The Development And Validation Of The Test Of Astronomy STandards (TOAST)

Stephanie J. Slater, Ph.D., CAPER Center for Astronomy & Physics Education Research, USA

ABSTRACT

The Test Of Astronomy STandards (TOAST) is a comprehensive assessment instrument designed to measure students' general astronomy content knowledge. Built upon the research embedded within a generation of astronomy assessments designed to measure single concepts, the TOAST is appropriate to measure across an entire astronomy course. The TOAST's scientific content represents a consensus of expert opinion about what students should know from three different groups: the American Association for the Advancement of Science, the National Research Council, and the American Astronomical Society. The TOAST's reliability and validity are established by results from Cronbach alpha and classical test theory analyses, a review for construct validity, testing for sensitivity to instruction, and numerous rounds of expert review. As such the TOAST can be considered a valuable tool for classroom instructors and discipline based education researchers in astronomy across a variety of learning environments.

Keywords: Astronomy Education; Assessment; TOAST

In science education and public outreach, whether we are teaching, evaluating programs, or conducting research, measuring student thinking is arguably the most important thing we do. Reliable data informs our instructional approaches, allows us to report to administrative and funding entities, and provides us with the data needed to move our research agenda forward. Unfortunately, measuring student thinking is a tricky business. Good measurement requires that we correctly perform multiple complex steps with little or no error. First, we have to specifically define the thing that we are interested in measuring, describing both the expected range and depth of knowledge to be assessed. Once we have done that, we have to create assessment tasks that are tightly aligned with that definition, making sure that our tasks do not miss any of the domain that we are trying to measure or go beyond the bounds of that domain. As we write these tasks, we have to take care that we are not accidently measuring some trait that is not of interest, and we must write these tasks such that we can easily make sense of the data that they produce. Completing any one of these steps without substantive error requires both time and skill. Completing the entire series of tasks in order to create valid and reliable measures of student thinking is very difficult.

For introductory astronomy courses, there is a particular need to have high quality assessments that can be used for both education research and course evaluation purposes. In response to this need, the astronomy education research community has invested a great deal of intellectual energy, expertise, and resources into constructing assessments of students learning. In addition to the broad content Astronomy Diagnostic Test 2 (ADT2) (Zeilik, 2002), these instruments measure student thinking related to a discrete concept from the astronomy domain. The list of single concept inventories in astronomy includes instruments related to students' knowledge of lunar phases (Lindell & Sommer, 2004), the greenhouse effect (Keller, 2006), light and spectra (Bardar, 2006), stars and their properties (Bailey, 2012), and gravity (Williamson, 2013). These instruments can be judged to reasonably measure their specific content domain. However, they are not reasonable measures of the construct that more often interests instructors, researchers, and administrators: students' general astronomy content knowledge. This is true simply because these instruments do not measure that construct.

In order to make a claim about students' general astronomy content knowledge, we have to sufficiently sample across that domain so that we can know that we are not measuring an anomaly. We must measure across the

domain, sampling those learnings that we believe are core to an understanding of the subject. We cannot measure less, and we have to be careful not to measure more. Further, in order to facilitate comparison of student thinking across different learning environments, we need an instrument that is easy to administer and interpret. Such an instrument has been missing from the astronomy education research toolkit.

The Test Of Astronomy STandards, or the TOAST, is an instrument developed to address this need. The TOAST is a third-generation assessment built upon more than twenty years of research in student cognition and assessment in astronomy, specifically designed to measure student learning across the breath of introductory astronomy courses. This paper provides a description of the theoretical need for a broad content assessment in astronomy, with an in-depth analysis of the danger of using single-concept inventories as measures of general astronomy content knowledge. The scholarship underpinning the TOAST, and the work done to construct and validate the TOAST for introductory astronomy survey courses settings, are also described.

BACKGROUND

The Necessity Of Broad Content Coverage In Assessment

If we want to understand or make claims about students' "general astronomy content knowledge," or about teaching and learning across the breath of an introductory astronomy course, we have to use assessments that carefully measure that construct. This idea is embedded in the educational landscape, as a common sense notion, and a basic tenet of assessment. This common sense notion can be observed playing out in common practice because we instinctively believe that student knowledge is uneven across any given field. Common experience, and perhaps our own experience, leads us to believe that a physics major that has, for instance, mastered kinematics, may at the same time have little or no grasp on electricity and magnetism. This can be true even if the student has been taught both concepts. Because of these inconsistencies, at the end of courses, we often ask students to demonstrate their mastery of the content via cumulative, rather than single topic, exams. When students graduate from their undergraduate programs, they are often asked to take a comprehensive exam like the GRE to demonstrate their mastery of the broad domain. We do not accept their course grade from a single class as evidence that they grasp the fundamental concepts in the field.

From a theoretical perspective, broad content assessment is a basic requirement because student and instructor factors vary across the science content. A full discussion of this phenomenon goes beyond the scope of this paper, but it is helpful to consider at least one factor related to both students and instructors. When we consider students, we instinctively and correctly understand that each individual has varying strengths and weaknesses, and that those characteristics may interact with the content in unpredictable ways. To take an example from another scientific field, in a chemistry course, students are normally expected to learn a variety of skills, including how to complete stoichiometry problems and to predict the polarity of molecules. While both of these tasks require spatial reasoning abilities, stoichiometry problems draw on a student's ability to disembed relevant constructs (Bodner & McMillen, 1986), whereas predicting polarity leans on a student's ability to rotate figures (Wu & Shah, 2004). Very few people have equal levels of ability in these two areas. Therefore, even in the presence of consistent instruction, students will engage in these concepts in ways that are as variable as their innate abilities.

To complicate matters, instruction is rarely consistent over a course as instructor variables are also heterogeneous across the domain. Although there are many ways to frame instructional inconsistency, it may be easiest to consider the concept of pedagogical content knowledge (Shulman, 1986). Although there are many ways to define pedagogical content knowledge (PCK), for the purposes of this discussion, PCK might be thought of as the pragmatic knowledge that teachers use to guide their actions in science classrooms (Rowan et al., 2001), or as the fusion of an instructor's understanding of pedagogy with their understanding of the scientific content. This holistic definition can be broken down into variables that are easier to contemplate. We might say that good PCK in any given science requires the following four proficiencies:

1. A solid knowledge of the content (i.e., facts and concepts).
2. An understanding of the structure of the scientific discipline.
3. Knowledge of common student learning difficulties related to specific scientific concepts.
4. A toolkit of teaching strategies that address students' learning needs.

Even in an introductory course, instructors' PCK varies across the breadth of the content for each of these variables. Given four variables, there is room for wide variation in instruction across a course (Baxter & Lederman, 1999). Taken together, student and instructor variables can combine to create uneven learning experiences across the breadth of a course, which a single topic concept inventory cannot possibly detect.

Broad content assessments are also needed to avoid, as much as possible, the improper influence of assessment on goals and instruction. A large body of research indicates that, given awareness of the content of an externally administered assessment, administrators and instructors will alter course objectives and content in an attempt to improve test scores (Corbett & Wilson, 1988). This is particularly true in cases where assessment scores will potentially be used to rate school and instructor quality (Madaus, 1988). In some cases, instructors will adopt instructional activities that closely mirror the types of questions they believe will appear on the assessment, and they will remove course material that is unlikely to be covered on the test (Smith & Rottenberg, 1989). A narrow assessment that does not address the core learnings of a domain stands poised to negatively influence instruction and return invalid indicators of student learning (Shepard, 1991).

The problems inherent in using a single-concept inventory also impact course evaluations and research studies. The desire to demonstrate quality teaching is a powerful motivator for instructors, such that they can unknowingly make important changes to their teaching. Known as the "teacher-expectancy effect," or the "Pygmalion Effect," the desire to believe that we are helping our students, and our need to manage our job and ego pressures, can lead us to, perhaps unknowingly, provide additional interventions related to test content (e.g., time on task, improved pedagogy, instructor enthusiasm) (Slater, Bailey, & Slater, 2012). As in the case of externally driven assessment, if the instrument covers the breadth of the course, the Pygmalion Effect can result in dramatic improvement to instruction across a whole term. However, if the assessment only covers a small portion of the content, we cannot assume that the Pygmalion Effect engenders any positive instructional change beyond that content unit. Instead, as described previously, we have to assume that the increased attention to teaching the tested material may come at a cost to other portions of the curriculum.

Given variation in student and teacher factors across the curriculum, and the tendency of assessments to skew classroom practice, single concept inventories are poised to provide us with misrepresentations of content learning, and more importantly, have the potential to alter instruction in ways that are unexpected and undesirable. If we are interested in overall astronomy knowledge, we cannot use a single concept inventory as a proxy, as that data may be both misleading and invalid.

