

This article was downloaded by: [University of Arizona]

On: 19 March 2012, At: 10:05

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



International Journal of Science Education

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/tsed20>

Development and Validation of the Star Properties Concept Inventory

Janelle M. Bailey^a, Bruce Johnson^b, Edward E. Prather^c & Timothy F. Slater^d

^a Curriculum & Instruction, University of Nevada Las Vegas, Las Vegas, USA

^b Department of Teaching, Learning, and Sociocultural Studies, The University of Arizona, Tucson, AZ, USA

^c Steward Observatory, The University of Arizona, Tucson, AZ, USA

^d Secondary Education, University of Wyoming, Laramie, WY, USA

Available online: 28 Jul 2011

To cite this article: Janelle M. Bailey, Bruce Johnson, Edward E. Prather & Timothy F. Slater (2011): Development and Validation of the Star Properties Concept Inventory, *International Journal of Science Education*, DOI:10.1080/09500693.2011.589869

To link to this article: <http://dx.doi.org/10.1080/09500693.2011.589869>



PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings,

demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

RESEARCH PAPER

Development and Validation of the Star Properties Concept Inventory

Janelle M. Bailey^{a*}, Bruce Johnson^b, Edward E. Prather^c and Timothy F. Slater^d

^a*Curriculum & Instruction, University of Nevada Las Vegas, Las Vegas, USA;*

^b*Department of Teaching, Learning, and Sociocultural Studies, The University of Arizona, Tucson, AZ, USA;* ^c*Steward Observatory, The University of Arizona, Tucson, AZ, USA;* ^d*Secondary Education, University of Wyoming, Laramie, WY, USA*

Concept inventories (CIs)—typically multiple-choice instruments that focus on a single or small subset of closely related topics—have been used in science education for more than a decade. This paper describes the development and validation of a new CI for astronomy, the *Star Properties Concept Inventory* (SPCI). Questions cover the areas of stellar properties (focussing primarily on mass, temperature, luminosity, and lifetime), nuclear fusion, and star formation. Distracters were developed from known alternative conceptions and reasoning difficulties commonly held by students. The SPCI was tested through an iterative process where different testing formats (open-ended, multiple-choice + explain, and multiple-choice) were compared to ensure that the distracters were in fact the most common among the testing population. Content validity was established through expert reviews by 26 astronomy instructors. The SPCI Version 3 was then tested in multiple introductory undergraduate astronomy courses for non-science majors. Post-test scores (out of 23 possible) were significantly greater ($M = 11.8$, $SD = 3.87$) than the pre-test scores ($M = 7.09$, $SD = 2.73$). The low post-test score—only 51.3%—could indicate a need for changing instructional strategies on the topics of stars and star formation.

Keywords: *Assessment; Earth science education; Astronomy; College non-science majors*

Introduction

Although studies in astronomy education research go back many years and are increasing in scope and number (Bailey & Slater, 2003; Lelliott & Rollnick, 2010), the results

*Corresponding author: Curriculum & Instruction, University of Nevada Las Vegas, CEB 354, 4505 S. Maryland Pkwy, Box 453005, Las Vegas, NV 89154-3005, USA. Email: janelle.bailey@unlv.edu

of such studies are only recently impacting the introductory astronomy course for non-science majors in meaningful ways. Hereafter, we refer to this introductory course, a popular one within USA's tertiary institutions (Fraknoi, 2001; Partridge & Greenstein, 2003; Rudolph, Prather, Brissenden, Consiglio, & Gonzaga, 2010), as 'ASTRO 101'. Both traditional and action research studies in ASTRO 101, as well as their brethren in other science disciplines, are affecting how instructors design their courses to better facilitate meaningful learning. A challenge in such studies is finding appropriate assessment tools that can be used to measure variables of interest.

A *concept inventory* (CI) is one possible tool for assessing knowledge and learning in introductory science courses. A CI is typically a multiple-choice instrument that focusses on a single topic or small subset of closely related topics, containing numerous questions on each idea in order to gauge a student's understanding of the content in a variety of ways (Bailey, 2009). By using a CI, an instructor can get a better indication of what specific ideas students understand when they enter the course (when used as a pre-test) and what they continue to struggle with after instruction (post-test). This can then lead to improvements in the course materials or presentation of those materials on the relevant topics.

In astronomy to this point, a few assessment instruments have been developed to inform instruction when used as a pre-test/post-test, though only two of these satisfy the description of CI as we use it here. The Lunar Phases Concept Inventory (LPCI; Lindell, 2001; Lindell & Olsen, 2002) has been validated for use with college students, but has not been widely adopted. This may be the result of the limited time spent on the topic of lunar phases in a college-level course. The Light and Spectroscopy Concept inventory (LSCI) has been developed and tested on both small and large scales (Bardar, Prather, Brecher, & Slater, 2007; Prather, Rudolph, & Brissenden, 2009). The nature of light and how it is used to learn about astronomical objects and phenomena are among the most critical ideas in astronomy, and so the LSCI can serve as a valuable tool to evaluate student understanding of these ideas (Bardar, Prather, Brecher, & Slater, 2005).

An additional assessment, the astronomy diagnostic test (Deming, 2002; Hufnagel, 2002; Hufnagel et al., 2000; Zeilik, 2003), has been more widely used as pre-test/post-test to gauge student learning over a course (Brogt et al., 2007). However, this instrument was designed to look at student understanding of a variety of topics commonly covered in K-12 science, and its usefulness to look at a single content topic is limited. Likewise, the newer Astronomy and Space Science CI (Sadler et al., 2009) and Test Of Astronomy STandards (TOAST; Bailey, Slater, & Slater, 2011) focus more broadly on K-12 astronomy standards, and so do not follow the definition of CIs provided above.

Common ASTRO 101 topics go far beyond light and lunar phases, however (Slater, Adams, Brissenden, & Duncan, 2001), and so additional CIs are of interest to instructors. One such topic is stars. On a macroscale level, stars can be considered the building blocks of our universe. Their nuclear processes over extensive lifetimes convert light elements (hydrogen and helium) to heavier ones (including carbon and oxygen, through the periodic table, to uranium). Stars' gravitational interactions

with their environments can contribute to large-scale structure, such as stellar and planetary systems, star clusters, and galaxies. Our star, the Sun, provides the energy used in our planet's natural cycles. Thus, stars are an important part of the ASTRO 101 course (further illustrated by textbook coverage, as described below).

Prior research has shown that students enter ASTRO 101 courses with a number of alternative conceptions about these objects (Bailey, Prather, Johnson, & Slater, 2009). Although a detailed treatment of these alternative conceptions is beyond the scope of this paper, let us consider three examples. (a) Students frequently describe stars as being powered by burning or other chemical reactions, rather than nuclear fusion (itself sometimes confused with fission). (b) Students frequently characterize star formation as 'gas (and dust) coming together', although only a small percentage cite gravitational forces as the causal mechanism. Alternative ideas include magnetism and rotational forces. (c) Students believe white stars (not a valid astronomical classification) or red stars to be hottest, rather than blue. These and other student ideas are explored in greater detail in Bailey et al. (2009) and references therein.

Purpose of the Study

Because CIs have proved useful to instructors in a variety of domains, we sought to develop and test one for star properties, a topic with no previous instrument or even an appreciable number of questions on comprehensive assessments. The purpose of this study was to create and test a CI to serve as a measure of student learning about the properties and formation of stars. As part of the development process, we asked two research questions:

- (1) To what degree is the instrument valid and reliable, and what evidence supports this claim?
- (2) How do ASTRO 101 students compare on the instrument pre-test to post-test?

This paper first describes additional background about CIs and their development. We next discuss the creation of this instrument, which included multiple stages in an iterative development process. The methods and results of each version are described in turn, using a chronological presentation. Finally, we return to the research questions and discuss implications of this study for future efforts in research and practice of astronomy teaching and learning.

Background

Concept Inventories

A CI is a multiple-choice instrument that focusses on a narrow concept (or small set of related concepts) and whose distracters (i.e. wrong answers) are based upon known student learning difficulties (Bailey, 2009). A second group of instruments, known as 'two-tier diagnostic tests', are similar in scope to CIs as described above (Treagust, 1986, 1988). However, in this case, each question has a second part in which students also select a multiple-choice option that best matches their reasoning process—in

other words, why the student chose their answer. ‘Three-tier diagnostic tests’ add another layer by asking students to rate their confidence in their answer to each question (Caleon & Subramaniam, 2010).

CIIs typically are used for two main purposes. First, science education researchers use them as common assessments that can be used to evaluate achievement gains and to compare such gains over different situations. For example, a meta-analysis by Hake (1998) and a national study by Prather, Rudolph, Brissenden, and Schlingman (2009) both use CIIs as the basis to compare a set of very different classrooms. Second, individual instructors use CIIs as a way of evaluating their own instructional effectiveness and diagnosing common student problems (e.g. LoPresto & Murrell, 2009).

