Concept Inventories in Higher Education Science

Julie Libarkin

A manuscript prepared for the National Research Council Promising Practices in Undergraduate
STEM Education Workshop 2
Washington, D.C., Oct. 13-14, 2008

I. INTRODUCTION

Concept inventories (CIs) are multiple-choice assessment tests ideally designed for two learner-focused purposes. At their most useful, CIs can be used to diagnose areas of conceptual difficulty prior to instruction, and evaluate changes in conceptual understanding related to a specific intervention. Some CI developers (e.g Klymkowsky, and Garvin-Doxas. 2008) focus predominantly on diagnosis, while other efforts (e.g., Anderson et al., 2002, Libarkin and Anderson, 2007) work towards assessment tools that can serve the dual purposes of assessment as well as diagnosis. Regardless of the ultimate purpose of a CI, they are a valuable and necessary first-step in efforts to investigate learning in science fields across institutional settings.

CIs in higher education science are a burgeoning field for both practitioners and researchers. Generally acknowledged to be the first CI to emanate from a science domain, the Force Concept Inventory (Hestenes and Wells, 1992) provided the physics community with a snapshot of student learning in introductory courses. The investigation of college student conceptual understanding by members of the physics community in the mid-late 1980s was at least partially responsible for a focusing of the physics education community on conceptual change. Research on student conceptions in physics increased dramatically after 1985 (Libarkin and Kurdziel, 2001), and a wide array of innovations in physics instruction have subsequently utilized the FCI or newer CIs as independent methods of evaluation. For example, Workshop Physics, Physics by Design, and Lecture Tutorials in Physics have all been evaluated in part through FCI pre- and post-intervention testing.

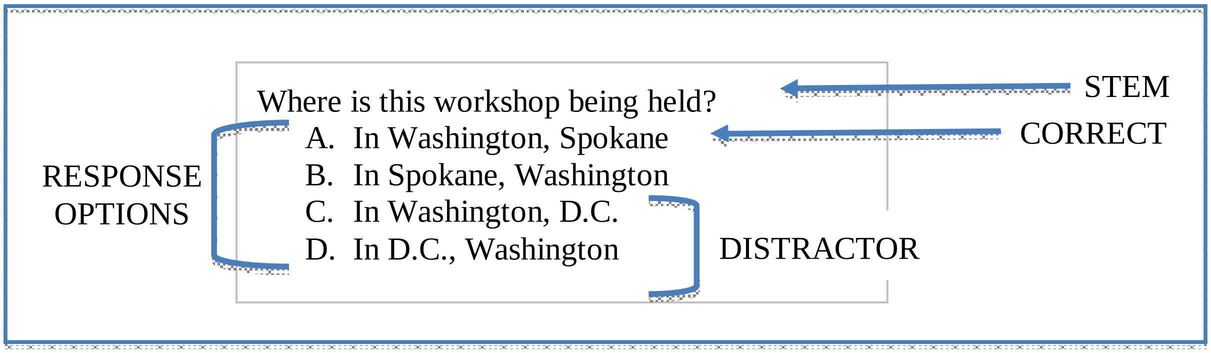
Where is this workshop being held?
A. In Washington, Spokane
B. In Spokane, Washington
C. In Washington, D.C.
D. In D.C., Washington
STEM
CORRECT
RESPONSE OPTIONS
DISTRACTOR

Figure 1. Components of a multiple-choice question.

The use of diagnostic multiple-choice items has a long history in science education (e.g., Treagust, 1986). As with any domain, test developers have a specific language for describing the components of a multiple-choice question. For clarity, Figure 1 has been provided to offer a schematic illustration of this terminology. The question, also called an item, consists of both a stem and response options. The stem refers to the statement that precedes the choices, or response options, in a multiple-choice question. Response options are further sub-divided in the correct response and the incorrect response options. The incorrect response options may be collectively called: distractors (preferred here), distracters, or foils. All CIs, with the exception of the Geoscience Concept Inventory (GCI), identified for this paper follow a common format for overall test construction. In general, science CIs contain between 20 and 35 items. These items are designed to broadly sample the content of an entire discipline (e.g., ADT in astronomy; Hufnagel, 2002), broadly sample a specific topic (e.g., genetics; Bowling et al., 2008), or more narrowly sample a sub-topic within a larger domain (e.g., lunar phases; Lindell, 2004).