The Absence Of An FCI Analog In Astronomy

Despite the body of research on the need to align measurement with claims, many in the physics community have considered student scores on the Force Concept Inventory (FCI), an assessment of the narrow concept of force, as a proxy for learning across first semester physics courses (Hestenes & Halloun, 1995; Coletta & Phillips, 2005). Giving this practice cursory examination, some researchers and practitioners have set forth that the FCI is indeed a reasonable proxy due to its universal coverage in physics courses (Bardar, 2006). While force is unquestionably a universal topic in introductory physics courses, this line of reasoning is superficial and has led to the fallacious belief that we might construct an analogous assessment instrument in astronomy. Arguments related to the validity of using the FCI as a proxy for general physics learning go beyond the scope of this paper, but it is relevant to question whether or not an FCI analog has been or can be created for astronomy. An analysis of the force concept and the FCI instrument indicates that force and the FCI possess three characteristics that may indeed make FCI scores a reasonable proxy for overall mastery of the material. Unfortunately, no topic in astronomy shares these characteristics, making it impossible to construct an instrument analogous to the FCI for astronomy.

A key characteristic that differentiates force from any of the content units in introductory astronomy is that force and a Newtonian framework are considered to be essential to an understanding of non-relativistic motion (Laws et. al., 1999). Beyond Newton's Laws, the remainder of a first semester physics course traditionally includes work, energy, momentum, gravity, and rotational motion. These concepts build upon a generative command of Newton's Laws. As such, it is reasonable to assume that a student who fails to master force will struggle with the material in the rest of the course, and at best, will only be able to develop a superficial understanding of the content.

In astronomy, there is no topic that is essential to an understanding of the rest of the material, and very little building of knowledge occurs. Instead, an introductory astronomy course is typically a series of fairly disconnected units. Celestial motion, stars, and light and spectra are all core to the domain of astronomy, yet few if any Astronomy 101 courses require students to use these concepts in subsequent units, in a generative way. As taught in introductory astronomy courses, a student's grasp of celestial motion does little help them learn cosmology, their mastery of stellar properties has little to do with their understanding of the solar system, and their conception of light and spectra does not impact their knowledge of lunar phases. The foundational, generative nature of the force concept is not shared by any of the concepts taught in an introductory astronomy course.

There are two other reasons that the field of astronomy education research is unlikely to construct an FCI analog. The first is that every student who enters a physics classroom has spent a lifetime developing a scientifically inaccurate model of force, which they have successfully used to manage everyday life. Indeed, evidence suggests that even infants have working models of physical phenomenon involving force and motion, and that they use these models to make sense of the world around them (Gopnik et. al., 1999). We use our "baby physics," or Aristotelian worldview, to survive the day, and our models are reinforced every time we successfully manage an object in flight or avoid running into objects. There is no similar topic in astronomy. While students certainly enter astronomy courses with preconceptions that they apply to the domain's content, these preconceptions do not form coherent models that are constantly used to negotiate real-life. There is no evidence that babies have working models of the greenhouse effect or lunar phases, or that they use their knowledge of these topics to negotiate the real-world. For most students, the first time they cognitively interact with stellar properties, or light and spectra, is when they enter their college astronomy courses. They have no robust misconceptions or models of this content, which we can see in students' pre-test data on the Light and Spectra Concept Inventory (LSCI) (Bardar, 2006). Across the board, pre-test averages for students on the LSCI are approximately 25%, or the score that students would earn if they were guessing. This is true whether the students are at community colleges or at prestigious four-year institutions (Prather et. al., 2009). Without evidence to the contrary, it is reasonable to suggest that the LSCI, unlike the FCI, is not measuring a construct that involves prior or robust conceptions. The lack of robust, deeply entrenched conceptions differentiates force from any topic in astronomy, and the FCI from the astronomy concept inventories.

Finally, the FCI measures a shift in students' ontological understanding of the Universe. Student thinking about force, as measured in the FCI, is not a collection of loosely associated ideas. On the contrary, student thinking related to force tends to fall into two coherent worldviews: Aristotelian and Newtonian. The FCI specifically measures the ontological shift between these two ways of interpreting the known world. According to the authors, a poor score on the FCI indicates that the student possesses a clearly non-Newtonian worldview. A score near 60% indicates that the student has an emerging conception of a Newtonian world, and a score over 85% indicates that the student is a "Newtonian thinker" (Hestenes & Halloun, 1995). No topic in astronomy requires students to restructure their ontological systems, as typically taught at the introductory level.

Without weighing in on the appropriateness of using the FCI as proxy for learning across a course, an argument can be made that the FCI instrument measures a potent construct based in three unusual characteristics of the concept of force. We do not have a similar topic in astronomy, as the content and structures of astronomy courses lack a topic that serves as the basis for the rest of the curriculum, and that builds upon our oldest and most entrenched conceptions of the world, and that actually requires a shift in the way that we view the world. As such, an FCI analog is not possible in the astronomy content domain.

Existing Tests

Although it is inappropriate to use a single-concept instrument to evaluate student learning across an astronomy course, many well-constructed single-concept inventories exist in astronomy that provide indispensible insight into student thinking. These instruments, in addition to the ADT2, provided the basis upon which a broad-content survey such as the TOAST could be constructed, and are listed Figure 1.

Astronomy Surveys and Concept Inventories		
ADT2:	Astronomy Diagnostic Test 2	(Zeilik, 2002)
LPCI:	Lunar Properties Concept Inventory	(Lindell, 2004)
GECI:	Greenhouse Effect Concept Inventory	(Keller, 2006)
LSCI:	Light and Spectra Concept Inventory	(Bardar, 2006)
SPCI:	Star Properties Concept Inventory	(Bailey, 2012)
NGCI:	Newtonian Gravity Concept Inventory	(Williamson, 2013)

Figure 1: Astronomy Surveys And Concept Inventories

Each of these instruments successfully differentiates students who hold scientifically correct conceptions in astronomy from students whose understanding of the astronomical content is dominated by alternative, non-scientific thinking. In addition, the ADT2 set out a distinct set of design principles that flow like DNA through a second generation of assessments, the single-topic concept inventories. Each of these instruments are short, ranging from 20 to 33 questions, and are written in natural, as opposed to jargon-laden, language. Each test item is constrained to assess understanding of a single idea, facilitating interpretation of results. Most importantly, each of these assessments is rooted in research on student thinking, collectively representing thousands of hours of research, in the form of interviews, surveys, and pilot testing of items. Much of the similarity in these instruments may be due to an overlap in the teams who created them. In each, Timothy Slater (e.g., ADT2, LPCI, GECI, LSCI, SPCI, and NGCI) or Edward Prather (e.g., GECI, LSCI, SPCI, and NGCI) were either formally part of the authoring team or were engaged in the process of drafting content parameters or test items. In other cases, lead authors on one concept inventory served as an expert consultant on another.

While these concept inventories are a closely related family of assessments, there are important differences between these assessments that stem from seemingly small variations in the intentions that underlie their construction. These differences group the instruments into three categories. The first group of instruments was the product of efforts to understand students' mental cartoons of a single phenomenon. Lindell's LPCI and Keller's GECI began with very open-ended probes into students' scientific understanding of lunar phases and the greenhouse effect, respectively. After an iterative process of refining the range and domain of student thinking, themes were characterized. The parameters for the content addressed in test items are rooted in that work, rather than from expert opinion as to the core scientific competencies required for mastery, best suiting these instruments to diagnose students' mental models. In contrast, the second group of assessments began with expert conceptions of the concept's core competencies. Bailey's SPCI and Williamson's NGCI used expert conceptions related to stars and gravity, respectively, to create survey and interview protocols that would reveal student thinking. Questions were crafted to probe the range of knowledge that an expert would have related to the concept, while the range of student thinking was used to craft the possible responses for each question. These instruments are useful for collecting information on students' prior knowledge and on the efficacy of instruction in moving students towards expert conceptions of the content. The third group of instruments began as tools for evaluating the effectiveness of particular curricula. The ADT2 is a revision of the Astronomy Diagnostic Test (ADT), which originated in efforts to determine the usefulness of concepts maps to overcome the most robust misconceptions in the domain of astronomy. The LSCI began as an evaluation tool for the *Project LITE* curriculum materials, and in the process of revision, "a few items were also borrowed…from the *Lecture Tutorials* [curriculum] project" (Bardar, 2006, p. 70). As a result, the test items on both instruments are influenced by the range and domain of specific teaching interventions, rather than from an expert-derived notion of the concepts' core competencies. They are useful in differentiating students who have had this instruction from those who have not, and they may also be useful for determining the fidelity of implementation of those curricula (Prather et. al., 2009).

Given the massive amount of effort and expertise invested in these works, the TOAST was deliberately constructed using these instruments as source material, whenever possible. As such, the TOAST should be considered a third generation assessment in astronomy, building on, and built from, two previous generations of scholarship.

