A DIRECT EXAMINATION OF COLLEGE STUDENT MISCONCEPTIONS IN ASTRONOMY. II. VALIDITY OF THE ASTRONOMY BELIEFS INVENTORY

Andrej Favia[1], Neil F. Comins[1], and Geoffrey L. Thorpe[2], *University of Maine*

Received January 22, 2015; Revised March 12, 2015; Accepted April 10, 2015

Abstract: This is the second in a series of papers in which we examine the persistence of 215 common misconceptions using a new instrument, the Astronomy Beliefs Inventory. In the first paper, we showed that the data in the ABI meet the statistical criteria for an internal validity analysis. In this paper, we examine the concurrent validity of the ABI by comparing student responses from the instrument with data from four instructional instruments, including multiple-choice exams and free-response questions, administered in their introductory astronomy course during the Fall 2013 semester. Two of the instruments were administered prior to instruction on the associated topics. We show that the validity of the ABI is supported through significant correlations with all four of the instruments under consideration.

Keywords: Students - Non-science Majors - General Astronomy - Astro 101 - Misconceptions - Undergraduate Education

INTRODUCTION

Misconceptions about astronomy are acquired at all ages through a variety of incorrect thought processes and incorrect information provided by the media and other sources (e.g., Comins, 2001). Children, who lack background information and experience thinking logically and scientifically, are especially at risk of accepting misconceptions as fact. In their study, Smith III, diSessa, and Roschelle (1993) observed that misconceptions persist even after instruction, and that misconceptions "have a strong influence on how student learning is currently evaluated." That is to say, misconceptions (which comprise deeply held incorrect beliefs) interfere with learning. The persistence of misconceptions about astronomy are so widespread that many researchers are still publishing studies that recognize the prevalence of misconceptions and their influences on student learning (Bailey & Slater, 2004; Bailey, Prather, Johnson, & Slater, 2009; Comins, 2001; Lombardi, Sinatrab, & Nussbaum, 2013, and references therein).

Various researchers tend to interpret misconceptions from the standpoint of how they think students learn. For example, Vosniadou (1994) suggests that learning involves the strategic refinement of one's own mental models. In their study, Bailey et al. (2009) refer to "the constructivist movement of the late 20th century" (p. 2), a paradigm in which effective instruction requires building upon pre-instructional student ideas. Teaching from a constructivist framework supports the use of inquiry-based tutorials and interactive clicker questions, with class-wide emphasis on group interactions (Slater & Adams, 2002). As Smith III et al. (1993) state, learning "involves the acquisition of expert concepts and the dispelling of misconceptions" (p. 122). Bransford (2000) emphasizes that the constructive process may enhance the students' post instruction understanding of astronomical concepts. Constructivism

[1] Department of Physics and Astronomy
[2] Department of Psychology

thus takes place at the intersection of what students may already know and how students can construct more sophisticated models.

Multiple researchers have alternate, often loosely-worded, definitions of the term "misconception." As examples, Sadler (1998) defines misconceptions as "conceptions that are quite different from their teachers' or those of scientists." Vosniadou (1994) defines misconceptions as "attempts to reconcile scientific information with false information regarding how one observes the world." Comins (2001) presents a list of definitions for the term "misconception," which he developed after asking researchers at a conference on misconceptions in science and mathematics in 1994. Comins presents a list of thirteen such definitions (p. 55), which include: "any belief that is untrue," "a mistaken idea," "an incorrect mental construct," "a misunderstanding," "only incorrect beliefs that are deep-seated in our minds," "a conception or belief that produces a systematic pattern of errors," and others. For our study, we adopt the definition of misconception used by Comins: "any deeply held belief that is inconsistent with currently accepted scientific concepts. Deeply held beliefs are distinctly different from superficial ones, such as details we might memorize for an exam and then promptly forget" (p. 56). This definition is consistent with the premise that misconceptions are deep seated and avoids grounding the term in a particular learning framework. The interpretation of misconceptions being *deep-seated* is also supported by the results of other studies on misconceptions in astronomy (Sadler, 1998; Bailey et al., 2009).