The CI differs from other kinds of assessments used by instructors. Self-created exams, especially final exams, often include a wider variety of content with fewer items (perhaps only a single question) on any given concept. Distracters may be based upon the instructor’s experience with incorrect student ideas, but may not take full advantage of careful research that identifies alternative conceptions. Such exams also typically do not undergo the more rigorous testing that is found with CIIs and similar assessments. The CI also does not attempt to measure any affective variables, such as interest, motivation, or achievement goal orientation; the final layer of the ‘three-tier diagnostic test’ is the exception here (Caleon & Subramaniam, 2010). Students often are assessed through other types of instruments as well, such as the SAT[®] or college entrance exams. These, however, tend to measure a larger scope of general knowledge and skills and rarely directly inform the content of an individual course in the way that a CI can.

CIIs exist on a wide variety of topics. In addition to the astronomy instruments described above (i.e. the LPCI and LSCI), popular CI topics include force and motion (Force Concept Inventory [FCI]; Halloun & Hestenes, 1985), natural selection (Anderson, Fisher, & Norman, 2002), and geoscience (Libarkin & Anderson, 2005). Other CIIs and related assessments can be found for many subtopics of physics, and have been or are being developed in other disciplines such as biology, chemistry, mathematics, and engineering. Treagust (1988) described how such tests originated and how they differed from the traditionally more qualitative research on student misconceptions. More recently, Sadler et al. (2009) described the use of ‘distracter-driven multiple-choice’ questions (p. 2) over the last several decades these questions are the type used in CIIs.

Treagust (1988) also described a development process for the two-tier diagnostic test, which has been used (or modified) to create a number of instruments. He presented three broad areas required for creating these tests: ‘defining the content’ (p. 161), ‘obtaining information about students’ misconceptions’ (p. 162), and ‘developing a diagnostic test’ (p. 163). These areas were then divided into 10 individual steps, listed in Table 1.

In contrast to Treagust’s (1988) presentation of recommended steps, Lindell, Peak, and Foster (2007) reviewed the design and validation methodologies of 12 CIIs in physics and astronomy as presented. They looked at five broad areas: determining

Table 1. Steps to developing a CI, from Treagust (1988, pp. 161–164)

-
1. Identifying propositional knowledge statements
 2. Developing a concept map
 3. Relating propositional knowledge to a concept map
 4. Validating the content
 5. Examining related literature
 6. Conducting unstructured student interviews
 7. Developing multiple-choice content items with free response (to explain the choice made)
 8. Developing the two-tier diagnostic tests
 9. Designing a specification grid
 10. Continuing refinements
-

the concept domain; creating the test specifications; reporting item statistics; defining the field testing population; and reporting reliability and validity statistics. Lindell et al. (2007) found that CI developers used and reported, either in publication or through personal communication, a wide variety of design methodologies.

Although the CI development described in this paper did not follow Treagust's (1988) steps exactly as prescribed, the process did share many of the same elements. In particular, Treagust's items 1, 4–7, and 10 (Table 1) were addressed in our method in an iterative process. All five of the areas described by Lindell et al. (2007) were included in the development process and are summarized here (see Bailey, 2006, for full details).

Test Development

Our instrument development process was guided by classical test theory (CTT), which assumes that a test taker will produce an observed score that is a sum of a theoretical (but immeasurable) true score plus a random error score (Allen & Yen, 1979; Osterlind, 2010). The CTT's assumptions and common measures, such as reliability and validity, will be described as they relate to this instrument in the results below. Item response theory (Embretson & Reise, 2000) has also been applied to Version 3 of this instrument; those results are described elsewhere (Wallace & Bailey, 2010).

Overview of the Design and Testing Process

In general, our testing process followed an iterative design in which a version of the instrument was created, administered to a sample of participants, and analyzed, with revisions made based upon the outcome of the analysis. The subsequent version was then administered to a sample of participants and the process repeated. Three versions were tested and will be discussed below. In addition, we conducted interviews with participants as part of the Version 1 process, and experts reviewed a Version 2.5 in between the administrations of Versions 2 and 3. Figure 1 illustrates this process. We describe each iteration in chronological order below (i.e. discussion of the methods and results of Version 1 is presented before moving on to Version 2).

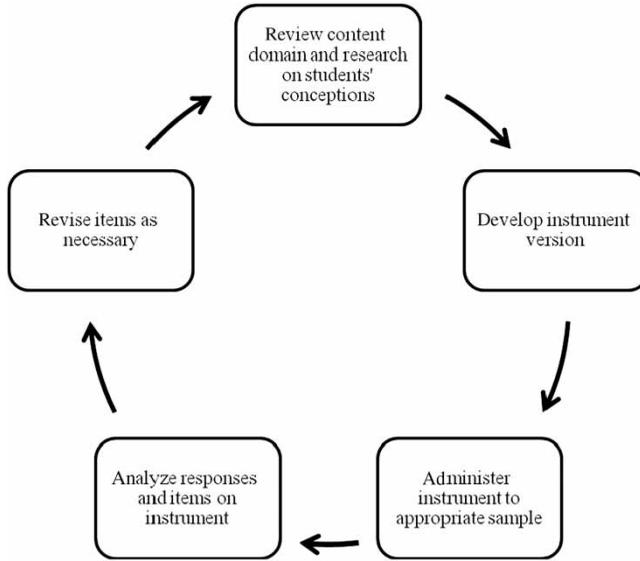


Figure 1. A diagram illustrating the iterative development process of the Star Properties Concept Inventory

Limitations of the Research Design

Some limitations of this research design potentially constrain the generalizability of the findings. The first is that all of the data were collected at a single institution. While there is no reason to believe that the students at this institution are dramatically different from students across the USA in terms of their demographics, motivation toward general education coursework, or astronomy knowledge (Rudolph et al., 2010; Slater et al., 2001), testing this claim was beyond the scope of this study. A validation study that includes ASTRO 101 courses from multiple institutions is currently in progress.

Interview volunteers were recruited without any compensation for their time and effort. Because of this, there is a self-selection effect that may bias the data (Seidman, 1998). Although the recruitment of interviewees used a stratified, random-sampling approach in order to include a range of student performances, the method is still limited by the students who chose to participate in the interview portion of the study.

The final limitation in the generalizability of this study is in the possible match to instructors' needs. If an instructor's course design differs dramatically from what has been assumed in the design of the research study (e.g. she has an appreciably larger emphasis on the solar system or, conversely, her course includes only stars and galaxies), she may determine that the alignment of content is inadequate for her use. While part of the goal of the expert review was to reduce this possibility, this is itself limited by a somewhat small sample size ($n = 26$) relative to the population of astronomy instructors.

Setting and Participants

We conducted this study at a large research university in the southwestern USA. The institution enrolls more than 28,000 undergraduate students, approximately 53% of whom are female and 47% male. About 65% of the undergraduate students are Caucasian, with another 15% Hispanic. Other ethnicities comprise about 14% of the population, with 6% unspecified. More than two-thirds of the undergraduates are aged 18–21.

The first group of participants in this study comprised undergraduate non-science majors enrolled in an ASTRO 101 course. Students in this course are typically, though not exclusively, in their first year of college and frequently are enrolled in the course to satisfy a general education requirement in the natural sciences (Deming & Hufnagel, 2001; Rudolph et al., 2010). The students in these courses are generally representative of the university's undergraduate population in terms of gender, age, and ethnicity.

A second group of students participated in this study in order to increase the number of participants, particularly in early development stages. These students were enrolled in a different general education natural science course, focussed on topics other than astronomy and which did not include any formal instruction on star properties or formation during the semester. These courses will be called 'Earth Science 101' (hereafter 'ES 101') for simplicity, as this reasonably describes the majority of the participating classes. The demographic distribution of these courses, in terms of gender, age, class, and ethnicity, is approximately the same as the ASTRO 101 courses.

The general education natural science courses at this institution are predominantly lecture-based survey courses, typically serving 100–300 students per section. Lectures are held in large, auditorium-style classrooms, and there is no separate laboratory component to the course. ASTRO 101 introduces students to a wide range of foundational topics related to observational and theoretical astronomy, using both historical and contemporary contexts as appropriate. ES 101 courses focuss on topics such as physical geography, earth resources, and environmental science, again using both historical and contemporary contexts.

We invited ASTRO 101 and ES 101 instructors to participate through an electronic mail message describing the study purpose and requirements, with follow-up communication individually as needed. Fifteen instructors provided access to their courses over two semesters, for a total of 16 sections and more than 2,000 students. Table 2 summarizes the numbers of ASTRO 101 and ES 101 sections and students participating in the study.

Version 1

Before describing the methods and results of our instrument's Version 1, we describe the scientific context of this study by looking at the ASTRO 101 course and what place stars have within the curriculum. This process helped us to define the construct of stars as it is used in the instrument.

Table 2. Number of respondents by semester of administration and class

Classes		Spring 2005		Fall 2005	
		ASTRO 101	ES 101	ASTRO 101	ES 101
PRE	Sections	6	2	5	3
	<i>n</i>	796	169	690	411
POST	Sections	6	1	5	2
	<i>n</i>	469	76	489	155

Scientific Context of the Study

The ASTRO 101 course is generally an introductory survey of astronomy, including all major topics in the field from both modern and historical contexts (Slater et al., 2001). At some institutions, the course may be split into two terms (quarters or semesters). In these cases, the course is often divided into ‘solar system’ and ‘stars, galaxies, and cosmology’ portions, though details may vary between institutions. Such splits are often parallel to those found in popular astronomy textbooks (Bruning, 2002, 2006a, 2006b).