Determining The Parameters Of "General Astronomy Content Knowledge"

Common experience suggests that it is nearly impossible to get any two experienced instructors of astronomy to come to agreement with regard to the "correct" content of an introductory astronomy survey course. This is understandable, given that astronomy is the science that literally encompasses the entire Universe as its domain. However, there are at least two ways to approach the question of what "students should know about astronomy." On one hand, we could engage in a discussion of what should be covered and what content should be excluded, defining the entire curricular scope for introductory astronomy courses. We could make the judgment that instructors should teach the celestial sphere but not the Hubble tuning fork; stellar evolution, but not stellar magnitudes; and so on. It is unlikely that such a discussion would be pleasant or ever come to resolution. On the other hand, we could assert that there are a few fundamental ideas that every person should grasp in order to master the construct of "general astronomy content knowledge," and that beyond that, instructors should have the freedom to supplement the basics with additional instruction on the topics of interest to themselves and their students. This second approach to defining the core content of an introductory astronomy course is not only a more pleasant path to take, but hundreds of experts had spent untold hours engaged in this work, prior to the construction of the TOAST.

Over the past twenty years, at least three entities have taken up the serious work of defining the core concepts of astronomy for the non-astronomer. This work is documented in the National Research Council's *National Science Education Standards (NSES)* (1996), the American Association for the Advancement of Science's *Project 2061: Benchmarks for Science Literacy (Benchmarks)* (1986), and the American Astronomical Society's *Goals for Astro 101 (Goals)* (Partridge & Greenstein, 2004). While the *NSES* and *Benchmarks* documents are intended to guide instruction for K-12 rather than college settings, they nonetheless describe the bare minimum that a scientifically literate individual should know about astronomy, as judged by hundreds of experts in astronomy and astronomy education. While it can be argued that students should learn more than this in a college level course, it cannot reasonably be argued that they should leave a college level course knowing less. Moreover, research studies across a variety of settings indicate that students leave high school and enter introductory astronomy courses without mastering these concepts. (cf Balfour & Kohnle, 2010; Kalkan & Kiroglu, 2007; Trumper, 2000; Turkoglu, Ornek, Gokdere, Suleymanoglu & Orbay, 2009); Zeilik & Morris, 2003). Clearly, if students are to learn these concepts, introductory astronomy courses will have to take a lead role. The importance of this role is magnified by data indicating that 40% of future teachers take introductory astronomy courses, and that many of these teachers will serve in elementary schools where the majority of K-12 astronomy is taught (Lawrenz, Huffman, & Appeldoorn, 2005). If pre-service teachers leave their astronomy courses without a grasp of the K-12 science, they will not be a position to properly teach astronomy to the next generation of students.

METHODS AND RESULTS

The development of an assessment instrument is more a matter of engineering than science in that the goal is to construct a tool that can serve a given purpose, using a certain set of design specifications. Rather than articulating a conventional set of research questions and a methodology, this section describes the performance specifications of the TOAST and the development and validation of the TOAST, which occurred in three phases: the creation of a consensus document providing explicit criteria for the scientific content; the selection and creation of items aligned with the criteria; and testing of the TOAST in order to determine the extent to which it met the design specifications. Results are given in the relevant section, where appropriate.

Performance Specifications For The TOAST

The TOAST is intended to serve as a measure of students' "general astronomy content knowledge" at the level of a college introductory astronomy course. As such, the content of the TOAST should reflect an expert consensus of the knowledge that a layperson should possess in astronomy, and its test items should constitute a reasonable sample of that knowledge. The TOAST should measure conceptual knowledge, rather than knowledge that is rooted in scientific vocabulary, and is intended to be sensitive to instruction and experience in astronomy. The TOAST should not unintentionally measure other common factors such as reading ability, and should avoid biases toward any particular curriculum package. The TOAST should adhere to general principles of good assessment practices, and specifically conform to accepted practices in astronomy education. The TOAST is

intended to be short in length (e.g., no more than 30 test items) and be written in the natural language of target students. Test items on the TOAST should only assess one concept, and distractors should reflect students' most common alternative conceptions. Finally, the TOAST should be able to discriminate among populations that have received different levels or a different quality of exposure to general astronomy content.

Development And Validation Of The TOAST

The criteria for the TOAST is derived from a consensus of three previously cited, expert position statements related to the core ideas in astronomy. This consensus document was reviewed for scientific coherence by an additional body of 28 experts in astronomy and astronomy education. The source documents: the *National Science Education Standards* (NSES), developed under the supervision of the National Research Council (NRC) (1996); *Project 2061: Benchmarks for Science Literacy*, developed by the American Association for the Advancement of Science (AAAS) (1986); and the American Astronomical Society's (AAS) *Goals* for introductory astronomy courses (Partridge & Greenstein, 2004), represent the combined wisdom of our field. Deriving a consensus from these three documents involved compromise and judgment calls, largely due to the different approaches that these three organizations used in the construction of their documents. Nonetheless, after eliminating all material not directly related to astronomy content knowledge, the three documents demonstrate a remarkable degree of agreement. The few astronomical concepts that did not appear in all three documents were deemed to be outside of the core of astronomy knowledge and were removed, with one exception. "Scale," or "a broad understanding on the scope" of the Universe was unique to the goals set forth by the AAS. The decision was made to retain the idea of scale in our consensus document as it represents a concept that is critical, and perhaps unique, to astronomy. The remaining concepts were restructured into coherent criteria using a reiterative, inductive card sorting approach. The resulting consensus document parsed out the criteria for the construct of general astronomy content knowledge into 11 core ideas, hereafter referred to as "criteria." These criteria were further grouped into three meta-criteria: Physical Laws and Processes, the Structure and Evolution of the Universe, and Patterns in the Sky. A table indicating which TOAST items are aligned with the three meta-criteria and 11 criteria are shown below in Table 1.

Table 1: Criteria And Test Item Details For The Test Of Astronomy STandards (TOAST)

Meta-Criteria	Criteria	TOAST Survey Items
Physical Laws and Processes	Gravity	Questions 20, 21
	EMR & EMR production	Questions 23, 25, 26, 27
	Fusion & the formation of heavy elements	Questions 8, 22, 24
The Structure and Evolution of the Universe	The Evolution of the universe	Questions 9, 15
	Star and stellar evolution	Questions 13, 14, 16, 17
	The evolution and structure of the solar system	Questions 18, 19
	Seasons	Questions 7, 12
	Scale	Questions 10, 11
Patterns in the Sky	Yearly patterns	Questions 2, 4
	Daily patterns	Questions 1, 6
	Moon phases	Questions 3, 5

An exhaustive document illustrating the agreement between the NRC's, AAAS', and AAS' scientific content, and the resulting criteria, was subjected to review by the expert group. The group of experts included 28 research astronomers, introductory astronomy instructors and active members of the astronomy education community, associated with astronomy research institutions, research and liberal arts universities, and community colleges nationwide. Rather than focusing on the content validity of the science, which had been soundly argued for by the experts of the NRC, AAAS, and AAS previously, this collection of experts was asked to focus on the degree to which the science content had been logically grouped into concepts. In every case, the experts agreed with the arrangement of the scientific content. The document illustrating the alignment of the astronomy content from the *Goals*, *Benchmarks*, and *Standards* is quite lengthy and is not included in this paper. Copies are available by corresponding with the author.

Aligning TOAST Test Items To The Criteria

Using the TOAST's scientific content criteria, 27 test items were sourced or constructed to measure a reasonable sample of general astronomy content knowledge. The psychometric criteria for item selection were consistent with those used in the construction of the ADT2 and the existing astronomy concept inventories, as described by Slater and Adams (2003). All items test one concept per question, are written such that the correct answer can be known before reading the choices, and avoid scientific jargon. Items were sourced from the existing astronomy assessments, wherever possible. When these sources failed to provide items that would assess students' understanding of the criteria, instruments that had been constructed for use with introductory astronomy courses were searched, such as those provided in Slater & Adams (2003, Appendix A). These items were compared to findings from research in student learning in astronomy and modified if necessary. In the case of two criteria, cosmology and the structure and formation of the solar system, high quality test items that reflect our current understanding of students' conceptions were not immediately available. Four new test items were constructed to assess these criteria. These questions were written in light of what we know of student thinking from the research team's teaching experience, and from the small amount of available research on the subject (Prather, Slater, & Offerdahl, 2002).