Interestingly, use of CIs for diagnosis and assessment did not filter into wide use and acceptance by scientists engaged in higher education instruction until an instrument (FCI; Hestenes and Wells, 1992) originated from within a science domain. This single instrument and its impact on physics education research, initiated a cascade of CI development initiatives. The development of a valid and reliable assessment instrument is not an easy task, as has been noted across science and science education (e.g., Treagust, 1986; Libarkin and Anderson, 2007). Ultimately, any instrument, whether designed to measure physical variables (e.g., height or weight), affective variables (e.g., attitude, confidence), cognitive variables (e.g., conceptions, latent traits), or other dimension, must be an accurate measure of the trait being investigated. The quality of a research study depends fundamentally on the validity and reliability of the tools being used. With this is mind and focusing on CIs, it is important to consider current approaches to CI construction, appropriate psychometric standards, and overarching community needs and potential.

The majority of CIs in science were developed to evaluate conceptual understanding of novice college students. Typical targeted courses include entry-level courses for non-science majors, introductory level service courses enrolling new majors, majors in other sciences, pre-service teachers, or non-science majors, as well as pre- and in-service courses for elementary and secondary science teachers. In these contexts, CIs are often used to either provide insight into pre-instruction alternative conceptions, or to investigate the efficacy of nontraditional instructional interventions. CIs can be powerful assessment vehicles, particularly when coupled with other metrics such as interviews or observations (e.g., Elkins and Elkins, 2007).

A review of peer-reviewed literature, web-published CIs, and new NSF grants illustrates value science disciplines place on concept inventories as assessment tools (Tables 1 and 2). At least 23 distinct CIs have been or are being developed across the sciences, including CIs in astronomy, physics, geoscience, chemistry, and biology. With the exception of the GCI in geosciences, each CI is a stand-alone tool containing 17 to 43 items (Table 1). The GCI consists of a bank of interrelated items from which faculty can generate small sub-tests. Libarkin and Anderson (2007) recommend sub-tests of only 15 questions, to avoid subject fatigue. The predominance of independent and unrelated CIs generally means that different groups within (and between) disciplines develop, disseminate, and use these tools, often without dialogue between different instrument developers. The variability in content and psychometric expertise within development teams, the nature of the initial purpose for individual CI development, the implemented theoretical and empirical scale development perspectives, and the diversity of pilot populations targeted in initial CI construction leads to an astonishing diversity in development and validation approaches. Specific mechanisms inherent to development of valid, reliable, and conceptually significant CIs are discussed later in this paper, including recommendations for common standards and approaches.

Table 1. Comprehensive list of published concept inventories in science*

CONCEPT INVENTORY	**DESCRIPTION/AVAILABILITY**	**REFERENCES**
Physics		
Mechanics Baseline Test; Force Concept Inventory (FCI)	FCI most commonly used: 29 items. Modified versions available through the authors only	Hestenes et al., 1992 Hestenes and Wells, 1992
FMCE: The Force and Motion Conceptual Evaluation	47 items; See Workshop Physics site[1]	Thornton and Skoloff, 1998
Thermal Concept Evaluation	26 items; heat energy and temperature	Yeo and Zadnick, 2001
BEMA: Brief Electricity and Magnetism Assessment	Basic concepts in electricity and magnetism for calculus-based physics	Ding et al., 2006
CSEM: Conceptual Survey in Electricity and Magnetism	32 items; Basic concepts in electricity and magnetism	Maloney et al. 2001
Chemistry (Unclear if any published instruments exist)		
Astronomy (see http://astronomy101.jpl.nasa.gov/tips/index.cfm?TeachingID=32)		
Astronomy Diagnostic Test (ADT v.2.0)	21 items	Hufnagel, 2002
Lunar Phases Concept Inventory	20 items; http://www.camse.org:591/moon/	Lindell and Olsen, 2002
Light and Spectroscopy Concept Inventory	28 items	Bardar, 2005
Biology		
Inventory of Natural Selection	20 items	Anderson et al., 2002
Biology Concept Inventory	30 items; http://bioliteracy.net/	Klymkowski et al
Diagnostic Question Clusters: Biology	Unclear, no additional information available	Wilson et al., 2006
Host Pathogen Concept Inventory	Unclear, http://www.life.umd.edu/hpi/publication.html	Smith et al., 2007; Marbach et al., 2007
Genetics Concept Assessment	25 items	Smith et al., in press
Genetics Literacy Assessment Instrument	31 items	Bowling et al., 2008
Geoscience		