The analysis of the data for this paper concerns the college-level astronomy course for non-majors, as taught at the University of Maine during the Fall 2013 semester. The class is taught in the style of a modified lecture. Common misinformation held by students about a relevant topic is presented in class, after which an explanation is provided to demonstrate why this information is, in fact, wrong. Students are also asked to respond to misconception-based, free-response "attendance questions" on a topic to be addressed in the subsequent class. For example, if the question is "How many zodiac constellations are there?" then the subsequent lecture would include a discussion about zodiac constellations. The answer is 13, not 12, as is often commonly thought by students in the course, in part because of the common association of "12 months for the 12 zodiac constellations." The subsequent lecture would have extra instruction on (i) the duration of each zodiac, which is not 30 days, and (ii) why there are *not* 12 zodiac constellations. Extra credit is also awarded to students for identifying their own misconceptions, as well as for identifying misinformation presented in other current sources (e.g., in a television show). We note that the course, as taught during the Fall 2013 semester, did not utilize inquiry-based tutorials. About half of the students in the course are non-science majors, and given the discussion by Clark et al. (2012), perhaps only the very top students in the class would have benefited more in learning the same scope of topics taught in the class from a constructivist-oriented learning environment.

We introduced the Astronomy Beliefs Inventory (ABI) in 2014 as a new instrument to explore the persistence of misconceptions (Favia, Comins, Thorpe, & Batuski, 2014, hereafter, Paper I). The ABI is a comprehensive inventory of 215 misconceptions held by college students taking an introductory-level course in astronomy. According to Strike and Posner (1992), conceptual change requires that the learner become dissatisfied with current concepts, and that learning occurs when a new conception seems to solve the current problem and encourages new ways of looking at natural phenomena. In other words, students have to be presented with strong evidence to contradict their preconceived notion, and then believe the new information. Because the ABI asks students to report what they believe for a variety of topics, the ABI can cover a broad range of misconceptions very efficiently.

The ABI is unique in that it avoids the conventional pre/post-test approach in learning assessment by including a retrospective component. The design of retrospective studies, however, is subject to some issues regarding reliability. Briefly, memory is a reconstructive process (Olson & Cal, 1984). As Henry, Moffitt, Caspi, Langley, and Silva (1994) note, a

retrospective approach may be of questionable validity in some contexts, notably in the recall of personally-significant emotional and psychosocial material. The authors suggest that "the use of retrospective reports should be limited to testing hypotheses about the relative standing of individuals in a distribution" (p. 92). A comprehensive review by Brewin, Andrews, and Gotlib (1993), however, "suggests that claims concerning the general unreliability of retrospective reports are exaggerated" (p. 82). The likelihood of inaccurate responses may be significantly reduced by asking subjects to provide reports on a timeline for abandoning misconceptions. Additionally, the ABI consists of very brief statements (e.g., "Saturn's rings are solid") each of which enhances the recall of one's own past beliefs, compared to longer statements with the same content. These latter versions are more likely to invoke additional memories that might obfuscate the recollection.

The ABI probes whether or not students believe the presented misconceptions. As with all tests, responses are affected by how accurately students fill in the correct bubbles on bubble sheets and how they interpret the statements in the inventory. About 10 departmental faculty with training in physics education research evaluated a pool of 235 items, and after receiving feedback, 20 items were discarded due to issues with clarity in the wording, leaving the 215 statements that we chose to constitute the ABI. In Paper I, we showed that students provided the correct bubble and statement interpretation over 90% of the time for a random subset of five of the ABI statements. Compared to traditional surveys, the ABI has the additional aspect of asking students when (before or during college) they learned a misconception. The ABI measures what students believe after instruction in class (anywhere between a week and several months, depending on the topic). The administration of a delayed post-test has been employed elsewhere in the literature to study student preconceptions in astronomy (Lombardi et al., 2013; Prather et al., 2004).

The one-semester, survey astronomy course upon which our results are based was taught at the University of Maine using the latest editions of the textbooks *Discovering the Universe* (Comins & Kaufmann III, 2012) or *Discovering the Essential Universe* (Comins, 2013). The ABI was administered to students between the last lecture and the final exam. Statements in the ABI were labeled, e.g., sA46, for statement number 46. (We omitted the labels when we administered the ABI in class.) Previously, the ABI was written and administered during different semesters in either of two formats: (i) as all inaccurate statements, from Fall 2009 to Fall 2012, or (ii) as a random distribution of scientifically correct and inaccurate statements, during the Spring 2013 and Fall 2013 semesters. The labels of those statements that were rephrased from incorrect to correct inherited an "n" at the end. For example, sA46 is "Jovian planets (Jupiter, Saturn, Uranus, Neptune) have solid surfaces," whereas sA46n, in Appendix A, is "Jovian planets (Jupiter, Saturn, Uranus, Neptune) have gaseous or icy surfaces."