Prior to expert review, all test items were subject to review for needed modification related to readability, such that reading ability, beyond an eighth grade level (i.e., 14 years of age), should not factor into test results. In the ADT2 and the existing concept inventories, an attempt had been made to employ natural student language; however, no formal attempt had previously been made to validate these efforts quantitatively. As part of the TOAST development, readability was measured for each test item using the Flesch-Kincaid Grade Level test (Kincaid, Fishburne, Rogers, & Chissom, 1975). This test rates text, giving a US school grade level score; a score of 8.2 represents text that is written at the reading level of an eighth student in the second month of school. Although short sections are difficult to rate on this scale, any portion of the TOAST with a Flesch-Kincaid score of 8.0 or higher, or an eighth grade reading level or higher, was revised to make the section easier to read. For example, one of the TOAST's test item's prompt reads: "Which sentence best describes why the Moon goes through phases?" One of the scientifically inaccurate distractor responses initially read: "The sunlight reflected from Earth lights up the Moon but is less effective when the Moon is lower in the sky than when it is higher in the sky." This response had a Flesch-Kincaid score of 10.3. To increase its accessibility, this response was edited, turning one long sentence into two shorter sentences: "The sunlight reflected from Earth lights up the Moon. It is less effective when the Moon is lower in the sky than when it is higher in the sky." This alteration lowered the reading score of the response to 4.7. The entirety of the TOAST has a Flesch-Kincaid score of 6.8. As such, reading ability is not likely to inadvertently influence a typical college student's ability to demonstrate their general astronomy content knowledge on the TOAST.

The scientific content criteria and the related test items were then returned to the expert panel in order to determine whether or not the test items represent a fair test of the content. The questions in the TOAST need to probe the central ideas of each criteria and represent a reasonable sample of the targeted domain. This is difficult given that the TOAST only has room for a few questions per content criteria. For instance, in Bailey's (2014) Star Properties Concept Inventory (SPCI) students were asked 30 questions related to stars and star formation. The TOAST only had room for two or three of these questions. A judgment was made that the relationship between a star's mass and its life span is more crucial to the stars concept than the function of convection cells within a star. The TOAST addresses the former but not the latter topic. A similar decision-making process had to be repeated multiple times for each criterion. The expert panel was asked to comment on the degree to which these decisions were correctly made. Supplied with the TOAST's content criteria, the table of AAS/AAAS/NRC scientific content, and the test items, the experts were asked three questions:

1. "Does the test item assess an important idea related to the astronomy concept, or should the test item be replaced with another, more relevant question?"
2. "Is the test item an appropriate measure for students' knowledge at the introductory science level?"
3. "Do you have any concerns or comments related to the test items?"

Relevant to the first question, all items, with the exception of one, were judged to be core to the content being assessed. This question, related to gravity, was removed and replaced with a question that was judged to more closely target the domain at the introductory level. Relevant to the second question, all items were deemed to be appropriate measures at the introductory level. From the third question, two concerns were raised. In one case, three of the experts expressed a concern about the tie between a question related to light and spectra, and a curriculum product, *Lecture Tutorials for Introductory Astronomy* (Adams et. al., 2003). Although the connection between the two was not readily apparent to the TOAST writing team, a review of Bardar's (2006) dissertation indicated that some of the items used in the LSCI were sourced from the *Lecture Tutorials* project (2006). This gave the concern credence, and the item was replaced. In another case, there was some concern over the wording in a question related to the Big Bang theory. Initially, one of the distractors read: "The event that created all matter and space from an infinitely small dot of energy." Responding to expert review, the word "created" was replaced with the word "formed."

Testing Of The TOAST

Testing of the TOAST instrument, in its entirety, was conducted in four phases: expert testing, target audience testing, testing for sensitivity to instruction, and a construct validity check, comparing initial results to the existing literature on student conceptions in astronomy.

Ensuring Scientific Validity Through Expert Testing

The astronomers who previously served as expert reviewers generously agreed to take the TOAST assessment, as if they were students, and to provide think-aloud accounts of their thoughts as they answered a selection of questions. This phase of testing was intended to validate the scientific content of the TOAST. It was observed that the group of 17 astronomers received an average score of 98% on the TOAST, with no astronomer answering more than two test items incorrectly. Follow-up interviews were conducted in order to determine the reason that certain items were not answered correctly. Each astronomer was asked to provide a think-aloud describing how she or he would answer two questions. In the event that the astronomer incorrectly answered a question, they were asked to talk about that question; if the astronomer did not miss two questions, they were asked to talk through their answers to questions that other astronomers missed. The astronomers were not told why these questions had been selected, or that they had incorrectly answered the questions.

The think-aloud process indicated that astronomers answered test items incorrectly either because they initially gave the task insufficient attention, or because they possessed the same alternative conceptions seen in the general population. In particular, the questions most often missed by astronomers were related to motion of the celestial sphere, a topic rarely used by astronomers in their professional work. There was no evidence to suggest that the astronomers incorrectly answered the questions due to flaws in the test's scientific content, as for every question that was missed by an astronomer, several other astronomers provided think-alouds resulting in the selection of the scientifically accurate answer. This process provided evidence that the scientific content of the TOAST is functioning as needed.

Piloting The TOAST With The Target Population

Data was collected from a large sample of introductory astronomy students, as a pre-test before instruction, in order to quantitatively assess the TOAST's validity for this population, using the psychometric principles of Classic Test Theory (CTT) (Novick, 1966). The participants for this sample were 1104 non-science majoring undergraduates enrolled in introductory astronomy courses at two comprehensive doctoral-granting research universities, two liberal arts universities, and one community college. All data were visually inspected: Student responses with more than two pieces of missing data, and those that obviously demonstrated an attempt to make patterns out of the answer bubbles, were discarded. The remaining sample of 1066 students provided sufficient sample size to conduct item response analysis. Responses were analyzed using Remark Classic OMR v2. to calculate Cronbach alpha, item difficulty and discrimination, and the mean, median, and standard deviation of the original sample population.

As evidence of the degree to which the TOAST is internally consistent, the TOAST survey has a Cronbach alpha of 0.83. The Cronbach alpha score is a measure of inter-item correlations and can be interpreted as a measure of how well a set of items measures a single construct (UCLA, 2008) while a low score could indicate a variety of things, including test fatigue. The widely accepted social science cut-off value for a set of items to be considered a coherent scale is 0.70 or higher (Nunnaly, 1978). As such, this data indicates that the TOAST demonstrates a level of internal consistency and shows no signs of test fatigue for participants. This is likely associated with the easy readability and the short length of the TOAST.

Two additional aspect of Classic Test Theory were analyzed: item difficulty and item discrimination. Item difficulty is a measure of the proportion of the population who correctly answered the test question. This data is given in *p*-values in Table 2. An item difficulty *p*-value between .30 and .90 is typically desirable (Haladyna, Downing, & Rodriguez, 2002). The average *p*-value for the TOAST items is 0.46 with all items scoring in a desirable range. Item discrimination is a measure to which success on a given item correlates to success on the overall instrument. Item discrimination for each item is given in Table 2 as a point-biserial correlation, calculated as the Pearson correlation between responses to a particular item and scores on the total test. A point-biserial of 0.15 or higher is considered satisfactory. The average point-biserial on the TOAST was 0.42 with 0.28 representing the lowest point-biserial. Together, these measures indicate that the TOAST is measuring astronomy content knowledge at the correct level for the population and that individual items are functioning in a psychometrically appropriate manner.

In addition, point-biserial values for each item's distractors, or scientifically inaccurate responses, were calculated. In most cases, scientifically inaccurate responses should exhibit negative point-biserials, indicating that students that perform well on the TOAST tend to avoid incorrect answers. Only one distractor on the TOAST returned a positive point-biserial. When asked about the source of the "atoms in the plastic of your chair" 10% of students (n=25) chose Distractor A, saying that these elements were formed "in our Sun." This distractor exhibits a point-biserial of 0.04. Distractor A was designed to attract students who may misapply the culturally transmitted idea that heavy elements, like the carbon and oxygen atoms in plastic molecules, form in the cores of stars. This option was not chosen at a high frequency, but the students who did choose this answer performed moderately well on the remainder of the instrument. This makes sense; one must possess cultural knowledge (e.g., the knowledge of atomic fusion in stellar bodies) before they are able to apply it in inappropriate ways. No other distractors exhibited positive point-biserials. We judge that these statistics indicate that each item, and each distractor, is functioning as intended and makes a meaningful contribution to this instrument.

Table 2: Item Difficulty And Discrimination For TOAST Test Items.