Geoscience Concept Inventory	68 currently validated items; sub-test generation encouraged. A community revision and expansion effort is underway.	Libarkin and Anderson, 2005;
		Libarkin and Anderson, 2006;
		Libarkin and Anderson, 2007

Other Geoscience-related tests	Referred to in publications	Black, 2005; Gosselin and Macklem-Hurst (2002)

*This list was generated through review of the literature, NSF award search, and use of internet search engines; CIs not identified via these searches may have been overlooked. Inclusion on this list does not imply endorsement of an instrument's validity and reliability by the author.

[1] http://physics.dickinson.edu/~wp_web/wp_resources/wp_assessment.html

Table 2. List of unpublished‡ or science concept inventories under development.

CONCEPT INVENTORY	**DESCRIPTION/AVAILABILITY**	**REFERENCES**
Biotechnology Concept Inventory	Under development	NSF grant: DUE-0837021; funded in 2008 to Siegel and colleagues
ECCE: The Electric Circuits Conceptual Evaluation	Discussed elsewhere, but specific reference not found	____
Chemical Concepts Inventory	The CCI is available at: http://jchemed.chem.wisc.edu/JCEWWW/Features/CQandChP/CQs/ConceptsInventory/CCIIntro.html	Mulford, 1996
Chemistry Concept Inventory (ChCI)	Foundation Coalition lists this tool, but but specific reference not found: http://www.foundationcoalition.org/home/keycomponents/concept/chemistry_desc.html	____
Organic Chemistry	https://engineering.purdue.edu/SCI/workshop lists this tool, but specific reference not found	____
GIS Concept Inventory	Under development	NSF grant: DUE-0837259; funded in 2008 to Bampton and colleagues
Chemical Equilibrium	Discussed in literature but availability is unclear.	Voska and Heikken, 2000
Other instruments originating from Astronomy Education	Several CIs are listed on NASA site. Unpublished include two: greenhouse effect (GECI), star properties (SPCI); see http://astronomy101.jpl.nasa.gov/tips/index.cfm?TeachingID=32	____

‡Unpublished refers to disseminated concept inventories for which peer-reviewed publications describing their development were unavailable.

II. IMPACT AND EFFICACY OF CONCEPT INVENTORIES IN SCIENCE

The development of CIs for use in college science classrooms can have significant impact on the way in which a community values and practices science education research and science instruction. In physics, the widespread use of the FCI coupled with the documentation of different learning outcomes for different instructional approaches (Hake, 1998) has led to significant discourse about "best practices" in physics instruction (see Mestre, 2008 for an interesting and related discussion of instructional goals). More recently, the development of the GCI (Libarkin and Anderson, 2007) for geosciences fortuitously occurred at the same time as a new community of geoscience education researchers was emerging. In addition to work by the GCI developers, several studies investigating conceptual change in geoscience and which utilize the GCI have been published. These studies have considered the value of fieldwork in conceptual change (Elkins and Elkins, 2007), impact of peer instruction on student learning (McConnell et al., 2005), adaptation of Lecture Tutorials to geoscience content (Kortz et al., 2008), and impact of pre-service teacher education on conceptual understanding (Petcovic and Ruhf, 2008). These studies utilized a number of different analytical approaches, taking advantage of the average pre/post design most commonly implemented in CI use, as well as more detailed analysis of individual item performance (e.g., Petcovic and Ruhf, 2008). Similar widespread dissemination and use of CIs in other sciences highlights the need for such tools. The next section discusses the importance of considering the approaches used in developing CIs, to ensure that an instrument is valid and reliable prior to its use.