For either version, the students were asked to indicate when they first heard of each statement, if ever, and indicate if they used to believe them only as a child, adolescent, or as a result of taking the course. See Paper I for details. We found that the ABI is not biased in terms of the order of statement presentation or its relatively long length (215 statements). However, we found that the different formats of the inventory (all incorrect statements vs. a mixture of correct and incorrect statements) created a statistically significant difference in the fraction of misconceptions that students believed by the end of the course. This means that we would get a more accurate representation of misconception endorsement using correct/incorrect questions, which we did. As we will show in a future paper, however, *correlations* between misconceptions (the tendency that if you believe X, you also believe Y, and vice versa) are relatively unaffected by the different formats.

METHOD

The ABI was administered in six semesters at the University of Maine from Fall 2009 through Fall 2013, with a total of 639 subjects in the overall study. In this paper, we focus exclusively on data from the Fall 2013 semester (N = 110, of which 67 students are male, and 43 students are female). The overall age of the sample is 19.6 years, with a median of 19, a mode of 18, and a range from 18 to 41. The subjects' ethnicities are 82.7% Caucasian, 2.73% Hispanic, 1.82% Asian, and 12.7% other/unspecified.

The version of the inventory used in the Fall 2013 semester consisted of both true and false statements. The course was taught by Neil F. Comins (hereafter NFC). As noted above, when taking the ABI, students were asked to indicate if they ever believed each of the statements. Students were also given the option to report if they first believed a statement prior to instruction in college. In this paper, we examined the validity of the data provided by the students (i.e., are they trying to recall their beliefs or are they providing incorrect reports) by comparing the ABI responses with data from four other instruments administered that semester: (i) a cumulative set of three prelims (standard course exams) in the course each covering different topics that span the material presented in the course, with each prelim administered prior to the ABI, (ii) the final exam, administered a few days after the ABI, (iii) a special pretest on black holes and galaxies, administered prior to the presentation of this material in class, and (iv) written responses to select "attendance questions" asked by NFC in class, as described in the introduction. Attendance questions were administered at the end of each class (except during a test day). Table 1 presents the chronology of test administration during the Fall 2013 semester, with the exception of the attendance questions.

Table 1
Chronology of test administration during the Fall 2013 semester

Date	*Test*
October 3	Prelim 1 (standard course exam 1)
October 31	Prelim 2 (standard course exam 2)
October 31	BHGP (administered after Prelim 2)
December 5	Prelim 3 (standard course exam 3)
December 13	ABI
December 17	Final Exam

The framework in which we test the validity of the ABI is the method of concurrent validity. As described by Zumbo and Rupp (2004), "one of the current themes in validity theory is that construct validity is the totality of validity theory and that its discussion is comprehensive, integrative, and evidence based" (p. 84). Those authors suggest that there have been few advances in validity theory in recent years, and the emphasis is now on the consequences of testing, essentially meaning criterion-referenced validity. Kline (2005) adds that "one common way to assess the utility of test scores is to use them to predict other variables of interest" (p. 203). We test for concurrent validity in the ABI, using the approach suggested by Kline, by measuring the correlation between the data in the ABI and the data from each of the other four instruments, one at a time. The statistic that we use to determine the measure of the correlation is the Pearson product-moment coefficient (Onwuegbuzie, Daniel, & Leech, 2007), which we define as the correlation coefficient r. A test of the correlation coefficient (i)

measures the extent that two variables are linearly related, and (ii) calculates the significance of the correlation by reporting the p -value. The p -value is the probability that a result at least as extreme as observed happens by chance from a random selection of the available data. A correlation coefficient of $r = 1$ represents a perfect linear relationship, and values between $0 < r < 1$ indicate a positive correlation. We define a correlation as statistically significant if $p < .05$. For variables with dichotomous scoring (e.g., correct vs. incorrect), this correlation is called a phi coefficient.