	***p*-Value**	**Point-Biserial**		***p*-Value**	**Point-Biserial**
Item 1	0.37	0.44	Item 15	0.28	0.41
Item 2	0.39	0.43	Item 16	0.79	0.42
Item 3	0.57	0.39	Item 17	0.28	0.29
Item 4	0.66	0.46	Item 18	0.56	040
Item 5	0.61	0.33	Item 19	0.22	0.56
Item 6	0.23	0.38	Item 20	0.35	0.42
Item 7	0.53	0.28	Item 21	0.33	0.47
Item 8	0.43	0.58	Item 22	0.40	0.44
Item 9	0.47	0.32	Item 23	0.40	0.50
Item 10	0.63	0.40	Item 24	0.28	0.41
Item 11	0.60	0.39	Item 25	0.41	0.48
Item 12	0.36	0.50	Item 26	0.20	0.31
Item 13	0.32	0.61	Item 27	0.26	0.37
Item 14	0.39	0.43			

The mean score for this sample population was 12 correct out of 27 items, or 44%, while the median was 11 correct, and the standard deviation was 5.4 (n=1066). The low average score is deemed to reflect the attractive misconceptions found in many of the TOAST's distractors.

Testing For Sensitivity To Instruction

In addition to pilot testing with a group of experts and a student target group, data was collected from three additional populations in an attempt to determine whether or not the TOAST can discriminate between groups with obviously different levels of education and experience in astronomy. The participants included 313 amateur astronomers, 519 alpha teachers (i.e., teachers with a history of participation in astronomy education workshops), and 32 in-service, non-science teachers. It can be reasonably assumed that these groups represent different levels of prior engagement with astronomical concepts, through formal education, their own teaching of astronomy, and in the use of astronomical content in work and avocational settings. A comparison of tests results can be seen in Table 3.

Table 3: Average TOAST Scores For Five Distinctly Different Populations

Sample	Average TOAST Score
Professional Astronomers (n=17)	98%
Amateur Astronomers (n=313)	83%
Alpha Teachers (n=519)	66%
In-service Teachers (n=32)	51%
Astronomy 101 Students (n=1066)	44%

A one-way ANOVA analysis indicates that the differences in the TOAST scores between these groups are statistically significant [$F(4, 101) = 23469.46$, $p < 0.05$]. This is judged to mean that the TOAST is sensitive to instruction and informal experience in astronomy.

Comparing TOAST Results to the Research on Student Thinking

In addition to the previously described analyses, a qualitative analysis was conducted to judge the survey's construct validity, or the degree to which the TOAST supports and is supported by relevant theory (Cronbach & Meehl, 1955). As a means of judging this instrument's construct validity, the distribution of participants' responses to each question were compared to the literature on student learning on that topic. It is expected that student responses to questions addressing topics in which there are significant and robust misconceptions should show a distribution reflecting those misconceptions. The entirety of this analysis is too lengthy to include in the body of this paper; however, the following passage presents an example of the results for two questions on the TOAST: Questions 7 and 12. These two questions address seasons on Earth, as related to the effects of Earth's tilt and orbit. There is a large body of literature related to this topic, leading to the clear conclusion that the general populace believes that the distance between Earth and the Sun is the cause of the seasons (viz., Kikas, 2004; Klein, 1982; Nussbaum, 1979; Nussbaum & Novak, 1976; Parker & Heywood, 1998; Sneider & Pulos, 1983; Schneps & Sadler, 1988; Slater, S., 2007; Trumper, 2000; Vosniadiou & Brewer, 1992). This alternate conception is so robust that learners will create a variety of hybrid, or synthetic, conceptions in order to retain this belief in the face of conflicting observation, data, or cultural knowledge. For instance, when students "learn" that Earth's tilt is responsible for Earth's seasons, many interpret this to mean that the difference in the distances between the northern and southern hemispheres, and the sun, is sufficient to cause summer and winter. In other words, students may report that Earth's tilt causes the seasons, but upon further investigation it becomes clear that many believe the tilt results in substantial differences in distance, and that difference in distance is responsible for the seasons. In the absence of compelling educational interventions, student responses on these TOAST questions should reflect an adherence to distance-related mental models. Questions 7 and 12, with data on students' responses are given in Figures 2 and 3, followed by an interpretation of the results.

7. Imagine that Earth was upright with no tilt. How would this affect the seasons?
   A. We would no longer experience a difference between the seasons.
   B. We would still experience seasons, but the difference would be *less* noticeable.
   C. We would still experience seasons, but the difference would be *more* noticeable.
   D. We would continue to experience seasons in essentially the same way we do now.

Response	Frequency of response	Percent of total responses	Point-biserial
A	**558**	**52.3**	**0.28**
B	292	27.4	-0.05
C	161	15.1	-0.21
D	55	5.2	-0.11
	1066	100.0	

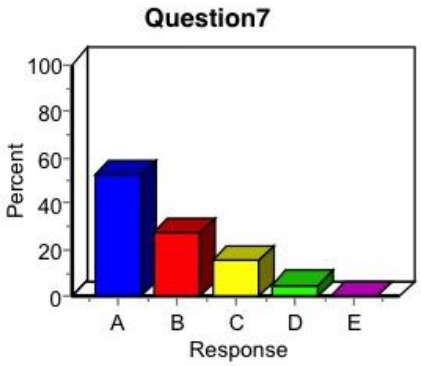


Figure 2: TOAST Question 7, With Results

12. Imagine that Earth's orbit were changed to be a perfect circle about the Sun so that the distance to the Sun never changed. How would this affect the seasons?
   A. We would not be able to notice a difference between seasons.
   B. The difference in the seasons would be *less* noticeable than it is now.
   C. The difference in the seasons would be *more* noticeable than it is now.
   D. We would experience seasons in the same way we do now.

Response	Frequency of response	Percent of total responses	Point-biserial
A	211	19.8	-0.31
B	322	30.2	-0.09
C	156	14.6	-0.22
D	377	35.4	0.50
	1066	100.0	

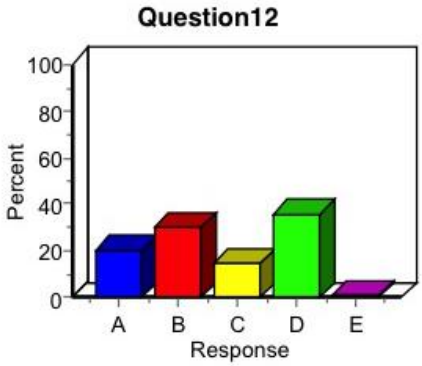


Figure 3: TOAST Question 12, With Results

In the results to Question 7, nearly 52.3% of participants correctly indicated that if the Earth had no tilt, there would be no perceptible seasons. Only 5.2% of participants chose Distractor D, indicating no knowledge of the important relationship between Earth's tilt and our seasons. What is interesting is that 42.5% of students indicate they possess some knowledge that this is an important cause and effect relationship, but are not quite sure of the consequences of making changes to the system. In Question 12, 35.4% of students indicated that they possess an accurate understanding of the relative importance of distance in determining changes in Earth's seasons. However, 62.31% of students chose a response in which changes to the shape of Earth's orbit would result in changes to our seasons. Taken together the results from these two questions tell a story that is clearly reflected in the literature: Many students do not possess a clear understanding of the cause of seasons, and that student understandings can roughly be divided into scientific, naïve (inaccurate), and synthetic (hybrid) conceptions (Vosniadou & Brewer, 1992). Similar results were seen in other test items; as such, these TOAST items demonstrate strong construct validity. The Test Of Astronomy STandards, or the TOAST, is provided in the Appendix.

On Abstaining From Rasch Modeling

Recently, Rasch modeling has experienced a surge in popularity as a means of analyzing assessments (Rasch 1960). Rasch modeling, like classical test theory and the Pearson's correlations used in this study, can be employed to provide information about how test items function, how subjects respond to those test items, and can give some indication whether or not an item should be retained in an assessment. However, Rasch modeling is not suited to the TOAST as Rasch assumes that the thing you are trying to measure is unidimensional and that all test items should be equally difficult. If astronomy education research and common experience have taught us anything, it is that general astronomy content knowledge is not unidimensional. Some concepts in astronomy are far more difficult to learn than others, depending upon the cognitive factors that influence mastery (e.g., cognitive structures, phenomenological primitives, identity structures) (Slater, 2013). As Goldstein (1979) points out, once one determines that responses to different items are determined by different processes, one ought not to expect all of the items to fit a simple unidimensional model like Rasch. As such, the TOAST is poorly suited for Rasch modeling. However, this should not be considered a weakness of the TOAST, as results from Rasch modeling are not intended to be used to remove an item from an assessment without a thorough examination of the item, looking at best practices of assessment, the nature of the content, and the ways in which students interact with that content. In the absence of Rasch modeling, the previously described validation steps have been used to examine each of the TOAST's test items multiple times, providing for great confidence that all items are functioning as intended.