III. EVALUATIONS OF CONCEPT INVENTORY DEVELOPMENT AND USE

The need for assessment tools for undergraduate science is felt most strongly by scientists engaged in undergraduate instruction, while the scholars most qualified to develop valid and reliable instruments are far removed from the pre-requisite science content understanding needed to develop a content-specific CI. For many CI development initiatives, this has led to the development of tools that are either not embedded in standards for instrument design, or which are not content appropriate for the undergraduate sciences. As noted by other scholars (Etkina, et al., 2005; Mestre, 2008) STEM professionals are trained in disciplinary content areas, not in assessment. Given this disconnect between those individuals most invested in CI development and the communities most qualified to develop valid and reliable tools, users of CIs must themselves gain understanding of instrument standards in order to ensure that only valid and reliable tools are being used for assessment purposes.

Users of CIs play an important role in development of valid and reliable instruments, and development teams often overlook the role of the user in ensuring that questions are both appropriate and well written. Users can consider the importance of item topics for the courses being taught, can gain an understanding of the importance of word choice in question development, and can easily provide expert insight as they engage in question review. In this paper, the discussion of validity and reliability is limited to three forms of validity (construct, content, and communication) that are absolutely necessary for development of effective measurement instruments. Other forms of validity can and should be considered in instrument design (e.g., Trochim, 2004).

In the context of CI development, my experiences suggest that construct, content, and communication validity should all play a central role in the conceptualization of instrument content. *Construct validity* is concerned with whether or not strong support for the content of the items exists. *Content validity* (also called face validity) considers whether or not the test items actually measure the latent trait being measured, and generally considers this from the perspective of the test developer. On the other hand, *communication validity* considers the test-taker perspective, asking whether or not the test-taker interprets the items in the same way as intended by the test developer (e.g., Lopez, 1996).

Experts, both in the content area and in test development, play an important role in establishing these three forms of validity. Rather than trying to keep these technical definitions of validity straight, we can rephrase the definitions above and ask ourselves three simple questions, phrased here with respect to geoscience:

1) Construct: Is the topic covered by this item important for geosciences understanding?
2) Content: From the perspective of an expert geoscientist, does the itemactually measure some aspect of geoscience understanding?
3) Communication: Would a test-taker interpret this item <u>in the same way</u> as intended by the test developer?

These questions can be answered through established research approaches in education, psychology, and related fields, and particularly through careful attention to established approaches for instrument design. A brief discussion of suggested "best practices" in CI development is provided below.

A fourth type of validity, ***cultural validity***, is worth mention here, especially as we move towards international collaborations in science education research. Cultural validity (Solano-Flores and Nelson-Barber, 2001) is an important and often ignored aspect of test development. The way in which different peoples will understand and interpret questions depends in part on the societal and geographical lens through which they view the world. Cultural validity comes into play both when considering exemplars of common objects and geographic perspective. For example, students from large urban settings may be unfamiliar with the concept of a "rowboat", and may have little experience with the Milky Way. In the same way, students in Australia may have little understanding of the shape of an American football, and will experience the seasons at different times than students in the Northern Hemisphere. The need to accommodate different cultural and geographical perspectives may require modification and re-validation of existing instruments, such as the on-going initiative to create a Southern Hemisphere Edition of the Astronomy Diagnostic Test (http://www.physics.usyd.edu.au/super/ADT.html).