The internal consistency of a set of data is reported by coefficient alpha (α), , sometimes called Cronbach's alpha, which is a measure between 0 and 1 indicating how closely related a set of variables are in a group (Schmitt, 1996). The value $\alpha \approx 0$ indicates dissimilar items, whereas the value $\alpha \approx 1$ indicates very similar items, with $\alpha = 1$ referring to two or more identical data sets. Values of $\alpha \geq .70$ represent a group of items with "adequate" internal consistency. If the group consists of items pooled together from different topics, then it is possible for the pooled data to exhibit lower consistency than when comparing the data within only one topic. The effect is that α could be lower for the pooled group than for the data in the sets taken individually.

As described in Paper I, a master code was developed for all responses to the ABI, to indicate relative degrees of misconception retainment. We define retainment as "the tendency for students to hold on to a misconception from either their childhood or during some point in the course" (Favia et al., 2014). Three of the four instruments (prelims, final exam, and the pretest on black holes and galaxies) are scored dichotomously, with "1" for a correct answer, and "0" for an incorrect answer. As we discuss in the section on attendance, originally three categories were established for each attendance question. For the purpose of assessing the concurrent validity of the ABI, we reduced the number of attendance question categories down to two, based simply on "correct" and "incorrect" answers. Effectively, all four instruments are scored dichotomously.

Table 2
Codes for comparing the data between the ABI and each instrument

Prelims	0: disbelieved correct statement or endorsed inaccurate statement 1: endorsed correct statement or disbelieved inaccurate statement
Final Exams	0: disbelieved correct statement or endorsed inaccurate statement 1: endorsed correct statement or disbelieved inaccurate statement
Black Holes and Galaxies Pretest	0: disbelieved correct statement or endorsed inaccurate statement, prior to college 1: endorsed correct statement or disbelieved inaccurate statement, prior to college
Attendance Questions	0: disbelieved correct statement or endorsed inaccurate statement, prior to college 1: endorsed correct statement or disbelieved inaccurate statement, prior to college

To make comparisons between each of the four instruments and the ABI, we also coded the data in the ABI into two categories. The ABI was administered after instruction in the course, as were the prelims (one for each set of topics) and the final exam. To make post-

instruction comparisons between the ABI and either of these tests, we recoded the ABI data into "1" if the student endorsed a scientifically-correct statement or disbelieved an inaccurate statement, and "0" if the student disbelieved a scientifically-correct statement or endorsed an inaccurate statement. Using the data from all students, we then correlated the scores (0 or 1) for each item on the test with the scores for the associated item on the ABI. On the other hand, the pretest and the attendance questions were administered prior to instruction on their respective topics. To make pre-instruction comparisons between the ABI, which was administered at the end of the course, and these two pre-instructional tests, we recoded the ABI data into "1" if the student claimed to believe a scientifically-correct statement or disbelieve an inaccurate statement at a time prior to college, and "0" if the student claimed to disbelieve a scientifically-correct statement or believe an inaccurate statement at a time prior to college. Table 2 presents a summary of the codings used for comparing the data between the ABI and each instrument.

RESULTS

Course prelims

Table 3
Correlations between prelim questions and their associated ABI statements

Prelim	*Question*	*ABI Statement*	*r*	*p*
1	14	sA50n	-.041	.670
1	38	sA80	.255	.007*
2	2	sA164n	.187	.051
2	25	sA50n	-.032	.740
2	30	sA172	.069	.474
2	39	sA76	.252	.008*
2	41	sA46n	-.040	.678
3	7	sA150	-.084	.384
3	7	sA222n	-.037	.703
All questions		All statements	.239	.012*

Three prelims (standard course exams, 50 multiple-choice questions each with five answers) were administered in AST 109. For our purposes, the prelims serve as a check on the consistency of one's recollections, because some prelim questions are based on the common misconceptions explored in the ABI. Our expectation is that students who answer a prelim question correctly ought to endorse the correct science for the associated ABI statement. The prelim questions are scored as "correct" and "incorrect." We initially identified 12 prelim questions associated with one or more associated ABI statements. Cronbach's alpha for the prelim questions was 0.48. Because the prelims are just course exams, they are not optimally designed as research instruments. To improve the reliability of the test, we removed four questions (one of which was associated with two ABI statements) to bring alpha up to 0.52. As discussed earlier, we expect a low value for alpha, because the prelim questions under consideration are pooled together from a broad variety of topics, as would be expected for a multiple-choice exam, and so are not generally expected to correlate internally with each other. In total, we made nine comparisons between prelim questions and their associated ABI

statements. A list of ABI statements and their associated prelim questions is presented in Appendix A.