We know from prior research that stars (including their properties and evolution, as well as the Sun) are one of the most frequently addressed topics in the typical ASTRO 101 course (Slater et al., 2001). However, to better understand the extent to which stars and star formation are covered, we also looked at textbook treatment of these topics. A survey of the 23 introductory astronomy textbooks published at the time shows that, on average, approximately 25% of a book’s text (150 pages or about six chapters) is dedicated to stars; the Sun may be covered here or as part of the 28% of the text (approximately 171 pages) that covers the solar system (Bruning, 2006b). Chapters relating to stars include some of the following subtopics, as discussed in three popular textbooks and identified through section titles, text boxes, and boldfaced vocabulary words: evolutionary processes, Hertzsprung–Russell (H–R) diagram, hydrostatic equilibrium, luminosity–distance relationship, nuclear fusion, spectra and chemical composition, stellar lifetimes, stellar structure, temperatures and colors (Bennett, Donahue, Schneider, & Voit, 2004; Kaufmann & Freedman, 1999; Zeilik, 2002). cursory inspection of other textbooks shows similar coverage. In contrast to the seemingly long list of subtopics, it is important to note that the typical one-semester ASTRO 101 course will likely cover stars and star formation in less than 2 weeks, or fewer than 6 contact hours with students.

Despite such large coverage in textbooks, a definition of a star is not always stated explicitly. For example, only 1 of the 3 textbooks used to create the list above includes the term ‘star’ in its glossary. Bennett et al. (2004) define a star as:

A large, glowing ball of gas that generates energy through nuclear fusion in its core. The term *star* is sometimes applied to objects that are in the process of becoming true stars (e.g., protostars) and to the remains of stars that have died (e.g., neutron stars). (p. G-11)

In the other 2 textbooks, stars are not explicitly defined. Rather, the meaning is defined implicitly through the main text.

In addition to the basic idea of what a star is, the ASTRO 101 student is typically expected to understand, after instruction, a variety of content relating to the properties of and models used to describe stars. For example, instructors may want their students to understand that there are observable differences between stars (such as in apparent brightness, color, or spectral signature), and that these differences can be interpreted to mean that there are related physical variations, some of which support the idea of multiple evolutionary stages. Students might also be expected to learn the relationships between such properties; how stars form and evolve over time; or the processes of nuclear fusion at different stages of stars' lives. These and related ideas formed the basis for the content of this study.

Methods

In the first step of the instrument creation, we developed 30 multiple-choice questions, taken from our own previous work (such as course exams), modified from other sources (such as textbook questions), or newly created for this instrument to fill in any content gaps. Previous review of the literature and research conducted as an earlier phase of this study showed that students have a number of alternative conceptions about stars (Bailey, Prather, Johnson, & Slater, 2009). This work used open-ended questions and a small subset of interviews to investigate students' pre-instructional ideas about stars. The responses gathered in the previous study formed the basis of the questions' distracters (i.e. incorrect choices) in the present study described here. Using research-based distracters helps to ensure that instrument addressed a wide range of student ideas.

At the same time, questions and possible responses were written with attention to and use of students' natural language wherever possible. This language frequently comes from the alternative conceptions research and development process, through interviews or responses to open-ended questions. Using students' natural language provides access to students' pre-instructional beliefs, in that students may have ideas about concepts, objects, or processes but do not know the scientific language to match. Conversely, natural language also provides a way to investigate students' conceptual understanding after instruction, something that might be masked by rote memorization of scientific terminology. In other words, students may use the correct vocabulary but still hold alternative ideas about what that vocabulary means, as has been shown in many studies in physics (e.g. Halloun & Hestenes, 1985; Prather, 2000; Prather & Harrington, 2001). As just one example relevant to this instrument, consider the term 'nuclear fusion'. Although we used this phrase in some questions, we also wrote similar questions that instead used phrases like 'the combining of elements into new elements'. Thus, even if a student did not know the scientific term they may still be able to recognize the process. Additional instances where the use of natural language is particularly important will be described in the Results sections below.

In addition to the 30 content questions, students were asked three other questions. The first (Question 31) served as the initial recruitment for interviews that would be conducted later in the semester (described below). Question 32 asked the student's gender, in order to identify whether results differed between female and male participants. Finally, Question 33 asked whether the student previously had taken an astronomy course. No other demographic questions (e.g. ethnicity, prior physics, or mathematics coursework) were included because they were beyond the scope of this study. We named the 33-item instrument the 'Star Formation Concept Inventory (SFCI), Version 1'.

During Spring 2005, the SFCI Version 1 was administered in three different formats, in order to investigate the validity of the items and their distracters. By looking at participants' responses to different question formats, we compared the results to ensure that the question was being answered in the same way, and if not, we revised the question accordingly. In Version 1a, questions were presented in a multiple-choice format. In Version 1b, the same multiple-choice questions were given, but participants were additionally asked to explain the reasoning behind their choice for each question. Finally, Version 1c contained only the stem of the multiple-choice question (i.e. no distracters were provided) and participants were asked to provide short answers with explanations. Wording remained the same when possible, though some questions required slight changes for clarity. An example of one question in each of the three different formats is presented in Table 3.

We created multiple forms, containing a subset of questions, for each version in order to reduce and balance administration time. For example, Version 1b-1 contained only questions 1–8 plus the common questions 31–33; a total of 4 forms were used for Version 1b while 6 were used for Version 1c and 2 for Version 1a. The three formats were photocopied and mixed into stacks, so that the versions were randomly distributed to the participants.

SFCI Version 1 was administered to 796 students in 6 ASTRO 101 sections and another 169 students in 2 ES 101 sections as a pre-test on the first day of class. The breakdown of students who completed each form is given below as part of Table 4. Because no question was repeated within a given format (i.e. Versions 1a, 1b, or 1c), each question was answered by approximately 160 students in multiple-choice format and another 80 students in open-ended format. Names were collected on the responses in order to recruit potential interview volunteers, but were dissociated from the surveys after identification numbers had been randomly assigned.

Interviews were conducted as a way to investigate the internal consistency and validity of the instrument. By determining whether students interpreted the CI questions in the same way at different times, we can have confidence in the instrument's ability to test the same constructs over time (i.e. providing evidence of reliability). Ensuring construct validity (i.e. ability of the instrument to measure the defined construct) is supported through interviews because we can find out if students are interpreting the questions in the way they were intended.

On Version 1, respondents were asked to indicate their willingness to participate in an interview (Question 31). Volunteers from the participating ASTRO 101 courses

Table 3. Sample questions from different formats of SFCI Version 1

Form	Sample question
Version 1a	<p>Arrange the following stars by surface temperature from hottest to coldest: white stars, blue stars, and red stars.</p> <p>Hottest → coldest</p> <p>a. red > white > blue</p> <p>b. red > blue > white</p> <p>c. white > red > blue</p> <p>d. blue > white > red</p> <p>e. blue > red > white</p>
Version 1b	<p>Arrange the following stars by surface temperature from hottest to coldest: white stars, blue stars, and red stars.</p> <p>Hottest → coldest</p> <p>a. red > white > blue</p> <p>b. red > blue > white</p> <p>c. white > red > blue</p> <p>d. blue > white > red</p> <p>e. blue > red > white</p> <p>Explain the reasoning behind the choice you made.</p>
Version 1c	<p>Arrange the following stars by surface temperature from hottest to coldest: white stars, blue stars, and red stars. Why did you choose the order you did?</p>

were then selected using a stratified, random-sampling approach (Seidman, 1998). Based upon their score on the SFCI Version 1, students were ranked as high-, middle-, or low-scoring. Between 10 and 25 volunteers from each scoring group were randomly selected and solicited via electronic mail. Participants were reminded that interviews were voluntary and that they could refuse at this time if they desired. Two or three participants from each scoring group and each form of Version 1c were interviewed, for a total of 18 interviews. Each interview lasted between 20 and 45 minutes and was audio recorded.

Table 4. Number of students completing different formats of SFCI Versions 1a and 1b ($n = 480$)

Form	n	M (%)	SD (%)	Skewness	Kurtosis	Cronbach's α
1a-1	78	40.3	12.6	0.258	0.001	0.112
1a-2	80 ^a	42.3	13.2	-0.324	-0.192	0.284
1b-1	79	31.7	18.7	0.139	-0.504	0.263
1b-2	81	47.4	19.1	0.128	-0.074	0.375
1b-3	81	49.5	17.2	0.121	-0.600	0.168
1b-4	81	29.6	19.0	0.384	0.071	0.309
All MC forms combined ^b	480	40.1	18.4	0.053	-0.024	—

^aEighty-one participants completed this form; one was removed as an outlier whose score was greater than 3σ above that form's mean.

^bBetween 78 and 82 participants completed each of the six forms in Version 1c in addition to this total, for a grand total of $n = 965$.

