Are They All Created Equal? A Comparison of Different Concept Inventory Development Methodologies

Rebecca S. Lindell, Elizabeth Peak and Thomas M. Foster

Physics, Astronomy, Chemistry, Biology, Earth Science Education Research Group
Southern Illinois University Edwardsville, Edwardsville, IL 62025, USA

Abstract. The creation of the Force Concept Inventory (FCI) was a seminal moment for Physics Education Research. Based on the development of the FCI, many more concept inventories have been developed. The problem with the development of all of these concept inventories is there does not seem to be a concise methodology for developing these inventories, nor is there a concise definition of what these inventories measure. By comparing the development methodologies of many common Physics and Astronomy Concept Inventories we can draw inferences about different types of concept inventories, as well as different valid conclusions that can be drawn from the administration of these inventories. Inventories compared include: Astronomy Diagnostic Test (ADT), Brief Electricity and Magnetism Assessment (BEMA), Conceptual Survey in Electricity and Magnetism (CSEM), Diagnostic Exam Electricity and Magnetism (DEEM), Determining and Interpreting Resistive Electric Circuits Concept Test (DIRECT), Energy and Motion Conceptual Survey (EMCS), Force Concept Inventory (FCI), Force and Motion Conceptual Evaluation (FMCE), Lunar Phases Concept Inventory (LPCI), Test of Understanding Graphs in Kinematics (TUG-K) and Wave Concept Inventory (WCI).

Keywords: Instrument Development, Concept Inventory Development
PACS: 01.40.Fk, 01.40.gf, 01.40.G-, 01.50.Kw

INTRODUCTION

Since the creation of the Force Concept Inventory (FCI) [1] the development of research-based distracter driven multiple-choice instruments has surged. Now nearly every scientific discipline has multiple concept instruments available for their use. A quick Google search yielded concept instruments in Physics [2], Astronomy [3, 4], Engineering [5], Biology [6], Chemistry [7], and Geoscience [8]. Nowhere was the prevalence more evident than at the 2006 Physics Education Research Conference, where 4 out of the 5 invited talks spoke about concept inventory development in their specific scientific fields.

The problem with the development of all of these instruments is that there does not seem to be a concise definition of what exactly a concept inventory actually measures. Many users of these instruments lump all conceptual tests under the classification of "concept inventory," while others argue that we need to differentiate the instruments based on standard definitions of surveys and instruments as defined by the education research community [9].

Not only is there a discrepancy in definitions of the term concept inventory, there also seems to be discrepancies in the methodologies utilized to create these inventories. The purpose of this project is to compare the different methodologies of conceptual instruments in Physics and Astronomy to determine the similarities and differences in the development methodologies.

For the purposes of this project we are adopting the following general definition of a concept inventory.

Concept Inventory: "A multiple-choice instrument designed to evaluate whether a person has an accurate and working knowledge of a concept or concepts."[10]

Results from this study yielded information about the different methodologies utilized to develop concept inventories, as well as providing evidence for the need for a new classification scheme for concept inventories.

METHODOLOGY

To begin this study, we selected a sub-population of concept inventories and assessments as listed on the North Carolina State University's Assessment Instrument Information website [2]. To further narrow our focus, the following criteria were used.

- Each instrument needed to focus on student understanding of either Physics or Astronomy concept(s).
- The methodology must have been published and/or communicated to us through private communication and the inventory made available.

This left us with a total of 12 "concept inventory" methodologies to analyze. The inventories are listed in Table 1.

TABLE 1. Concept Inventory Methodologies Analyzed In This Study

Instrument	# of Items	# of Concepts[1]
Astronomy Diagnostic Test (ADT) [3]	21	10
Brief Electricity and Magnetism Assessment (BEMA) [11]	30	3
Conceptual Survey in Electricity and Magnetism (CSEM) [12]	32	2
Diagnostic Exam Electricity and Magnetism (DEEM) [13]	66	
Determining and Interpreting Resistive Electric Circuits Concept Test (DIRECT) [14]	29	1
Energy and Motion Conceptual Survey (EMCS) [15]	25	2
Force Concept Inventory (FCI) [1]	30	1
Force and Motion Conceptual Evaluation (FMCE) [16]	43	2
Lunar Phases Concept Inventory (LPCI) [4]	20	1
Mechanics Baseline Test (MBT)[17]	36	3
Test of Understanding Graphs in Kinematics (TUG-K) [18]	21	1
Wave Concept Inventory (WCI) [19]	20	1

[1] We classified the number of concepts in terms of broad concepts and not specific sub-concepts.

To compare the methodologies, we began by studying the instrument design process as outlined in Crocker and Algina [20]. These steps are listed in Table 2. While there are many different texts and articles on test development and design, many focus on how to develop classroom tests and not educationally valid instruments designed for large-scale use. For the purposes of this study, we will refer to methodologies for educationally valid instruments as instrument design methodologies, as opposed to test development. Similar rubrics can be obtained from other test theory texts. For those unfamiliar with instrument design, Table 3 provides an overview of key concepts [20].