Limitations Of The TOAST

The TOAST is limited by the same limitations inherent in all of research studies underlying its construction. This research is based in the knowledge, thought processes, and abilities of fairly homogenous samples of convenience. Further, while pilot testing of concept inventories involved large, general education populations, the interview protocols were enacted with students who were overwhelmingly white, full-time, and middle socio-economic class. These college students were comfortable communicating in English, in academic settings, such that they were willing to volunteer and have extended one-on-one communication with a person in a superior position. These students were also willing to be vulnerable, revealing their thinking about scientific content to a person in a position to judge them, indicating a certain amount of confidence. As a consequence, it is possible that the research that underlies the TOAST provides an incomplete picture of the full range of student thinking in astronomy. As the content contained in the TOAST is tested across more variable populations, it is hoped that we will get a better idea of how other populations of students interact with the science of astronomy. As that happens, a revised TOAST, or its descendent, might be able to make more accurate measurements of student thinking.

Further, while evidence has been put forth that the TOAST is a valid and reliable assessment in a variety of situations that does not mean that the TOAST is valid and reliable for every application. Validity and reliability are things that must be argued for in each new situation. This is true for every assessment tool, including the TOAST. As users consider which assessment they should use, there are at least two questions that they should consider. The first is: "Does the new population interact with the material in the same way as the populations reported in this paper?" In particular, if the new population does not communicate in English at an eighth grade level or better, results from the TOAST will not just reflect students' knowledge of astronomy, they will also reflect students'

reading ability. In such cases, when working with students who are second language learners, have learning disabilities, or who have a history of poor performance in school tasks, the author would encourage instructors and researchers to consider using alternative assessments, particularly those assessments that are open-ended and that allow students to represent their knowledge using drawings or other inscriptions. Secondly: "Does this assessment measure content that is valuable in this setting?" Alignment between an assessment and the knowledge that we wish to measure is crucial. As Erik Brogt and colleagues found in their study of data from the ADT2 (2007), the most important indicator of learning gains is the extent to which the assessment matches the intend teachings in a course. If the TOAST, or any other assessment fails to align with the domain or exceeds the depth of the intended curriculum, the resulting data is not likely to represent a valid measure.

DISCUSSION

The TOAST is primarily intended to be used as a measure of students' mastery of the core concepts associated with an introductory astronomy course. As indicated by the 2012 *Decadal Survey* (Deustua, Noel-Storr, & Foster, 2009), the TOAST is unique in its ability to provide this type of instructional assistance due to its close tie to the criteria set forth by the astronomy education community. In that capacity, it may be used to multiple ways. First, the TOAST may be used to improve instruction in an individual instructor's course. As a background knowledge check at the beginning of a course (Cross & Angelo, 1993), this survey can provide instructors with an idea of their students' prior conceptions across the entire content of an introductory astronomy course. Education research literature across many fields indicates that an awareness of and response to students' prior knowledge is critical for effective instruction (Bransford, Brown, & Cocking, 1999). In conjunction with the use of the TOAST as a background knowledge check, the TOAST can be administered after instruction. This can provide instructors with evidence of the effectiveness of instruction within their own classrooms. When used to compare different instructional techniques employed across sections or semesters, individual teachers can gain insight into the impact of instructional strategies within their own teaching context. As a variation, the TOAST can be used to measure the impact of adjusting instruction in one of the three criteria "meta-categories." For instance, an instructor might note that students' understanding of patterns in the visible sky does not seem to improve after lecture-based or lecture and discussion-based instruction, and may decide to add a laboratory or homework component for that portion of the course. The instructor may use the related TOAST questions in a pre- and post-test fashion to help determine if this intervention is helpful in their particular context, similar to the appropriate use of a concept inventory.

In education studies, the TOAST can be used to make empirically valid claims about the quality and impact of instruction across astronomy courses, and can be used to probe various factors that interact with students' abilities to learn astronomical concepts. A number of published studies provide evidence of the TOAST's ability to provide meaningful research data. Inge Heyer's (2009) work on the interaction between college students' spatial reasoning skills and their abilities to learn astronomy; Theresa Moody's (2010) work on the impact of professional development on teachers' confidence in teaching astronomy; Dan Lyon's (2011) work on how inquiry - oriented labs improve student understanding; Debra Stork's (2014) work on K-12 teachers' knowledge of astronomy as related to national standards and frameworks; and Coty Tatge's (Tatge & Slater, 2014) current work on the impact of culture and language in the teaching and learning of astronomy, provide some notion of the range of studies that might benefit from the broad snapshot of astronomical content knowledge measured by the TOAST.

Although the TOAST is not the "FCI for Astronomy," the successful assessment development process for making the TOAST holds promise for other discipline based education research endeavors. The TOAST's creation was largely possible because it leveraged the extensive survey, interview, and item development on single-topic concept inventories of earlier astronomy education researchers. The physics education research community has created several single-topic concept inventories poised to be folded together into a comprehensive, third-generation physics survey in the same way as the TOAST. Similarly, the life sciences have a large library of single-topic concept inventories that too might benefit from the TOAST development process. In the end, revisiting the community's consensus of what constitutes broad scientific knowledge in a domain, such as was systematically done in generating the TOAST's criteria, is a vital response for all disciplines as new science education standards and frameworks repeatedly appear on the horizon.

AUTHOR INFORMATION

Dr. Stephanie J. Slater is the Director of the CAPER Center for Astronomy & Physics Education Research, an international collaborative of scholars studying contemporary issues in science education and cognition. Her research focuses on the intersection between learning science and socio-cultural issues. E-mail: stephanie@caperteam.com

REFERENCES

Adams, J.P., Prather, E.E., & Slater, T.F. (2003). *Lecture Tutorials for Introductory Astronomy.* Upper Saddle River, New Jersey. Prentice Hall

American Association for the Advancement of Science (1986). *Project 2061: Benchmarks for Science Literacy*. Washington, DC.

Bailey, J. M., Johnson, B., Prather, E. E., & Slater, T. F. (2012). Development and validation of the star properties concept inventory. *International Journal of Science Education*, *34*(14), 2257-2286.

Balfour, J., & Kohnle, A. (2010). Testing conceptual understanding in introductory astronomy. *New Directions*, (6), 26-29.

Bardar, E. M. (2006). *Development and analysis of spectroscopic learning tools and the light and spectroscopy concept inventory for introductory college astronomy*. Ph.D. Dissertation, Boston University. (Order No. 3214908). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (305364831). Retrieved from http://search.proquest.com/docview/305364831?accountid=14793

Baxter, J. A., & Lederman, N. G. (1999). Assessment and measurement of pedagogical content knowledge. In *Examining pedagogical content knowledge* (pp. 147-161). Springer Netherlands.

Bodner, G. M., & McMillen, T. L. (1986). Cognitive restructuring as an early stage in problem solving. *Journal of Research in Science Teaching*, *23*(8), 727-737.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press.

Brogt, 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.

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172-1182.

Corbett, H. D., Wilson, B. L., & Educational Resources Information Center (U.S.). (1989). Raising the stakes in statewide mandatory minimum competency testing. Philadelphia, PA: Research for Better Schools.

Cross, P. & Angelo, T. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers* (Second Edition). San Francisco: Jossey-Bass Publishers.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin.* 52, 281-302.

Deustua, S., Noel-Storr, J., & Foster, T. (2009). In Support of Astronomy Education Research. In *astro2010: The Astronomy and Astrophysics Decadal Survey* (Vol. 2010, p. 9P).

Goldstein, H. (1979). Consequences of Using the Rasch Model for Educational Assessment. *British Educational Research Journal*, *5*(2), 211-220.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow & Co.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*. 15(3), 309-333

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory. *The Physics Teacher*, *33*(8), 502-506.

Heyer, I. (2012). *Establishing the empirical relationship between non-science majoring undergraduate learners' spatial thinking skills and their conceptual astronomy knowledge* (Order No. 3507367). Ph. D. Dissertation University of Wyoming. ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (1015156166). Retrieved from http://search.proquest.com/docview/1015156166?accountid=14793

Kalkan, H., & Kiroglu, K. (2007). Science and nonscience students' ideas about basic astronomy concepts in preservice training for elementary school teachers. *Astronomy Education Review*, *6*(1), 15-24.

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. Ph. D. Dissertation University of Arizona. (Order No. 3237466). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (305353131). Retrieved from http://search.proquest.com/docview/305353131?accountid=14793

Kikas, E. (2004). Teachers' conceptions and misconceptions concerning three natural phenomena. *Journal of Research in science Teaching. 41* (5), 432-448.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Klein, C. (1982). Children's concepts of the Earth and the sun: A cross-cultural study. *Science Educator. 65*, 95–107.

Lawrenz, F., Huffman, D., & Appeldoorn, K. (2005). Enhancing the Instructional Environment: Optimal Learning in Introductory Science Classes. *Journal of College Science Teaching*, *34*(7), 40.

Laws, P., Sokoloff, D., & Thornton, R. (1999). Promoting active learning using the results of physics education research. *UniServe Science News*, *13*, 14-19.

