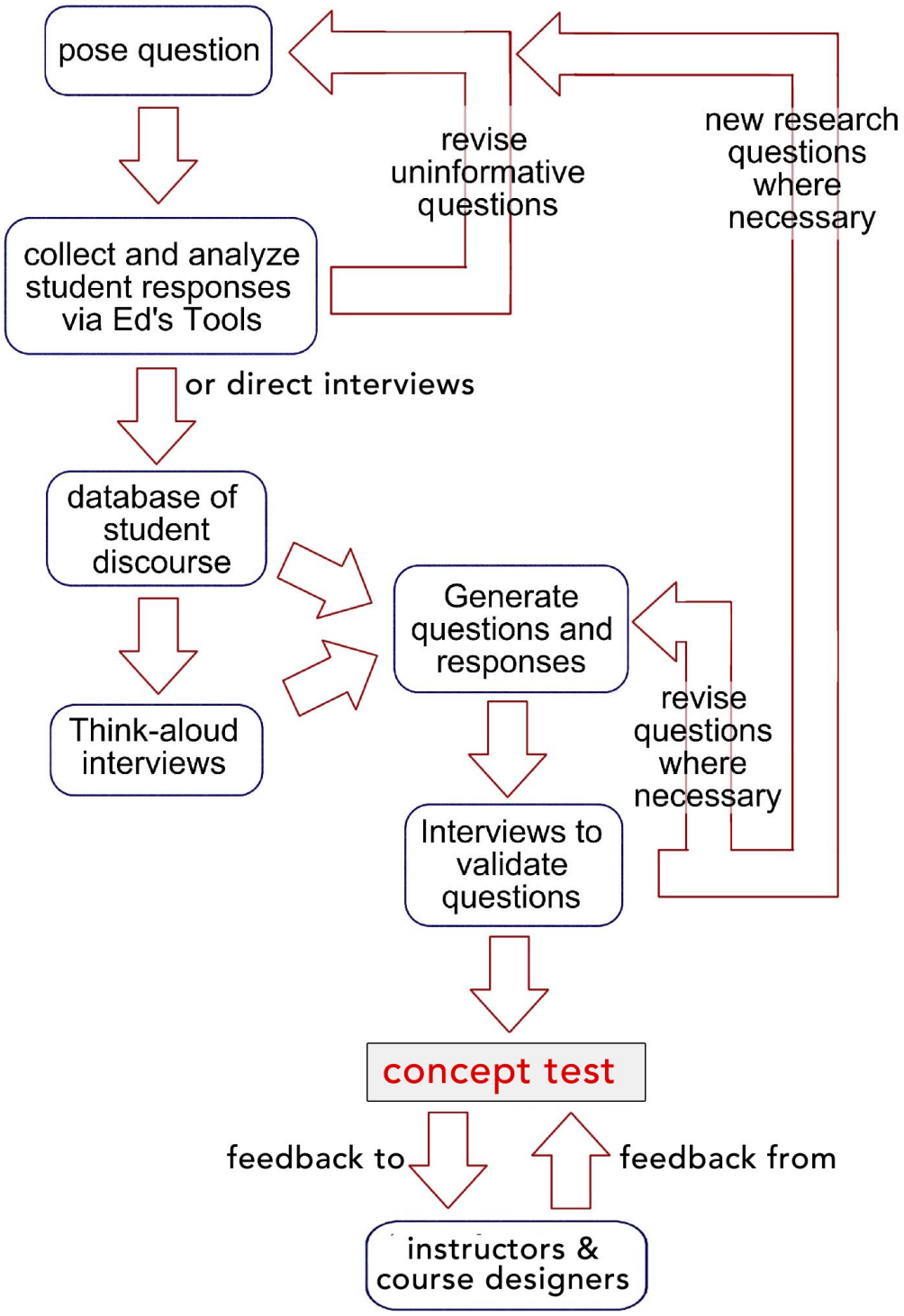
(A) Domain: Individuals and the causes of their structure, function, and variation. These principles apply to all organisms.