During the interviews, participants were asked to do three things. First, they responded to the same survey form as they had completed at the start of the semester, now thinking aloud as they answered the questions and elaborating on their responses wherever possible. Next, we compared the participant's interview responses with his or her pre-test and discussed any differences with them. Finally, interview participants were asked to look at the equivalent multiple-choice questions (from Version 1a) and describe what they would have chosen (and why) if they had received the multiple-choice format instead. We created a table for each interviewee showing his or her original written response, interview response, and corresponding multiple-choice response, along with reasons for any noted differences, comparing the participant's written responses and our notes with audio recordings to verify accuracy. The results of the student interviews are described below, after a discussion of the results from the pre-test administration.

Results: Instrument analysis

We calculated a score (percent correct) for the multiple-choice format surveys (Versions 1a and 1b) based only upon the number of questions on that form (i.e. if the student had Version 1a-1, with questions 1–15, the score was calculated out of 15). Descriptive statistics for each form are provided in Table 4. The mean score for all participants on all forms was $M = 40.1\%$, $SD = 18.4\%$ ($n = 480$). Absolute values of skewness and kurtosis below 1, which were the case for each of our forms and for the multiple-choice format sample as a whole, are considered excellent indicators of a normal distribution of responses (George & Mallery, 2009).

Cronbach's α was calculated for each form to gauge reliability. Although each of these values was considerably lower than the value considered acceptable ($\alpha = 0.7$) (George & Mallery, 2009), at this point in the analysis we were more concerned about vetting individual questions than overall instrument reliability. As revisions were made to Version 1 to create Version 2, and so on, an increase in reliability was targeted (and achieved, as described below) in coordination with our efforts to increase validity evidence.

Version 1c responses were analyzed through a two-step process. We first scored the open-response questions by matching the students' written responses to the choices provided on the multiple-choice format used in Versions 1a and 1b. If the written response matched the correct answer, it was scored correct; if it matched one of the distracters or did not match any of the options, it was scored incorrect. For the second step, the collection of unmatched responses was analyzed for themes through a grounded theory approach (Glaser & Strauss, 1967). After themes were identified, we determined the frequency of each theme's occurrence within the open-ended responses. These frequencies were then compared with the distribution of responses to the choices provided on Versions 1a and 1b. If a newly emerged theme had a higher frequency of responses than one of the distracters, that distracter was changed to the higher-frequency response for Version 2.

An example of this process is illustrated by the analysis of the following question. The multiple-choice version of one question from Versions 1a and 1b is below, with the combined response frequencies for these versions indicated in parentheses.

Stars are made mostly of which element when they first form?

- a. helium (7%)
- b. hydrogen (60%)
- c. carbon (26%)
- d. oxygen (3%)
- e. silicon (1%)

The phrasing of the question was changed slightly for Version 1c, using the word ‘elements’ instead of the singular ‘element’.

For Version 1c responses that could be matched to the multiple-choice options, the distribution was: (a) helium, 10%; (b) hydrogen, 32%; (c) carbon, 11%; (d) oxygen, 7%; and (e) silicon, 0%. Other responses were included by 65% of the participants (the total was greater than 100% because students could list more than one element in the open-ended format). Included in the ‘others’ category were the elements nitrogen, iron, boron, and mercury, plus other substances such as carbon dioxide, ammonia, water, air, dust, rocks, and minerals. Because nitrogen was provided by about 5% of the respondents on Version 1c, we replaced option ‘(e) silicon’ on Version 2. The new question then read:

Stars are made mostly of which types of atoms when they first form?

- a. oxygen
- b. nitrogen
- c. carbon
- d. helium
- e. hydrogen

Note that options a–e were re-ordered in this and other questions to better distribute the correct answers across the instrument.

Results: Student interviews

Overall, students were quite consistent with their responses between the pre-test instrument administration and the interviews. Internal consistency was measured by calculating the correlation between the written pre-test and the oral interview responses. Participants were also asked to indicate the multiple-choice response they would have selected; a total of 48 of the 90 questions (53%) were answered correctly over all 18 interviews. Correlations also were calculated between the oral interview and the multiple-choice responses and between the written pre-test and the multiple-choice responses. All three correlations were significant at the $p < 0.01$ level (Table 5), indicating consistent responses between the written pre-test and the oral interview.

Table 5. Bivariate correlations between different response types to Version 1c ($n = 90$ questions)

Response type	1	2	3
1. Written pre-test response	—		
2. Oral interview response	0.478**	—	
3. Oral multiple-choice selection	0.285**	0.535**	—

** $p < 0.01$.

In the cases where an answer to a question changed, students attributed the change to 1 of 3 things. In nearly half of the instances, participants indicated that they had recently learned about the question topic in class. Because of the recruitment and scheduling process, interviews were performed during weeks 4–8 of the semester, and therefore, such influence could not be avoided. In most of the rest of the cases where an answer changed, the participant indicated that he or she was guessing in both instances, and simply happened to guess different things. Finally, upon questioning the change in two cases, the participants expressed surprise at their original, pre-test answer and suggested that they had been mistaken on the pre-test.

Version 2

Methods

Upon analyzing the results of Version 1 and the interviews, we made changes to some of the questions on the CI to create Version 2. Eleven of the 30 content questions remained unaltered for Version 2 other than the correction of minor typographical errors. The stem or at least one distracter was changed on 13 questions to better represent students' ideas and language as determined from Version 1c and/or student interviews. We also revised six questions to clarify the scientific concept being tested or the question being asked. Across the entire CI, 'stellar' was changed to 'star's' and 'temperature' was changed to 'surface temperature'. Finally, we reordered all items to better separate questions dealing with similar concepts and varied question choices to better distribute the correct answers (over options a–e). These changes attempted to provide greater evidence of validity, by improving upon the measurement of the construct of stars, and reliability, by affording greater consistency across students and administrations.

Version 2 was divided into two different formats. Version 2a contained all multiple-choice questions, while Version 2b contained multiple-choice plus 'explain your reasoning'. The 30 content questions were again distributed over multiple forms (two for Version 2a and four for Version 2b) to reduce administration time.

Version 2 of the CI was administered as the post-test to the participating classes during the last 2 weeks of the Spring 2005 semester. A total of 545 participants completed the post-test. This was only 56% of the number of pre-test participants, and can be attributed to two factors. On an individual level, many students either dropped the course or were not in attendance on the day of the post-test. Post-test

attendance rates in each section ranged from 45 to 85% of the pre-test attendance, the result of both attrition and absenteeism. On a group level, one of the ES 101 instructors who had agreed to participate in the study was unable to schedule the post-test before the end of the semester, and so a large number of ES 101 students were unable to take the post-test. The number of students who responded to each of the different forms is presented in Table 6.

Results: Instrument analysis

As with Version 1, percentage scores for Version 2 were calculated out of the number of questions possible (e.g. out of 8 possible for Version 2b-1 but out of 15 possible for Version 2a-1). The mean score for all participants on all forms was $M = 54.9\%$, $SD = 22.6\%$ ($n = 545$). We found skewness and kurtosis absolute values again below the accepted value of 1 (George & Mallery, 2009), indicating normal distributions within each form as well as for the sample as a whole.

We again calculated Cronbach's α for each of the forms (Table 6). These values were appreciably higher than Version 1, indicating a move toward greater instrument reliability. However, none of the forms by itself had an alpha high enough to be considered acceptable (George & Mallery, 2009). Because the pre-test and post-test were not identical instruments, it was not appropriate to compare pre-test and post-test results to investigate possible learning gains at this time.

Version 2.5

Methods

Additional changes were made to the CI as a result of the Version 2 analysis to create Version 2.5. Nine of the 30 questions remained unchanged. Distracters were changed in two questions to include nuclear fission when nuclear fusion was also an option, in order to determine if students could distinguish between the two different processes rather than just recognizing the phrase without understanding it. Twelve questions were changed to improve clarity or to better reflect students' ideas and natural language as provided through explanations on Version 2b. Finally, 7 questions were removed for being too low-level or outside the scope of the topics of star properties

Table 6. Number of students completing different formats of SFCI Version 2 ($n = 545$)

Form	n	M (%)	SD (%)	Skewness	Kurtosis	Cronbach's α
2a-1	92	51.3	18.5	-0.143	-0.569	0.657
2a-2	91	53.0	21.6	-0.524	-0.650	0.755
2b-1	92	52.5	22.9	0.056	-0.386	0.593
2b-2	91	58.9	23.4	-0.136	-0.679	0.588
2b-3	90	54.6	22.1	-0.603	0.033	0.516
2b-4	89	59.2	26.3	-0.175	-0.617	0.640
All forms combined	545	54.9	22.7	-0.166	-0.420	—

and formation. Such changes served to increase the reliability and validity of the instrument.

Additionally, we endeavored to increase reliability and validity by adding three new questions that each compared the values of a property (surface temperature, luminosity, or diameter) between the Sun, a red giant, or a white dwarf. These were designed to investigate students' understanding of the nature of the Sun relative to other stars. The addition of these three questions was prompted by two issues: participating instructors' emphasis on how the Sun compares with other stars and participants' explanatory statements (in Version 2b and earlier in Version 1 interviews) that included phrases such as 'the Sun is an average star'. The resulting Version 2.5 contained 26 content questions, all in multiple-choice format. At this time, the name 'Star Properties Concept Inventory (SPCI)', was adopted to better reflect the content of the instrument.