Table 3 presents the correlations between prelim questions and their associated ABI statements, using the method described. The first and second columns, respectively designated "Prelim" and "Question," identify the prelim (which of the first three exams) and the question number from that exam. The third column, designated "ABI Statement," identifies the statement from the ABI that is associated with the prelim question. The correlation coefficient, r, and the p-value for each comparison are reported in the fourth and fifth columns, respectively. Statistically significant p-values are less than .05 and are marked with an asterisk. The final row compares the mean score on all prelim questions under consideration to the mean score on all related ABI statements. We remind the reader that statement labels ending in "n" are phrased as scientifically correct statements, and we have incorporated this into our coding scheme for consistency.

Table 3 shows that there is an overall significant correlation between prelim scores and ABI scores (p = .012). When comparing one question at a time, weak correlations are to be expected, since the data are scored dichotomously. However, when comparing the mean score on all prelim questions under consideration to the mean score on all related ABI statements, we observe a significant correlation, which suggests that student responses to the ABI partially reflect how students perform on the prelims. This result provides support for the validity of the ABI.

Final Exam

The final exam (100 multiple-choice questions each with five possible answers) serves as an additional check for the consistency of one's recollections, because some misconceptions are also shared between the final exam and the ABI. Our expectation is that students who endorse a misconception on the ABI would select an incorrect answer on the associated final exam question. The final exam questions are scored as "correct" and "incorrect." We initially identified 17 final exam questions associated with one or more associated ABI statements. Cronbach's alpha for the prelim questions was .50. To improve the reliability of the test, we removed five questions (two of which were associated with two ABI statements) to bring alpha up to .57. Again, alpha is low because of the variety of topics covered on the final exam. In total, we made 13 comparisons between final exam questions and their associated ABI statements. A list of ABI statements and their associated final exam questions is presented in Appendix B.

Table 4 presents the correlations between final exam questions and their associated ABI statements, using the method described. The first column, designated "Question," identifies the question number from the final exam. The second through fourth columns follow the same format as the third through fifth columns in Table 3. Table 4 shows that there is an overall significant correlation between final exam scores and ABI scores (p = .034), which suggests that student responses to the ABI partially reflect how students perform on the final exam. This result provides further support for the validity of the ABI.

In order to develop more detailed insights into certain pre-instructional beliefs, we next examine student perceptions of black holes and galaxies.

Table 4
Correlations between final exam questions and their associated ABI statements

Question	*ABI Statement*	*r*	*p*
3	sA111n	.225	.020*
9	sA2	.152	.116
10	sA185	.135	.165
21	sA46n	-.065	.504
22	sA172	.178	.065
23	sA171	.275	.004*
24	sA8	.336	<.001*
40	sA248	.186	.054
44	sA108n	.214	.027*
44	sA170	.188	.053
51	sA208	.097	.317
67	sA50n	.032	.747
98	sA262	-.037	.702
All questions	All statements	.204	.034*

Black Holes and Galaxies Pretest

Motivated by *a priori* knowledge of student misconceptions about black holes and galaxies, we developed an 18 multiple-choice question test, called the Black Holes and Galaxies Pretest (BHGP), specifically for the Fall 2013 semester. The BHGP was administered without penalty of course credit lost for incorrect responses. The design of the BHGP followed the "distractor driven" model by Sadler (1998), in the sense that we designed the incorrect answers to distract students from endorsing the correct information by prompting memory recall of their prior misconceptions. The questions and response options to the BHGP are reported in Appendix C. No posttest based on these questions was administered; instead, the ABI, administered at the end of the semester, served as the posttest, since responses to the associated statements about black holes and galaxies can be used to check the reliability of self-reports. The percent of correct responses for each question of the BHGP are presented in Table 5. The associated statements on the ABI to each question are also presented. The overall fraction of questions on the BHGP that students answered correctly was 44%.