1. An individual organism actively maintains its structural and functional integrity.
2. All organisms are composed of cells at some point in their life cycle.[1]
3. Maintenance of organismal integrity requires dynamic change.
4. Organismal functions trade-off against each other.[2]
5. Maintenance of organismal integrity is a function of interactions with the abiotic and biotic environment.
6. Organisms require external sources of materials and energy for maintenance, growth, and reproduction.[3]
7. Because organisms are changeable, external factors can force change.
8. Environmental heterogeneity in space and time leads to variation in life-history patterns.
9. Organismal reproduction is both a cause and consequence of evolutionary processes.
10. The properties of organisms are the result of evolution.

(B) The fundamental principles of the sub-theory of multicellular organisms. These principles apply to multicellular organisms only

11. Multicellularity allows for specialization of cells.[4]
12. Cell–cell interactions are necessary for cell specialization
13. Specialization of cells requires their spatial or temporal localization at some point in the life cycle.
14. Specialization of cells allows for modularity.[4]
15. Specialization of cells leads to emergent organismal properties.
16. Development requires heterogeneity in cellular or organismal composition.

1. Not clear about the meaning of "at some point".
2. Is this universally true, is the "optimization" of a trait (function) always constrained by "trade-offs"?
3. Is this another way of saying that organisms are open, non-equilibrium systems?
4. Unicellular organisms can also display multiple cellular specializations, e.g. actively dividing versus spores, swimming versus non-swimming variants
5. How is "modularity" defined? Different function for different cell types?


pose question
revise uninformative questions
new research questions where necessary
collect and analyze student responses via Ed's Tools
or direct interviews
database of student discourse
Think-aloud interviews
Generate questions and responses
revise questions where necessary
Interviews to validate questions
concept test
feedback to
feedback from
instructors & course designers

concept test	reference
Photosynthesis & Respiration	Haslam & Treagust (1987)
Diffusion & Osmosis	Odom & Barrow (1995)
Concept Inventory of Natural Selection (CINS)	Anderson et al (2002) Furtak et al (2011)
Biomechanics concept inventory (BMCI)	Knudson et al., 2003
Biology Concepts Instrument (BCI)	Garvin-Doxas & Klymkowsky, 2008; Klymkowsky et al., 2010
Genetics Concept Assessment (GCA)	Smith et al., 2008
Genetics Literacy Assessment (GLA)	Bowling et al., 2008
Molecular Life Science Concept Inventory (MLSCI)	Costa et al., 2008; Wright & Hamilton, 2008
Host-pathogen Interaction Concept Inventory (HPICI)	Marbach-Ad et al., 2009
Introductory Molecular & Cell Biology Assessment (IMCBA)	Shi et al., 2010
Experimental Design & Control Assessment (EDCA)	Shi et al., 2011
Enzyme-Substrate Concept Inventory (ESCI)	Bretz and Linenberger, 2012
Meiosis Concept Inventory (MCI)	Kalas et al 2013
Evo-Devo Concept Inventory(EDCI)	Perez et al 2013
Genetic Dominance Concept Inventory (GDCI)	Abraham et al., 2014
Experimental Design Concept Inventory (EDCI)	Deane et al., 2014
Genetic Drift Inventory	Price et al., 2014
Statistical Reasoning in Biology Concept Inventory (SRBCI)	Deane et al., 2014
Molecular Biology Capstone Assessment (MBCA)	Couch *et al.*, 2015
Central Dogma Concept Inventory	Newman et al., 2016
Homeostasis Concept Inventory	McFarland et al., 2017
Microbiology Concept Inventory	Seitz et al., 2017
Microbiology-2 Concept Inventory	Paustian et al., 2017
Biological Science Quantitative Reasoning Exam	Stanhope et al., 2017
List of available Chemistry Concept tests: here List of available Physics Concept Tests: here	