Lindell, R. S., & Sommer, S. R. (2004, September). Using the lunar phases concept inventory to investigate college students' pre-instructional mental models of lunar phases. In *2003 Physics Education Research Conferences: 2003 Physics Education Conference* (Vol. 720, No. 1, pp. 73-76). AIP Publishing.

Lyons, D.J. (2011). *Impact of backwards faded scaffolind approach to inquiry-based astronomy laboratory experiences on undergraduate non-science majors' views of scientific inquiry.* Ph. D. Dissertation University of Wyoming.

Madaus, G. E. (1998). The influence of testing on the curriculum. *Yearbook-National Society For The Study Of Education*, 2, 71-112.

Mangione-Leslie, K., Dockers, J., & Wavering, J. (2005). "What do they know? A look into preservice teachers' earth science content knowledge." Proceedings of the International Association of Science Teacher Education.

Moody, T. E. R. (2011). *An analysis of the effects of a five-day professional development experience on in-service k-12 teachers'content knowledge and self-reported confidence in their ability to teach astronomy content* Doctoral Dissertation, Ball State University.

National Research Council (1996). *National Science Education Standards.* Washington DC: National Academies Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18.

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nussbaum, J. (1979). Children's conceptions of the earth as a cosmic body: a cross-age study. *Science Eduction,* 63, 83–93.

Nussbaum, J. & Novak, J. (1976). An assessment of children's concepts of the earth utilizing structured interviews *Science Education, 60*, 535–50.

Parker, J. & Heywood, D. (1998) The earth and beyond: developing primary teachers' understanding of basic astronomical events', *International Journal of Science Education*, *20* (5), 503 - 520.

Partridge, B. & Greenstein, G. (2004). "Goals for "Astro 101": Report on Workshops for Department Leaders," *Astronomy Education Review*, *2*(2), 46. http://aer.noao.edu/cgi-bin/article.pl?id=64.

Prather, E. E., Rudolph, A. L., Brissenden, G., & Schlingman, W. M. (2009). A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *American Journal of Physics*, *77*(4), 320-330.

Prather, E., Slater, T. & Offerdahl, E. (2002). "Hints of a fundamental misconception in cosmology." *Astronomy Education Review*, *1*(2). http://aer.noao.edu/cgi-bin/article.pl?id=245.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., & Camburn, E. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study*. Ann Arbor: Consortium for Policy Research in Education, University of Pennsylvania.

Schneps, M., & Sadler, P. (1988). A private universe. Pyramid Films, Santa Monica, CA.

Shepard, L. A. (1991). Will national tests improve student learning?. *Phi Delta Kappan*, 232-238.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 4-14.
Slater (Parker), S. (2007). *The impact of a kinesthetic astronomy curriculum on high school students' conceptual understanding of the seasons*. Master's capstone. Montana State University. Bozeman, MT.
Slater, S. J., & Slater, T. F. (2013, December). Better Categorizing Misconceptions Using a Contemporary Cognitive Science Lens. In *AGU Fall Meeting Abstracts* (Vol. 1, p. 0758).
Slater, S. J., Slater, T. F., & Bailey, J. M. (2010). *Discipline-Based Education Research: A Scientist's Guide*. New York WH Freeman.
Slater, T. F., & Adams, J. P. 2003, *Learner-Centered Astronomy Teaching: Strategies for ASTRO 101*. Upper Saddle River, NJ: Prentice Hall.
Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational measurement: Issues and practice*, *10*(4), 7-11.
Sneider, C. B & Pulos, S. (1983). Children's cosmographies: Understanding the earth shape and gravity *Science Educator*. *67*, 205-221.
Stork, D. (2014). *Contemporary discipline-based astronomy education research study of K-12 teachers' astronomy knowledge using the Test Of Astronomy STandards*. Ph. D. Dissertation dissertation, University of Wyoming.
UCLA: Academic Technology Services, Statistical Consulting Group (2008). *SPSS FAQ: What does Cronbach's alpha mean?* from http://www.ats.ucla.edu/stat/spss/faq/alpha.html (accessed April 10, 2008.)
Tatge, C.B., Slater, S.J., & Slater, T.F. (2014, October). *Melhorando a educação em astronomia nos estados unidos pela investigação de outras culturas.* Paper presented at the Nacional De Educação Em Astronomia meeting of the Sociedade Astronômica Brasileira in Curitiba, Brazil.
Trumper, R. (2000). University students' conceptions of basic astronomy concepts. *Physics Education. 35*(1), 9-15.
Turkoglu, O., Ornek, F., Gokdere, M., Suleymanoglu, N., & Orbay, M. (2009). On pre-service science teachers' preexisting knowledge levels about basic astronomy concepts. *International Journal of Physical Sciences*, *4*(11), 734-739.
Vosniadou, S., & Brewer, W. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology. 24*, 535–85.
Williamson, K. E. (2013). *Development and calibration of a concept inventory to measure introductory college astronomy and physics students' understanding of Newtonian gravity* Ph. D. Dissertation Montana State University (Order No. 3608801). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (1496774510). Retrieved from http://search.proquest.com/docview/1496774510?accountid=14793
Wu, H. K., & Shah, P. (2004). Exploring visuospatial thinking in chemistry learning. *Science Education*, *88*(3), 465-492.
Zeilik, M. (2002). Birth of the astronomy diagnostic test: Prototest evolution. *Astronomy Education Review*, *1*(2), 46-52.
Zeilik, M., & Morris, V. J. (2003). An examination of misconceptions in an astronomy course for science, mathematics, and engineering majors. *Astronomy Education Review*, *2*(1), 101-119.

END NOTE

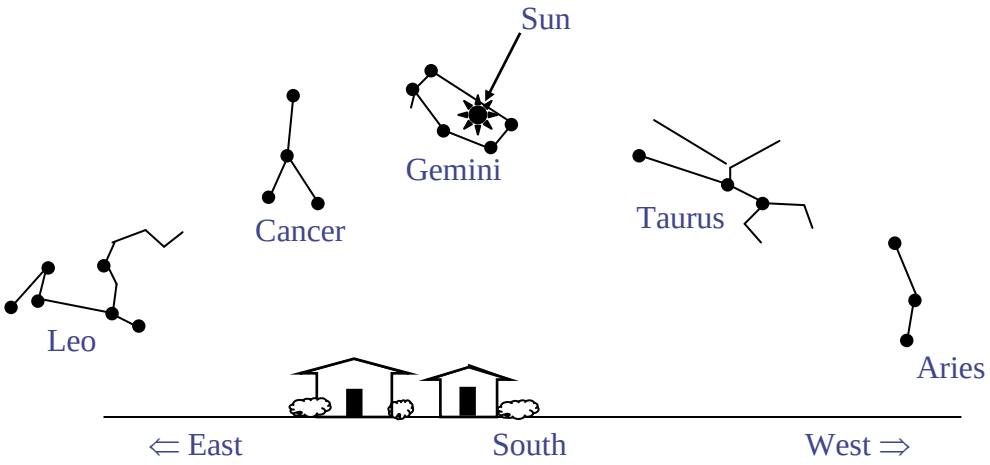
The toast appears on the following pages as an appendix. For a classroom ready photocopy, please contact the author.

APPENDIX

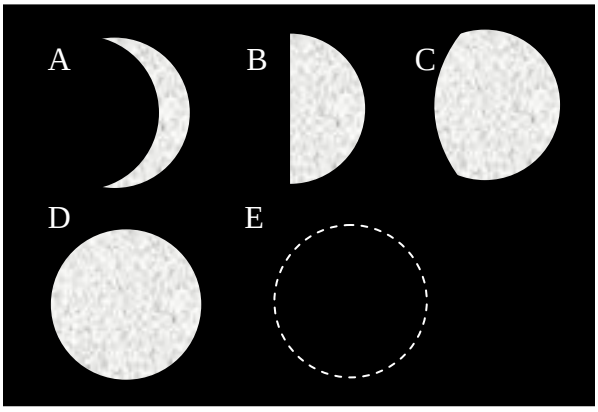
Use the drawing below to answer the next two questions.

1. If you could see stars during the day, the drawing above shows what the sky would look like at *noon* on a given day. The Sun is at the highest point that it will reach on this day and is near the stars of the constellation Gemini. What is the name of the constellation that will be closest to the Sun at sunset on this day?
   a. Leo
   b. Taurus
   c. Aries
   d. Cancer
   e. Gemini

2. This picture shows the position of the stars at *noon* on a certain day. How long would you have to wait to see Gemini at this same position *at midnight?*
   a. 12 hours
   b. 24 hours
   c. 6 months
   d. 1 year
   e. Gemini is never seen at this position at midnight.