TABLE 2. Steps in Instrument Design

1. Identify purpose
2. Determine the concept domain
3. Prepare test specifications
4. Construct initial pool of items
5. Have items reviewed - revise as necessary
6. Hold preliminary field testing of items - revise as necessary
7. Field test on large sample representation of the examinee population
8. Determine statistical properties of item scores - eliminate inappropriate items
9. Design and conduct reliability and validity studies

TABLE 3. Overview of Key Concepts in Instrument Design

Concept Domain: Refers to the concept/concepts that will be covered on an instrument. It represents the content that will be assessed by the instrument. May contain alternative and scientific understanding as well. Sometimes referred to as Construct Domain.

Test Specifications: Details how items will be constructed. Typically represented as a table or diagram. The test map discussed by Aubrecht and Aubrecht [21] is just a visualization of a test specifications.

Field Testing: The process by which items are tested using a sample population. Item statistics are calculated to determine validity of items. Invalid items are deleted or revised. Field testing is often repeated until all items meet test specifications.

Item Statistics: There are two main statistics calculated in instrument development: difficulty and discrimination. While many test theory texts provide guidelines for excepting or rejecting an item based on these statistics, these guidelines are based on the assumption that responses are randomly distributed among the distracters.

Difficulty: How difficult is the item? Measures the percentage of respondents that answer item correctly.

Discrimination: Refers to how well the item discriminates between the upper quartile and lower quartile.

Validity: Represents how well an instrument measures the construct it is attempting to measure. There are three main types of validity: criterion, construct and content validity.

Criterion Validity: The degree to which scores on inventory predicts another criterion. Typically established through comparison to other standardized instruments or course grades.

Construct Validity: The degree to which scores can be utilized to draw an inference on the content domain. Typically established through open-ended student interviews.

Content Validity: The degree to which inventory measure the content covered in the content domain. Typically established through expert review.

Reliability: The degree to which scores are consistent.

Table 4. Methodology Comparison for Common Astronomy and Physics Concept Inventories[1]

	Concept Domain Determined by			Test Specifications				Item Statistics Reported			Field Testing					Validity Studies			Reliability Statistics Reported		
				Basis of Distracters		Distracter Correspondence to Alternate Models					Size			Location							
Instrument	Qualitative Study	Researcher	Existing Literature	Researcher's Understanding	Student understanding	Corresponds	Does Not Correspond	Difficulty	Discremenation	Concentration Analysis	>500 students	500 - 1000 students	>1000 students	Local	National	Criterion	Construct	Content	Cronbach Alpha	Kuder - Richardson	Point Biserial
ADT[2]	■	■	■	■			■	■	■	■			■		■		■	■			
BEMA		■		■			■	■	■			■		■				■		■	■
CSEM								■	■				■		■			■		■	
DEEM	■	■						■	■		■			■					■	■	
DIRECT		■	■				■	■	■				■		■		■	■		■	
EMCS		■			■		■	■	■				■		■		■	■	■		
FCI		■		■			■	■					■		■						
FMCE	■		■		■	■							■		■	■	■		■		
LPCI	■		■		■	■		■	■	■			■		■		■	■	■		
MBT		■		■	■		■						■	■			■	■		■	
TUG-K		■	■	■	■		■	■	■			■			■		■	■		■	■
WCI											■			■			■	■			

[1] All analysis based on research reported in original paper or personal communication. Blanks refer to non-reported information.

[2] Questions relating to Lunar Phases and Seasons were based on qualitative investigation conducted by R. Lindell; rest of concept domain determined by researchers.

For the purposes of this study, we focused on five different points in instrument design process: determining the concept domain, preparing the test specifications, determining the item statistics, field-testing of the items and conducting the reliability and validity studies.

RESULTS AND DISCUSSION

Table 4 shows the results of our analysis. As you can see there does not appear to be a consistent methodology being employed, nor do the different methodologies break down according to the accepted definitions of survey and inventory instruments. Below, we will discuss each of the five different points in the instrument design process separately.

Differences in Determining the Concept Domain

Developers utilized three different resources to define the concept domain. These included researcher's understanding, existing literature on students' understanding of the phenomena and performing a qualitative investigation of students' understanding. It must be noted, that we would argue that if the instrument is going to be utilized to diagnose specific alternative conceptions, the concept domain must be grounded in the views of the students and not the researchers. Of the twelve methodologies classified, only four utilized a qualitative investigation to help define the concept domain, and only two others utilized common student difficulties as reported in the literature.

Differences in Test Specifications

When analyzing the differences among the test specifications, we determined that there were two classes of differences: basis of the distracters and the correspondence of the different distracters to different alternate models.