In order to gather evidence about the content validity of the SPCI Version 2.5, we asked a panel of experts to review the instrument. Reviewers were recruited through the AstroLrner electronic mailing list (a virtual community of astronomy educators; Slater, 2010); volunteers were asked to both answer the questions and provide comments about the questions and instrument as a whole. A total of 26 instructors returned their responses and comments. Nineteen of the reviewers have doctorates, and 6 have master's degrees. Twelve of the reviewers' degrees are in astronomy or astrophysics, 7 are in physics, and 3 are in other sciences (with 4 in non-science fields). All were current instructors for ASTRO 101-type courses. Twenty-four of the reviewers teach at 2-year or 4-year colleges or universities, and their teaching experience ranges from 1 to 30 years. Scores were calculated for the reviewers as a way of double checking the instrument's answer key and to verify the panel's expertise in the content area (especially for those members who do not have science backgrounds). The average score on the 26 questions included in Version 2.5 was 24.6, or 94.6%. Any incorrect answers were compared with the reviewer's comments; in all cases the reviewer indicated some indecision between two responses and had selected the incorrect one.

Results: Expert review

The expert review provided a great deal of feedback on each question. These ideas were reviewed individually, and the collections of feedback for each question were considered as a whole. Changes that were made based upon the feedback included alterations to some of the vocabulary used (described further below), clarification of questions that inadvertently contained multiple ideas or ambiguous questions, and corrections of typographical errors.

Perhaps the biggest change came about through the comparison of correct scientific vocabulary relative to students' natural language. An example of this arose when discussing the luminosity of a star. Luminosity is a quantity that describes the amount of energy emitted by the star per unit time, typically measured in Watts or some comparable unit. This is a term that most students would not be expected to

know prior to instruction, and so it had not been used in Versions 1 or 2. However, the substitute language that had been used (such as brightness), although more ‘student-friendly’, was not accurate and could have been confusing for students after instruction. As a compromise, the phrase ‘energy output (luminosity)’ was used in all questions dealing with this concept. In another case, expert reviewers objected to the use of the color ‘white’ as an option in some questions, as white is actually a combination of all colors of light and is not scientifically accurate. However, it was a very common response from students when they were asked open-ended questions, indicating a lack of understanding about the colors of stars. ‘White’ was therefore retained as a distracter because of its prominence as an alternative conception.

Version 3

Methods

Based upon the results of the expert review, changes were made to create Version 3 of the SPCI (Appendix). In addition to changes for clarity on 11 questions as described above, two questions were reworded to avoid leading participants to a correct answer on other items. Three questions were determined to be ambiguous and were removed. Six questions remained unaltered.

After all changes were made, Version 3 contained 23 content and 2 demographic questions (gender and prior astronomy coursework). A single format was used, so that each participant had the same CI with all 25 questions in multiple-choice format. Students responded to the survey on Pearson NCS[®] scannable answer forms to facilitate rapid collection of response data (previously, participants had recorded their answers directly on the CI).

Version 3 was administered as a pre-test on the first day of class during the Fall 2005 semester, and as a post-test during the final 2 weeks of the semester. Five sections of ASTRO 101 and 3 of ES 101 participated in the pre-test; 1 of the 3 ES 101 sections did not participate in the post-test because of scheduling problems.

In Versions 1 and 2 of the CI, all responses were used to calculate the statistics reported. For the analysis of Version 3, an effort was made to remove any data which might have been incomplete or which indicated that the student did not take the survey seriously. For example, if a student left more than two questions unanswered, his or her data were removed (this was done to avoid bias on early questions as a result of students not finishing the entire CI). Likewise, answer sheets which had obviously been carelessly completed—such as by entering ‘Bs’ for a disproportionately large number of the 25 questions or providing answers that were not options for certain questions—were removed, as it was deemed unlikely that the student made an honest effort to answer the questions. From this point forward, we performed the analysis on only the ‘reduced’ groups. Finally, we further looked for matched pairs, including only those students who took both tests. The participant breakdown by class is described in Table 7, with the numbers of ‘raw’ (all participants), ‘reduced’ (where individual student results were removed for the reasons just described), and matched groups presented.

Table 7. Number of participants for SPCI Version 3, by course type

Administration	<i>n</i>	
	ASTRO 101	ES 101
Pre-test		
Raw	690	411
Reduced	586	334
Post-test		
Raw	489	155
Reduced	417	113
Matched	334	83

Results: Instrument analysis

Reliability of the instrument was assessed by calculating Cronbach's α . Using the $n = 417$ students who completed the SPCI at both times, we found that the pre-test administration yielded an α of 0.470 for the 23 content items of the SPCI Version 3. This value is less than what is typically considered acceptable; however, coefficient α values are sensitive to the homogeneity of the test-taking population (Thompson, 2003). This low α value may indicate a considerable amount of guessing, which raises some concern over the attraction of distracters and the instrument's ability to identify students' alternative conceptions. However, the overall instrument pre-test score (31%) is higher than what would be expected from guessing (22%, as determined from the proportion of 4- and 5-choice questions), indicating that the distracters are working.

The post-test α was 0.763, which is considered an acceptable level of reliability (George & Mallery, 2009). In other words, reliability was greater on the post-test because with some knowledge of stars gained from their course, student knowledge variance was adequately ascertained by the SPCI. The acceptable post-test coefficient α gave us some confidence that the SPCI was a consistent measure of students' understanding about star properties.

Item Difficulty and Item Discrimination

The item difficulty of a question, p , is defined simply as the proportion of students who got the answer correct; thus items with lower difficulty values are 'harder' questions than those with higher difficulty values (Allen & Yen, 1979). The difficulty of each content item was calculated for both pre-test and post-test, ASTRO 101 participants only. Pre-test item difficulty values ranged from 0.067 to 0.814, while the post-test difficulty values ranged from 0.161 to 0.918. Allen and Yen suggest that a range of item difficulties, with an average of about 0.5, is ideal. Although some of the items from SPCI Version 3 have low item difficulty, the overall average (0.51 at post-test) implies that the distribution of item difficulty is reasonable. No items were marked

for removal because of too-low or too-high item difficulty. This decision was supported by also evaluating the item discrimination (described below) and the nature of the highest-selected distracter.

Item discrimination indicates ‘the degree to which responses to one item are related to responses to the other items on the test’ (Allen & Yen, 1979, p. 120). The measure of item discrimination used in this analysis is the item/total-test-score point biserial correlation, r_{iX} . As with item difficulty, a wide range of point biserial values is desirable to ensure that performance on the test is indicative of the varying knowledge levels of all students. For the post-test administration to ASTRO 101 students, the range of point biserial values was 0.110–0.551. A suggested minimum value of point biserial is 0.2 (Ding & Beichner, 2009). Two of our questions fell below this value—items 3 and 13. Our analysis using item response theory provided further evidence that these two questions may be problematic (Wallace & Bailey, 2010), and they have been flagged for revision or elimination in an future iteration of the instrument.

Ideally the item difficulty should be higher (i.e. the questions are answered correctly by more students) on the post-test administration than the pre-test. For the ASTRO 101 students who completed the SPCI Version 3, this is true for all but three of the questions, whose item difficulty decreased. Looking only at the item difficulty, however, is insufficient to determine if the question should be removed from the test; item discrimination and the most common distracter should also be reviewed. The three questions were reviewed individually; we decided to keep two of the questions because of the small difficulty decrease and high point-biserial, as well as the questions’ ability to identify a student’s understanding of the details of stellar processes. The third question whose difficulty fell is item 13, for which the low item discrimination is also a concern, as described above.

Results: Measured learning gains

The mean pre-test score for all participants was $M = 7.12$ (out of 23 possible), $SD = 2.78$, while the mean post-test score was $M = 10.81$, $SD = 4.23$ ($n = 417$). Once again, skewness and kurtosis absolute values fell below the accepted value of 1 (George & Mallery, 2009), indicating normal distributions. The results of a paired-samples t -test of all participants show that the post-test scores are significantly higher than the pre-test scores, $t(416) = -17.429$, $p < 0.001$, Cohen’s $d = 1.03$ (large effect). Looking only at the ASTRO 101 scores yields similar results. Pre-test scores yield a $M = 7.09$, $SD = 2.73$, while post-test scores yield $M = 11.84$, $SD = 3.87$ ($n = 334$). A paired-samples t -test of the ASTRO 101 participants show that the post-test scores are significantly higher than the pre-test scores, $t(333) = -21.679$, $p < 0.001$, Cohen’s $d = 1.42$ (large effect).

Although it was originally our intention to use the ES 101 as a comparison group, the large discrepancies in group size caused certain assumptions, such as that of equal variance, to not be met. Further investigation of this issue would require the use of non-parametric tests, which are beyond the scope of the study (i.e. group comparisons would not address our research questions).