Experience in CI development, use, and revision suggests that authentic CI items, those that are meaningful and relevant to the population being studied, must be created with careful attention to scale development rules, validation approaches, and student perspective. In order to write authentic concept questions, two important steps must be followed. First, rules for multiple-choice question development are well known within the scale development community, and should be utilized in order to devise questions that are as psychometrically valid as possible. Second, data collected from the target population must be considered in addressing the rule: "Use plausible distractors" (Figure 2). These data will provide a deeper understanding of existing alternative conceptions, which can then be developed into plausible distractors. The use of qualitative data to develop distractors for assessment tests has become a standard practice in concept inventory development. I also strongly encourage the use of qualitative data in determining which content areas should be covered on an assessment instrument and in developing question stems. As noted above, test-takers and test developers may view the content area differently, and most likely will start out with a very different foundation upon which they build their knowledge. Further discussion of how to gain an understanding of alternative conceptions from qualitative data is beyond the scope of this paper, although it is worth noting that the field of alternative conceptions research is both longstanding and currently vigorous across all of the sciences.

Rules related to writing STEMS

1. Structure the stem as a question when at all possible. Use: “What is obsidian?”, rather than the completion form of “Obsidian is______”. If you use a completion form, keep the blank at the end.

2. Use unambiguous and simply worded stems. Use as few sentences as possible. Do not use parenthetical statements or unnecessary commas.

3. Use appropriate vocabulary. Avoid technical language for non-majors, for example. Make sure that the vocabulary is understandable to the target population.

Rules related to writing RESPONSE OPTIONS

1. Use plausible response options. Make sure that the distractors are meaningful to the population being tested.

2. Use 3 to 5 response options. More than five options adds no psychometric value and may produce confusion for the test-taker.

3. Avoid TYPE K format questions. (TYPE K: A list of statements is provided, and responses are a combination of statement choices).

4. Avoid absolutes and complexity in response options. Do not use “All of the Above”, “None of the Above”, and complex response format (e.g., ”a and c”, but not b”).

5. Keep the lengths and structure of response options similar. The longest or shortest answer is often the correct response. (Anecdote: If you choose all of the longest answers on the Force Concept Inventory, then you will score at the national average). Similarly, the more “technical” answer is often the correct response.

Figure 2. Principle rules guiding construction of item stems and response options for multiple-choice questions. Rules have been collated from several sources (Haladyna and Downing, 1989a; Haladyna and Downing, 1989b; Frey et al., 2005).

The field of test development provides us with a number of “rules” for writing assessment questions. These rules significantly increase the likelihood that a question will have satisfactory construct, content, and communication validity. Eight of these rules appear to be fundamentally necessary for construction of valid items, as suggested by the existence of research supporting the necessity of the rule (e.g., Haladyna and Downing, 1989b; Frey et al., 2005; Figure 2). Certainly, exceptions to these rules can be found in almost any validated test, suggesting that these are guidelines rather than strict and inviolable laws. Consistency within the community, however, would promote following of these rules where possible, and documentation of reasons for rule violation where necessary.

A synthesis of my personal perspective on best practices in CI development follows. Other CI developers and scholars will certainly hold a different perspective, particularly as perspectives will be context-dependent and inherently tied to individual disciplinary background. I see these suggestions as a basis for meaningful discussion among CI developers about metrics for determining CI validity and reliability. A set of community standards would be extremely beneficial for reviewers and users of CIs who may be less familiar with scale development theory and practice, and who need the development community to take ownership of CI oversight. An example of the method currently being used to revise and expand the GCI is provided (Figure 3) and in comparison to other CIs (Table 3) as an exemplar of one way in which these considerations might be addressed.