3. You look to the eastern horizon as the Moon first rises and discover that it is in the new moon phase. Which picture shows what the moon will look like when it is at its high point in the sky, later that same day?
   a. A
   b. B
   c. C
   d. D
   e. E



4. You are located in the continental U.S. on the first day of October. How will the position of the Sun at noon be different two weeks later?
   a. It will have moved toward the north.
   b. It will have moved to a position higher in the sky.
   c. It will stay in the same position.
   d. It will have moved to a position closer to the horizon.
   e. It will have moved toward the west.

5. Which sentence best describes why the Moon goes through phases?
   a. Earth's shadow falls on different parts of the moon at different times.
   b. The moon is somewhat flattened and disk-like. It appears more or less round depending on the precise angle from which we see it.
   c. Earth's clouds cover potions of the moon resulting in the changing phases that we see.
   d. The sunlight reflected from earth lights up the moon. It is less effective when the moon is lower in the sky than when it is higher in the sky.
   e. We see only part of the lit-up face of the Moon depending on its position relative to Earth and the Sun.

6. Imagine you see Mars rising in the east at 6:30 pm. Six hours later what direction would you face (look) to see Mars when it is highest in the sky?
   a. Toward the north
   b. Toward the south
   c. Toward the east
   d. Toward the west
   e. Directly overhead

7. Imagine that Earth was upright with no tilt. How would this affect the seasons?
   a. We would no longer experience a difference between the seasons.
   b. We would still experience seasons, but the difference would be *less* noticeable.
   c. We would still experience seasons, but the difference would be *more* noticeable.
   d. We would continue to experience seasons in essentially the same way we do now.

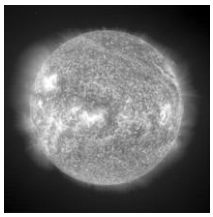
8. How does the Sun produce the energy that heats our planet?
   a. The gases inside the sun are burning and producing large amounts of energy.
   b. Gas inside the sun heats up when compressed, giving off large amounts of energy.
   c. Heat trapped by magnetic fields in the sun is released as energy.
   d. Hydrogen is combined into helium, giving off large amounts of energy.
   e. The core of the Sun has radioactive atoms that give off energy as they decay.

9. The Big Bang is best described as:
   a. The event that formed all matter and space from an infinitely small dot of energy.
   b. The event that formed all matter and scattered it into space.
   c. The event that scattered all matter and energy throughout space.
   d. The event that organized the current arrangement of planetary systems.

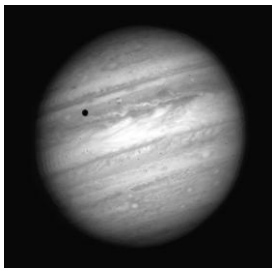
10. Which of the following ranks locations, from <u>closest to Earth to farthest from Earth</u>?
    a. the sun, the moon, the edge of our solar system, the north star, the edge of our galaxy
    b. the sun, the north star, the moon, the edge of our galaxy, the edge of our solar system
    c. the moon, the north star, the sun, the edge of our solar system, the edge of our galaxy
    d. the moon, the sun, the edge of our solar system, the north star, the edge of our galaxy
    e. the north Star, the Moon, the Sun, the edge of our galaxy, the edge of our solar system

Consider the six different astronomical objects (A-F) shown below.

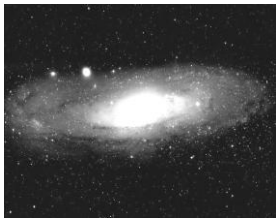
A. The Solar System



B. The Sun



C. Jupiter



D. Andromeda

E. Galaxy Cluster

F. Nebula

11. Which of the following is the best ranking (from smallest to largest) for the size of these objects?
    a. c<f<b<a<d<e
    b. e<d<f<a<b<c
    c. c<b<a<f<d<e
    d. f<c<b<a<d<e
    e. None of the above is correct.

12. Imagine that Earth's orbit were changed to be a perfect circle about the Sun so that the distance to the Sun never changed. How would this affect the seasons?
    a. We would not be able to notice a difference between seasons.
    b. The difference in the seasons would be *less* noticeable than it is now.
    c. The difference in the seasons would be *more* noticeable than it is now.
    d. We would experience seasons in the same way we do now.

13. 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.

14. Which one property of a star will determine the rest of the characteristics of that star's life?
    a. Brightness
    b. Temperature
    c. Color
    d. Mass
    e. Chemical makeup

15. Current evidence about how the universe is changing tells us that
    a. we are near the center of the universe.
    b. galaxies are expanding into empty space.
    c. groups of galaxies appear to move away from each other.
    d. nearby galaxies are younger than distant galaxies.

16. 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.

17. When the Sun reaches the end of its life, what will happen to it?
    a. It will turn into a black hole.
    b. It will explode, destroying earth.
    c. It will lose its outer layers, leaving its core behind.
    d. It will not die due to its mass.

18. If you were in a spacecraft near the Sun and began traveling to Pluto you might pass
    a. planets.
    b. stars.
    c. moons.
    d. two of these objects.
    e. all of these objects.

19. How did the system of planets orbiting the Sun form?
    a. The planets formed from the same materials as the sun.
    b. The planets and the sun formed at the time of the big bang.
    c. The planets were captured by the sun's gravity.
    d. The planets formed from the fusion of hydrogen in their cores.

20. Which of the following would make you weigh half as much as you do right now?
    a. Take away half of the earth's atmosphere.
    b. Double the distance between the sun and the earth.
    c. Make the earth spin half as fast.
    d. Take away half of the Earth's mass.

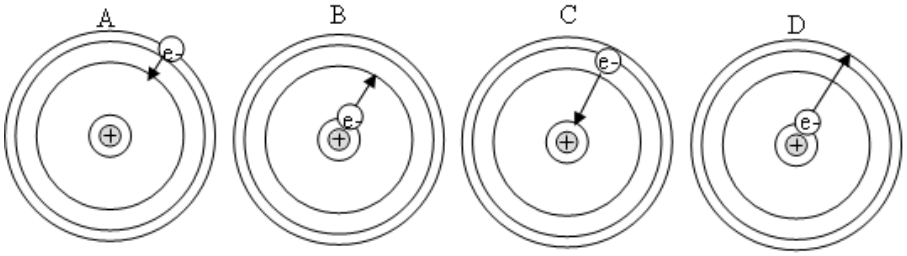
21. Astronauts "float" around in the space shuttle as it orbits Earth because
    a. there is no gravity in space.
    b. they are falling in the same way as the space shuttle.
    c. they are above earth's atmosphere.
    d. there is less gravity inside of the Space Shuttle.

22. Energy is released from atoms in the form of light when electrons
    a. are emitted by the atom.
    b. move from low energy levels to high energy levels.
    c. move from high energy levels to low energy levels.
    d. move in their orbit around the nucleus.

23. Which of the following would be true about comparing visible light and radio waves?
    a. The radio waves would have a lower energy and would travel slower than visible light.
    b. The visible light would have a shorter wavelength and a lower energy than radio waves.
    c. The radio waves would have a longer wavelength and travel the same speed as visible light.
    d. The visible light would have a higher energy and would travel faster than radio waves.
    e. The radio waves would have a shorter wavelength and higher energy than visible light.

24. The atoms in the plastic of your chair were formed
    a. in our sun.
    b. by a star existing prior to the formation of our Sun.
    c. at the instant of the Big Bang.
    d. approximately 100 million years ago.
    e. in a distant galaxy in a different part of the early universe.

Use the drawings below to answer the next <u>two</u> questions.

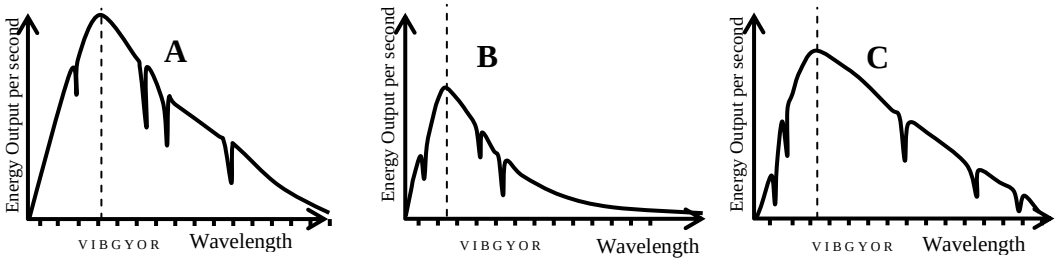


25. Which atom would be absorbing light with the greatest energy?
    a. A
    b. B
    c. C
    d. D

26. Which atom would emit light with the shortest wavelength?
    a. A
    b. B
    c. C
    d. D

---

27. The graphs below illustrate the energy output versus wavelength for three unknown objects A, B, and C. Which of the objects has the highest temperature?



    a. A
    b. B
    c. C
    d. All three objects have the same temperature.
    e. The answer cannot be determined from this information.