Examining the basis of the distracters, we discovered that there were two main differences. Of the eight methodologies, which reported the basis of their distracters, three based the distracters purely on students' understanding, as discovered through open-ended questions or student interviews. Another three based the distracters purely on the researchers' understanding and two used a combination of the students' and researchers' understandings. It is interesting to note that four out of the twelve methodologies failed to discuss this key feature.

Since many of the concept inventories claim to be able to diagnose specific alternative understanding of the concept, we specifically examined the methodologies to determine if each distracter corresponded to a specific alternative model. In our opinion only two of the instruments meet this criteria and only these two inventories can be reliably used to diagnose alternative understandings.

Differences in Item Statistics

It is interesting to note that three of the different methodologies failed to report any item statistics. Typically these statistics are considered the bare minimum that should be reported in any instrument development. Of the remaining nine methodologies, all but one reported the two standard statistics of difficulty and discrimination. Two instruments utilized the new technique of concentration analysis [22] to help evaluate the items.

Differences in Field Testing

We found that there were two classes of field-testing: size of population and location of the field tests. We found that eight of the twelve instruments were field-tested at the national level, but only one failed to receive at least 1000 data points. Of the four locally field-tested instruments, only one exceeded a 1000 data points.

Differences in Establishing Reliability and Validity

Of the ten methodologies that reported the results of their validity study, only one reported establishing criterion validity and no methodology reported establishing all three types of validity. Finally only two methodologies reported only the lowest content validity results.

Reliability statistics were reported for only nine of the different methodologies. The reliability was typically determined by calculating one or more of the following statistics: Cronbach alpha, Kuder-Richardson 20 or 21, or the Point Biserial coefficient.

CONCLUSIONS

In conclusion, we found that there are many different methodologies being utilized to develop concept inventories. We as a community need to determine guidelines for developing these instruments. We also find that the definition for concept inventories is way too broad and we need to introduce a new classification scheme. Lastly we strongly encourage developers to employ all of the steps in the design process, as well as to publish their methodologies so as the community can determine the appropriateness of utilizing the instruments.

REFERENCES

1. D. Hestenes, M. Wells and G. Swackhamer, *Phys. Teach* **30**, 141-158 (1992).
2. Assessment Instrument Information Page, North Carolina State University, http://www.ncsu.edu/per/TestInfo.html, accessed October 15, 2005.
3. B. Hufnagel, *Astro Ed Rev.* http://aer.noao.edu **1**(1), 47-51 (2002)
4. R. Lindell and J. Olsen, *Proc. 2002 PERC,* New York: PERC Publishing, NY, (2002).
5. Engineering Foundation Coalition, http://www.foundationcoalition.org/home/keycomponents/concept/index.html.
6. D. Anderson, K. Fisher and G. Norman, *J. Res. Sci Teach* **39** (10), 952-978 (2002).
7. D. Mulford and W. Robinson, *J. Chem. Ed.* **79**(6), 739-44 (2002).
8. J. Libarkin and S. Anderson, *J Geosci. Ed.* **53**, 394-401 (2005).
9. Wikepedia definition, http://en.wikipedia.org/wiki/Concept_inventory, accessed October 15, 2005.
10. Education researchers consider a survey to contain 5-7 items per concept, while with inventories you need 20+ items per concept. P. Heller, Personal Communication, 2006.
11. L. Ding, R. Chabay, B. Sherwood and R. Beichner, Personal Communication, (2006).
12. D. Maloney, T. O'Kuma, C. Hieggelke & A. Van Heuvelen, *Am. J. Phys* **69**, S12-S23 (2001).
13. J. Marx, Unpublished Dissertation*,*: Rensselaer Polytechnic Institute (1998).
14. P. Engelhardt & R. Beichner, *Am. J. Phys* **72**, 98-115 (2004).
15. C. Singh and D. Rosengrant, *Am. J. Phys* **71**, 607-617, (2003).
16. R. Thornton and D. Sokoloff, *Am. J. Phys* **66**, 338-352 (1998) and personal conversation with R. Thornton (2006).
17. I. Halloun and D. Hestenes, *Am. J. Phys* **53**, 1043-1055 (1985).
18. R. Beichner, *Am. J. Phys* **62**, 750-762 (1994).
19. R. Roedel, S. El-Ghazaly, T. Rhoads & E. El-Sharawy, *29th ASEE/IEEE Frontiers in Education Conference* San Juan, Puerto Rico, 10–13 November 1999.
20. L. Crocker and J. Algina, *Introduction to classical and modern test theory*., Reinhart and Winston, Inc, New York: Holt (1986).
21. G. J. Aubrecht, II and J. D. Aubrecht, *Am. J. Phys* **51**, 613-620 (1983).
22. L. Bao and E. F. Redish, *Am. J. Phys*. **69**, S45 (2001).