1. The topic that will be covered by a CI or set of CI items should be carefully considered prior to initiation of development. This will ensure that the CI is targeting the concepts that are most important for the targeted population and setting.
2. As much as possible, test items should be embedded in the experience of the testing population. This means that distractors, and stems if appropriate, should originate from interaction with the target population. Several recent efforts in CI development generated distractors based upon both review of conceptions literature and the experience of the developers; I would personally argue that the perception of the developer may be far removed from the reality of actual ideas held by students.
3. Items should be developed based upon existing and research-based standards in item development. Some of these standards, such as the rules depicted in Figure 3, are universal. Other rules will depend upon the test construction theory driving the work. For example, identification of co-existing ideas about a single concept might be addressed through use of items with multiple response options (“choose all that apply” items). In classical test theory, items with multiple response options are generally discouraged as they are generally more difficult and troublesome to score. Partial credit item response theory models, on the other hand, can easily address scoring concerns related to this type of item.
4. Validity and reliability are the “lens” through which we should consider the usability of an instrument. Validity helps us reconcile a measurement value with the true value of the trait being measured. For example, we need to understand how a score on the GCI represents the level of geoscience understanding actually held by the individual student. Similarly, reliability addresses how well we are able to reproduce a measure or repeat a study. A wide array of validity and reliability approaches, utilizing both qualitative and quantitative data, should be considered when developing a research design prior to CI development.
5. Finally, and assuming that the previous steps involve a range of validation processes and measures of reliability, CIs are ultimately developed to meet the needs of the faculty and scholars within our communities. Community input on the content, construction, review, and dissemination of CIs will ultimately result in instruments that more effectively meet a diversity of needs.

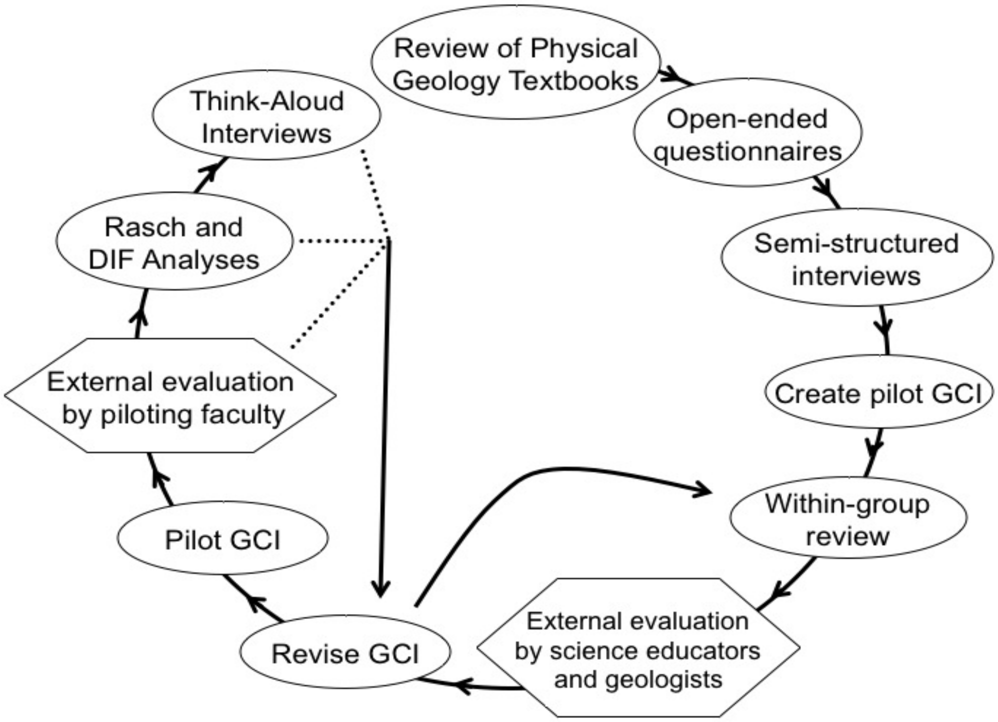
Review of Physical Geology Textbooks
Open-ended questionnaires
Semi-structured interviews
Create pilot GCI
Within-group review
External evaluation by science educators and geologists
Revise GCI
Pilot GCI
External evaluation by piloting faculty
Rasch and DIF Analyses
Think-Aloud Interviews

Figure 3. Schematic of ongoing development of the GCI

Table 3. Comparison of existing CI approaches to the Geoscience Concept Inventory (GCI).