Thematic Clusters

One of the important design considerations for the SPCI, as for CIs in general, was the need for multiple questions on any given topic (Nunnally, 1978). Additionally, ‘because items are random samples of information about a construct from a universe of information defining the construct, it is logical that more samples of the information represent the universe more thoroughly’ (Osterlind, 2010, p. 143). In order to fully explore the construct of ‘what is a star’, multiple questions from three major content themes were required within the SPCI: nuclear fusion as the defining characteristic of stars, with 5 questions; the process of star formation, with 5 questions; and star properties, with 13 questions. The intent here was not to develop formal subscales, but rather to fully explore the construct of interest.

The results for the different themes are generally the same as the overall scores, showing significant increases over instruction for the ASTRO 101 participants (Figure 2). This pattern of results on the three thematic clusters demonstrates that the items are working together as a single instrument in the way that was intended. If one of the themes, for example, had demonstrated a decrease in scores, those questions would have been re-evaluated on an individual basis to determine whether they needed to be rewritten (because they are too difficult or misleading) or removed (because they are beyond the scope of typical ASTRO 101 instruction). Again, these themes are presented only in an attempt to ensure full exploration of the construct of stars. They are not intended as separate subscales and should not be interpreted as such during the SPCI’s use for research or instruction.

Discussion

We focus our discussion on two main areas: the purpose and research questions that guided the study, followed by the implications of this instrument’s creation for the ASTRO 101 course. Recall that the study’s purpose was to develop a CI around the topic of star properties and formation. The development was guided by two research questions: (1) To what degree is the instrument valid and reliable, and

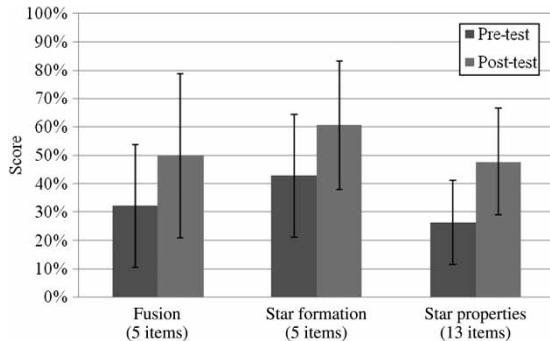


Figure 2. Comparison of mean pre-test and post-test scores on three main themes. Differences between the ASTRO 101 pre-test and post-test means are significant at the $p < 0.01$ level for all three themes

what evidence supports this claim? (2) How do ASTRO 101 students compare on the instrument pre-test to post-test?

Revisiting the Research Questions

Knowledge of student beliefs about stars properties and formation (Bailey et al., 2009) was used to inform the development of the SPCI through an iterative process (Figure 1). The SPCI Version 3 (Appendix) is a 23 multiple-choice question instrument intended to measure student learning gains over instruction, thereby providing a measure of instructional effectiveness. The SPCI uses students' natural language, avoids astronomical jargon wherever possible, and contains distracters that are based on known difficulties. This instrument, administered to ASTRO 101 students in Fall 2005, has been shown to be able to identify student learning gains over the period in which the ASTRO 101 students had instruction on these topics.

We also looked at how the ASTRO 101 students compared on the instrument pre-test with post-test. Although statistically significant learning gains were observed over the instructional semester, the mean post-test score (51%) for the ASTRO 101 course might be considered an unsatisfying result for instructors. Such gains are concurrent with other studies that show students only score about 50–60% correct after lecture (Bardar, 2006; Prather et al., 2004). This mean may cause instructors to reconsider their overall course design or pedagogical strategies. Suggestions for increasing instructional effectiveness are beyond the scope of this study and can be found elsewhere (see, e.g. Hake, 1998; Prather, Rudolph, & Brissenden, 2009).

Analyzing Instructional Effectiveness with CIs

The Force Concept Inventory (FCI; Halloun & Hestenes, 1985) has been heavily adopted by instructors of college-level physics courses across the country as a way of gauging the effectiveness of their instructional strategies. In a meta-analysis of the FCI and the earlier Mechanics Baseline test (Hestenes & Wells, 1992) used in this manner, Hake (1998) found that the 48 interactive engagement courses achieved a significantly higher gain over instruction than the traditional courses. Based upon results such as these, Hake (2005) calls for all disciplines in higher education to consider the development and use of multiple-choice instruments, in the vein of the FCI, to measure instructional effectiveness and to develop interactive engagement methods suitable for the discipline.

What is needed for astronomy, then, is a suite of CIs that more closely align with the topics that are typically included in the ASTRO 101 course. In addition to the LPCI (Lindell, 2001; Lindell & Olsen, 2002), LSCI (Bardar et al., 2005, 2007), and the SPCI (described here), Hornstein et al. have developed the Solar System CI, currently undergoing validation studies (Hornstein, Duncan, & Collaboration of Astronomy Teaching Scholars, 2009; Hornstein et al., 2010). An early version of the Cosmology Subject Inventory (McLin & Cominsky, 2008) is currently under revision and expansion. The Greenhouse Effect Concept Inventory (Keller, 2006; Keller, Prather, &

Slater, 2006) may be of interest to some astronomy instructors or to instructors of other introductory courses in planetary or earth science for non-science majors.

The use of CIs to determine the effectiveness of teaching on a focussed topic can be interpreted as an indication of teaching effectiveness as a whole, and this effectiveness has repeatedly been shown to be lacking (e.g. Hake, 1998; Prather, Rudolph, Brissenden et al., 2009). Widespread use of—and research using—these CIs in ASTRO 101 and similar classes is needed in order to better inform the development of interactive engagement methods for these courses for non-science majors. CIs for additional topics should be developed in conjunction with research on student understanding, so that the instruments accurately represent the range of student ideas about the topics both before and after instruction.

One might ask how many CIs are enough – a fair question. ASTRO 101, typically a survey of the whole universe, by definition includes a large number of topics. While it is not appropriate to use several CIs in a single semester, having a variety of options can allow instructors to make targeted improvements over time, or allow curriculum developers to test their materials with validated instruments. Additionally, the creation of more topical CIs can then contribute to new, whole-universe instruments, developed with individual instructors' needs in mind (see, e.g. the Geoscience Concept Inventory; Libarkin & Anderson, 2005).

Future Research and Conclusions

There are several avenues of further research that can be followed to build upon the results of this project. Currently underway is a study that investigates the reliability and validity of the SPCI when administered across multiple instructors and different types of institutions. An analysis of the data will help determine whether the students at the original testing institution were, in fact, reasonably representative of other ASTRO 101 students both before and after instruction, as well as the generalizability of the SPCI across a diverse group of students. Preliminary results suggest that the SPCI's properties remain stable over this larger, nation-wide sample, though detailed analysis is ongoing at the time of this writing.

The data used in this study have also been analyzed using item response theory (Wallace & Bailey, 2010). Many of the findings here have been corroborated through the IRT analysis, such as the need to revisit items 3 and 13 for possible revision or elimination. The national study data will also be analyzed through both item response theory and CTT to further understand the properties of the SPCI items.

Once a greater level of generalizability is determined, data may be analyzed to compare courses using interactive engagement and traditional lecture-based instruction (Hake, 1998; Prather, Rudolph, Brissenden et al., 2009). Given the prominence of stars in most ASTRO 101 courses (Slater et al., 2001), the use of the SPCI to investigate instructional effectiveness could become widespread across the astronomy community. If it is determined that, as has been shown about the teaching and learning of light (Prather, Rudolph, Brissenden et al., 2009), traditional astronomy courses are ineffective at promoting deep conceptual understanding about stars, then the

development of additional interactive engagement methods or focussed curricular materials designed to intellectually engage students could follow. Such improvements to courses benefit students at large, but it may be particularly beneficial for the pre-service teachers who take such courses (Lawrenz, Huffman, & Appeldoorn, 2005) to experience science taught through engaging pedagogical strategies.

While this study confirmed that the SPCI can reliably measure change in student understanding about stars over time, it also may provide the ability to identify those cases where instruction appears to have little effect on student understanding. By looking at post-test results, researchers can further investigate students' alternative conceptions that may be resistant to change. This can allow us to further target research, curriculum, and instruction in order to maximize conceptual change.

Another possibility is the expansion of this study to related topics. This research study was deliberately focussed on only a few basic properties of stars and the process of star formation. The breadth of student ideas about stellar evolution was not addressed. The investigation of these ideas could lead to the creation of an additional CI on stellar evolution, or one in which items from the SPCI and new questions are combined to make another instrument, perhaps one that can be customized to align more closely with an individual classroom, as described above (Libarkin & Anderson, 2005).

Overall, this research study has provided a way to measure what ASTRO 101 students understand about stars and star formation. The SPCI has been created and evaluated, and has shown to be a valid and reliable instrument that can measure learning on the topics of star properties and formation over an instructional period. The astronomical community is poised to use inventories such as these to examine instructional effectiveness as it moves toward learner-centered instruction.

Acknowledgements

This study was completed as part of the first author's doctoral dissertation while all authors were at the The University of Arizona. The authors would like to thank all of the instructors and students involved in the study. Doug Lombardi and Alice Corkill provided tremendous help with the statistical analysis. Doug Lombardi, Tamera Hanken, Tim Bungum, Iria Gonzalez, Kate Wintrol, Hasan Deniz and the anonymous reviewers provided valuable feedback on the manuscript at various times, for which we are greatly appreciative.

References

- Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press, Inc.
- Anderson, D.L., Fisher, K.M., & Norman, G.J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39(10), 952-978. doi: 10.1002/tea.10053
- Bailey, J.M. (2006). *Development of a concept inventory to assess students' understanding and reasoning difficulties about the properties and formation of stars* (PhD dissertation). The University of Arizona, Tucson, AZ, USA.

- Bailey, J.M. (2009). Concept inventories for ASTRO 101. *The Physics Teacher*, 47(7), 439–441. doi: 10.1119/1.3225503
- Bailey, J.M., Prather, E.E., Johnson, B., & Slater, T.F. (2009). College students' preinstructional ideas about stars and star formation. *Astronomy Education Review*, 8(1), 010110–010117. doi: 10.3847/AER2009038
- Bailey, J.M., & Slater, T.F. (2003). A review of astronomy education research. *Astronomy Education Review*, 2(2), 20–45. doi: 10.3847/AER2003015
- Bailey, J.M., Slater, S.J., & Slater, T.F. (2011). *Conducting astronomy education research: A primer*. New York, NY: W.H. Freeman and Company.
- Bardar, E.M. (2006). *Development and analysis of spectroscopic learning tools and the light and spectroscopy concept inventory for introductory college astronomy* (PhD dissertation, Boston University, United States – Massachusetts). Retrieved September 25, 2006, from Dissertations & Theses: Full Text. (Publication No. AAT 3214908).
- Bardar, E.M., Prather, E.E., Brecher, K., & Slater, T.F. (2005). The need for a Light and Spectroscopy Concept Inventory for assessing innovations in introductory astronomy survey courses. *Astronomy Education Review*, 4(2), 20–27. doi: 10.3847/AER2005018
- Bardar, E.M., Prather, E.E., Brecher, K., & Slater, T.F. (2007). Development and validation of the Light and Spectroscopy Concept Inventory. *Astronomy Education Review*, 5(2), 103–113. doi: 10.3847/AER2006020
- Bennett, J., Donahue, M., Schneider, N., & Voit, M. (2004). *The cosmic perspective* (3rd ed.). San Francisco, CA: Addison Wesley.
- Brog, E., Sabers, D., Prather, E.E., Deming, G.L., Hufnagel, B., & Slater, T.F. (2007). Analysis of the Astronomy Diagnostic Test. *Astronomy Education Review*, 6(1), 25–42. doi: 10.3847/AER2007003
- Bruning, D. (2002). 2002 survey of introductory astronomy textbooks. *Astronomy Education Review*, 1(1), 92–113. doi: 10.3847/AER2001008
- Bruning, D. (2006a). 2006 survey of introductory astronomy textbooks. *Astronomy Education Review*, 4(2), 54–90. doi: 10.3847/AER2005020
- Bruning, D. (2006b). Survey of introductory astronomy textbooks: An update. *Astronomy Education Review*, 5(2), 182–216. doi: 10.3847/AER2006026
- Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32(7), 939–961. doi: 10.1080/09500690902890130
- Deming, G.L. (2002). Results from the Astronomy Diagnostic Test national project. *Astronomy Education Review*, 1(1), 52–57. doi: 10.3847/AER2001005
- Deming, G.L., & Hufnagel, B. (2001). Who's taking ASTRO 101? *The Physics Teacher*, 39(6), 368–369. doi: 10.1119/1.1407134
- Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics – Physics Education Research*, 5(2), 020103. doi: 10.1103/PhysRevSTPER.5.020103
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fraknoi, A. (2001). Enrollments in astronomy 101 courses: An update. *Astronomy Education Review*, 1(1), 121–123. doi: 10.3847/AER2001011
- George, D., & Mallery, P. (2009). *SPSS for Windows step by step: A simple guide and reference, 16.0 update* (9th ed.). Boston, MA: Pearson Education.
- Glaser, B.G., & Strauss, A.L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Piscataway, NJ: Aldine Transaction.
- Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. doi: 10.1119/1.18809

- Hake, R.R. (2005). The physics education reform effort: A possible model for higher education? *The National Teaching & Learning Forum*, 15(1). Retrieved from <http://www.ntlf.com/FTPSite/issues/v15n1/physics.htm>
- Halloun, I.A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53(11), 1043–1055. doi: 10.1119/1.14030
- Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, 30(3), 159–166. doi: 10.1119/1.2343498
- Hornstein, S.D., Duncan, D., & Collaboration of Astronomy Teaching Scholars. (2009, January). *Development of a solar system concept inventory*. Paper presented at the 213th Meeting of the American Astronomical Society, Long Beach, CA.
- Hornstein, S.D., Prather, E.E., English, T.R., Desch, S.M., Keller, J.M., & Collaboration of Astronomy Teaching Scholars. (2010, January). *Continued development of the Solar System Concept Inventory*. Paper presented at the 215th Meeting of the American Astronomical Society, Washington, DC.
- Hufnagel, B. (2002). Development of the Astronomy Diagnostic Test. *Astronomy Education Review*, 1(1), 47–51. doi: 10.3847/AER2001004
- Hufnagel, B., Slater, T.F., Deming, G.L., Adams, J.P., Adrien, R.L., Brick, C., & Zeilik, M. (2000). Pre-course results from the Astronomy Diagnostic Test. *Publications of the Astronomical Society of Australia*, 17(2), 152–155.
- Kaufmann III, W.J., & Freedman, R.A. (1999). *Universe* (5th ed.). New York, NY: W.H. Freeman and Company.
- Keller, J.M. (2006). *Part I. Development of a concept inventory addressing students' beliefs and reasoning difficulties regarding the greenhouse effect, Part II. Distribution of chlorine measured by the Mars Odyssey Gamma Ray Spectrometer* (PhD dissertation, The University of Arizona, United States – Arizona). Retrieved February 25, 2007, from Dissertations & Theses: Full Text. (Publication No. AAT 3237466).
- Keller, J.M., Prather, E.E., & Slater, T.F. (2006, January). *Probing student understanding of the atmospheric greenhouse effect*. Paper presented at the 2006 Winter Meeting of the American Association of Physics Teachers, Anchorage, AK.
- Lawrenz, F., Huffman, D., & Appeldoorn, K. (2005). Enhancing the instructional environment: Optimal learning in introductory science. *Journal of College Science Teaching*, 34(7), 40–44.
- Lelliott, A., & Rollnick, M. (2010). Big ideas: A review of astronomy education research 1974–2008. *International Journal of Science Education*, 32(13), 1771–1799. doi: 10.1080/09500690903214546
- Libarkin, J.C., & Anderson, S.W. (2005). Assessment of learning in entry-level geoscience courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, 53(4), 394–401.
- Lindell, R.S. (2001). *Enhancing college students' understanding of lunar phases* (Ph.D. dissertation, The University of Nebraska - Lincoln, United States – Nebraska). Retrieved October 5, 2002, from Dissertations & Theses: Full Text. (Publication No. AAT 3022646).
- Lindell, R.S., & Olsen, J.P. (2002, August). *Developing the Lunar Phases Concept Inventory*. Paper presented at the 125th National Meeting of the American Association of Physics Teachers (Physics Education Research Conference), Boise, ID.
- Lindell, R.S., Peak, E., & Foster, T.M. (2007). Are they all created equal? A comparison of different concept inventory development methodologies. In L. McCullough, L. Hsu, & P. Heron (Eds.), *American Institute of Physics Conference Proceedings* (Vol. 883, pp. 14–17). Melville, NY: American Institute of Physics.
- LoPresto, M.C., & Murrell, S.R. (2009). Using the Star Properties Concept Inventory to compare instruction with lecture tutorials to traditional lectures. *Astronomy Education Review*, 8(1), 010105–010105. doi: 10.3847/AER2009014