CONCEPT INVENTORY DEVELOPMENT APPROACHES*	DEVELOPMENT OF THE GCI	COMMENTS ON GCI APPROACH
Predetermined content	Test content is based upon ideas presented by students	Questions are grounded in data gathered from college students
Alternative choices based on developer opinion, existing studies, questionnaires, and/or interviews (N=10-75)	1000+ questionnaires; 75+ interviews 10 institutions	Analysis (coding) of qualitative data allowed development of authentic “incorrect” choices
50-500 college students tested during piloting	Fall 2002: N = 2219 pre-tests F2003: N = 1500 pre-tests	For N >~300, statistical sampling of sub-populations is usually possible
Institutions of similar type or locality (N = 1-5)	Colleges: 5 community or tribal, 44 public or private, 60 courses, 8-250 students.	The GCI should be generalizeable to all populations of students.
Commonly, statistical analyses either not performed, or reporting of reliability statistics, difficulty scores, or linear bias scores only. Factor analysis performed in some cases.	Rasch analysis performed. Some items removed due to statistical bias as measured by Differential Item Function analysis.	Raw scores can be re-scaled relative to test difficulty, providing a more accurate measure of changes. Sub-tests are statistically comparable.

* Blend of development strategies utilized by CIs listed in Table 1.

IV. NEEDS and FUTURE DIRECTIONS

One limitation of existing CIs is the rigid character of the test content. That is, the developers predetermine the specific questions included on the inventory, and faculty or researchers interested in using the instrument are limited by the content of these questions. Most importantly, courses that do not cover the content included on these inventories are still limited to evaluation via these few existing instruments. A newly funded project to revise and expand the GCI offers a mechanism for developing flexibility in sub-test design while maintaining the ability to statistically compare results across content (e.g., Libarkin and Anderson, 2006; Table 3).

Mestre (2008) highlights a problem in STEM education reform that persists despite several decades of effort in CI use and development. As noted above, standard practice in CI development results in production of isolated CIs, often with specific relevance to a single course or sub-discipline. These CIs have no specific meaning relative to one another, inhibiting meaningful comparison across content. This results in understanding of student learning across very small time spans, from a few weeks during instruction to the more common semester long, pre/post evaluation. Rarely, students are given delayed CIs several months to a year post-instruction, providing some measure of short-term longitudinal effects. Investigation of conceptual change across a program is currently outside of the reach of any existing CI, although the ongoing effort to expand the GCI and apply item response theory techniques holds promise for integrated assessment within an undergraduate program of study.

At present, efforts to create concept inventories for assessment of learning in higher education science are highly concentrated within specific disciplines. The pioneering work of the Mechanics Baseline Test (Hestenes and Wells, 1992) and the FCI (Hestenes et al., 1992; Table 1) sparked the development of assessment instruments in a number of other science disciplines, including astronomy, biology, and the geosciences. Each of these efforts yielded assessment instruments that are being actively used by researchers within the discipline, and are yielding valuable information about the connection between teaching and learning in specific disciplines.

The wide variety of existing assessment tools across science disciplines (Table 1) highlights a significant question in higher education assessment, namely the relationship between scores on disparate assessment instruments. For example, the Force Concept Inventory and the Inventory of Natural Selection have both undergone validity and reliability study, and each is widely used by the physics and biology communities, respectively. However, learning gains measured by the FCI, and linked to specific teaching approaches such as collaborative learning (CL), are only applicable to physics instruction. Replicate studies of CL's impact on students' conceptions of natural selection are needed to determine if learning and CL approaches are linked in biology instruction. However, the question remains: How is a score of 50% on one CI related to a score of 50% on an independently developed second CI? Are learning gains or effect sizes measured by different inventories comparable? Meaningful and sustainable quantitative investigation of conceptual change in college science hinges on our ability to answer these questions.

Acknowledgements. This manuscript benefited from feedback provided by Margaret Hilton and the advisory board of the NRC.