- McLin, K.M., & Cominsky, L.R. (2008, January). *Teaching in-service and pre-service teachers modern cosmology, part I: A concept inventory*. Paper presented at the 211th Meeting of the American Astronomical Society, Austin, TX.
- Nunnally, J. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
- Osterlind, S.J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal* (2nd ed.). Boston, MA: Pearson.
- Partridge, B., & Greenstein, G. (2003). Goals for “Astro 101”: Report on workshops for department leaders. *Astronomy Education Review*, 2(2), 46–89. doi: 10.3847/AER2003016
- Prather, E.E. (2000). *An investigation into what students think and how they learn about ionizing radiation and radioactivity* (PhD dissertation, The University of Maine, United States – Maine). Retrieved from <http://proquest.umi.com/pqdweb?did=731957451&Fmt=7&clientId=17675&RQT=309&VName=PQD>
- Prather, E.E., & Harrington, R.R. (2001). Student understanding of ionizing radiation and radioactivity. *Journal of College Science Teaching*, 31(2), 89–93.
- Prather, E.E., Rudolph, A.L., & Brissenden, G. (2009). Teaching and learning astronomy in the 21st century. *Physics Today*, 62(10), 41–47. doi: 10.1063/1.3248478
- Prather, E.E., Rudolph, A.L., Brissenden, G., & Schlingman, W.M. (2009). A national study assessing the teaching and learning of astronomy. Part I. The effect of interactive instruction. *American Journal of Physics*, 77(4), 320–330. doi: 10.1119/1.3065023
- Prather, E.E., Slater, T.F., Adams, J.P., Bailey, J.M., Jones, L.V., & Dostal, J.A. (2004). Research on a lecture-tutorial approach to teaching introductory astronomy for non-science majors. *Astronomy Education Review*, 3(2), 122–136. doi: 10.3847/AER2004019
- Rudolph, A.L., Prather, E.E., Brissenden, G., Consiglio, D., & Gonzaga, V. (2010). A national study assessing the teaching and learning of introductory astronomy part II: The connection between student demographics and learning. *Astronomy Education Review*, 9(1), 010107–010115. doi: 10.3847/AER0009068
- Sadler, P.M., Coyle, H., Miller, J.L., Cook-Smith, N., Dussault, M., & Gould, R.R. (2009). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 national science standards. *Astronomy Education Review*, 8(1), 010111–010126. doi: 10.3847/AER2009024
- Seidman, I. (1998). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (2nd ed.). New York, NY: Teachers College Press.
- Slater, T.F. (2010). The AstroLrner e-community: A 10 year retrospective. *Astronomy Education Review*, 9(1), 010601–010604. doi: 10.3847/AER2009055
- Slater, T.F., Adams, J.P., Brissenden, G., & Duncan, D. (2001). What topics are taught in introductory astronomy courses? *The Physics Teacher*, 39(1), 52–55. doi: 10.1119/1.1343435
- Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability* (pp. 3–30). Thousand Oaks, CA: SAGE Publications.
- Treagust, D.F. (1986). Evaluating students’ misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, 16, 199–207. doi: 10.1007/BF02356835
- Treagust, D.F. (1988). Development and use of diagnostic tests to evaluate students’ misconceptions in science. *International Journal of Science Education*, 10(2), 159–169. doi: 10.1080/0950069880100204
- Wallace, C.S., & Bailey, J.M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1), 010116. doi: 10.3847/AER2010024
- Zeilik, M. (2002). *Astronomy: The evolving universe* (9th ed.). New York, NY: Cambridge University Press.
- Zeilik, M. (2003). Birth of the astronomy diagnostic test: Prototest evolution. *Astronomy Education Review*, 1(2), 46–52. doi: 10.3847/AER2002005

Appendix. Star Properties Concept Inventory, Version 3

Instructions: Indicate the best answer for each question on the separate answer sheet. There are 25 questions total.

1. When a star is first formed, it is made mostly of which of the following?
 - a. oxygen
 - b. nitrogen
 - c. carbon
 - d. helium
 - e. hydrogen

2. Which of the following causes a star's interior temperature to increase during its formation?
 - a. Nuclear fusion causes gravitational collapse, which generates heat.
 - b. Heat is generated when the star's gravity contracts.
 - c. Gravitational collapse involves the generation of heat from chemical reactions.
 - d. During collapse, gravitational potential energy decreases while its temperature increases.

3. The Sun's surface temperature is
 - a. near the high end of the range of surface temperatures.
 - b. near the low end of the range of surface temperatures.
 - c. near the middle of the range of surface temperatures.
 - d. about the same as the surface temperatures of all other stars.

4. During the majority of a star's existence, in which part of a star is its energy produced?
 - a. radiative layer
 - b. nucleosphere
 - c. core
 - d. throughout the star
 - e. on the surface

5. Star Y has twice the mass of star X. How will star X use up its fuel compared to star Y?
 - a. Star X will use up its fuel more than two times slower than star Y.
 - b. Star X will use up its fuel two times slower than star Y.
 - c. Star X will use up its fuel at the same rate as star Y.
 - d. Star X will use up its fuel two times faster than star Y.
 - e. Star X will use up its fuel more than two times faster than star Y.

6. The force that dominates the formation of a star is
 - a. static electricity.
 - b. gravity.
 - c. magnetism.

- d. pressure.
 - e. nuclear fusion.
7. The hottest stars are what color?
- a. red
 - b. white
 - c. blue
 - d. all stars have the same color regardless of their temperature
 - e. all stars have the same temperature regardless of their color
8. Why don't most stars collapse in on themselves under gravity's influence?
- a. Material churning in and out of the center of the star balances gravity.
 - b. The internal structure of the star holds the surface out and keeps it from collapsing.
 - c. Gravity from planets orbiting the star pulls outward on the star's material.
 - d. The force from particles ejected outward from the center of the star balances gravity.
 - e. Gas pressure caused by energy created in the star pushes outward to balance gravity.
9. Which of the following objects has the highest surface temperature?
- a. a typical red giant
 - b. a typical white dwarf
 - c. the Sun
 - d. These objects could have the same temperature.
10. Star C has a lifetime of 50 million years, while star D has a lifetime of only 10 million years. What can you say about the masses of these stars?
- a. Star C has the greater mass.
 - b. Star D has the greater mass.
 - c. Stars C and D have about the same mass.
 - d. There is not enough information given to answer this question.
11. What is the name given to a star as it is initially forming?
- a. protostar
 - b. nebula
 - c. supernova
 - d. star cluster
 - e. white dwarf
12. How does the Sun produce the energy that heats our planet?
- a. The gases inside the Sun are burning and producing energy.
 - b. Atoms are combined into heavier atoms, giving off energy.
 - c. Gas inside the Sun heats up when compressed, giving off energy.
 - d. Atoms are broken apart into lighter atoms, giving off energy.
 - e. The core of the Sun has radioactive atoms that give off energy as they decay.

13. Which of the following objects is most massive: a red giant, a white dwarf, or the Sun?
 - a. A red giant is always the most massive.
 - b. A white dwarf is always the most massive.
 - c. The Sun is the most massive.
 - d. These objects could have the same mass.
14. Stars begin life as
 - a. a piece off of a star or planet.
 - b. a white dwarf.
 - c. matter in Earth's atmosphere.
 - d. a black hole.
 - e. a cloud of gas and dust.
15. What is a star?
 - a. a ball of gas that reflects light from another energy source
 - b. a bright point of light visible in Earth's atmosphere
 - c. a hot ball of gas that produces energy by burning gases
 - d. a hot ball of gas that produces energy by combining atoms into heavier atoms
 - e. a hot ball of gas that produces energy by breaking apart atoms into lighter atoms
16. If a red star and a blue star have the same size (diameter) and are at the same distance from Earth, which one will appear brighter?
 - a. the red star
 - b. the blue star
 - c. Both stars will look the same.
 - d. There is not enough information given to answer this question.
17. How is the lifetime of a star related to its mass?
 - a. More massive stars live considerably longer lives than less massive stars.
 - b. More massive stars live considerably shorter lives than less massive stars.
 - c. More massive stars live slightly shorter lives than less massive stars.
 - d. More massive stars live slightly longer lives than less massive stars.
 - e. All stars have the same lifetimes regardless of mass.
18. Which of the following objects has the greatest actual brightness (luminosity): a red giant, a white dwarf, or the Sun?
 - a. A red giant always has the greatest actual brightness (luminosity).
 - b. A white dwarf always has the greatest actual brightness (luminosity).
 - c. The Sun always has the greatest actual brightness (luminosity).
 - d. These objects could have the same actual brightness (luminosity).
19. The light from stars that we see on Earth results from
 - a. reflection of sunlight.
 - b. chemical reactions inside the stars.
 - c. nuclear reactions inside the stars.

- d. burning of gases inside the stars.
 - e. burning on the surfaces of the stars.
20. How would you rank the surface temperatures of red, white, and blue stars?
Hottest → coldest
- a. white > blue > red
 - b. white > red > blue
 - c. red > blue > white
 - d. blue > white > red
 - e. blue > blue > white
21. Which of the following would most likely give off the most energy?
- a. a red star half the size (diameter) of the Sun
 - b. a red star 10 times the size (diameter) of the Sun
 - c. a blue star half the size (diameter) of the Sun
 - d. a blue star 10 times the size (diameter) of the Sun
22. Star P has three times the mass of star Q. How will the lifetime of Star Q compare to the lifetime of star P?
- a. Star Q's lifetime will be less than one-third as long as that of star P.
 - b. Star Q's lifetime will be one-third as long as that of star P.
 - c. Star Q's lifetime will be the same as that of star P.
 - d. Star Q's lifetime will be three times as long as that of star P.
 - e. Star Q's lifetime will be more than three times as long as that of star P.
23. Which of the following determines most characteristics and future events of a star's existence?
- a. surface temperature
 - b. size (diameter)
 - c. color
 - d. composition (type of atoms)
 - e. mass
24. What is your gender?
- a. Female
 - b. Male
25. Have you previously taken an astronomy course? (If you are taking this survey in an astronomy course, do NOT count it in your response.)
- a. Yes
 - b. No