To test for the concurrent validity between the BHGP and the ABI, we created two separate sets of questions, one set for the black holes questions, and one set for the galaxies questions. For each set of questions independently, we determined Cronbach's alpha, then improved the reliability of each set by removing selected questions. We then correlated the data from the questions in each reduced set with their associated ABI statements. Because the BHGP is a pre-instructional instrument, we recoded the data in the ABI in the manner presented in Table 2. Specifically, we coded the data to reflect whether or not students reported that they believed a misconception prior to instruction in college.

For the set of black hole questions, Cronbach's alpha was .249. To improve the reliability of the test prior to correlating the data between the black hole questions and the ABI, we removed three questions (one of which was associated with two ABI statements) to bring

alpha up to .408, which is still low. We attribute the low value of alpha to (i) the short length of the test and (ii) the different sets of black hole qualities, as presented within the test. For instance, questions 1, 3, and 7 pertain to the creation, composition, and detection of black holes, respectively, whereas questions 6 and 8 ask students to consider how black holes affect objects in their vicinity. These sets of qualities convey different aspects about black holes, and by asking about them together in the same test, the internal consistency of the test was lowered. In total, we made seven comparisons between black hole questions and their associated ABI statements.

Table 5
Black Holes and Galaxies Pretest questions, the percent of students who answered correctly, and the associated statement(s) in the ABI

ABI Statement(s)	*% Correct*	*Question*
sA232	40%	1. How are black holes created today?
sA233n, sA245, sA247	30%	2. What is the fate of black holes?
sA235, sA237n	73%	3. Black holes consist of:
sA246	14%	4. As detectable from our universe, what shape does a black hole have?
sA238	9%	5. A spacecraft is put into circular orbit around a distant black hole. Which one of the following outcomes will happen to the spacecraft?
sA240n	67%	6. What would happen to an asteroid that passed into a black hole?
sA242n	70%	7. How do astronomers detect black holes?
sA243n, sA244	65%	8. If we boarded a spaceship to journey into a black hole, what would happen to us?
sA218n, sA225n	74%	9. How many galaxies exist in the universe today?
sA219n	12%	10. Relative to the Milky Way galaxy, the solar system is located:
sA220	42%	11. How many distinct shapes do galaxies have?
sA221, sA224	53%	12. Where in the universe is the Milky Way located today?
sA222n	18%	13. Where is the Sun located relative to the Milky Way?
sA226	20%	14. How are galaxies distributed throughout the universe?
sA227n	19%	15. Which statement about observing stars in the Milky Way is most accurate?
sA228n	80%	16. Which one statement about galaxy properties is correct?
sA230n	82%	17. Which statement most accurately describes the contents of the Milky Way?
sA231n	34%	18. Which one of the following is true about the Milky Way?

Table 6 presents the correlations between the black hole questions and their associated ABI statements, using the method described earlier. The first column, designated "Question,"

identifies the question number from the BHGP. The second through fourth columns follow the same format as those in Table 4.

Table 6
Correlations between black hole questions and their associated ABI statements

Question	*ABI Statement*	*r*	*p*
1	sA232	.275	.004*
3	sA235	.139	.152
3	sA237n	.142	.143
6	sA240n	.209	.030*
7	sA242n	.124	.202
8	sA243n	.106	.278
8	sA244	.078	.424
All questions	All statements	.348	<.001*

Table 6 shows that there is a significant correlation between black hole question scores and ABI scores ($p < .001$), which suggests that student responses to the ABI partially reflect students' preconceived notions about black holes.

We then proceeded to analyze the galaxy questions in the same manner. For the set of galaxy questions, Cronbach's alpha was .435. To improve the reliability of the test prior to correlating the data between the galaxy questions and the ABI, we removed five questions (one of which was associated with two ABI statements) to bring alpha up to .534. In total, we made six comparisons between galaxy questions and their associated ABI statements. Table 7 presents the correlations between the galaxy questions and their associated ABI statements, using the method described. The first column, designated "Question," identifies the question number from the BHGP. The second through fourth columns follow the same format as those in Table 4.

Table 7
Correlations between galaxy questions and their associated ABI statements

Question	*ABI Statement*	*r*	*p*
11	sA220	.270	.005*
12	sA221	.246	.010*
12	sA224	.213	.027*
16	sA228n	.114	.241
17	sA230n	.100	.307
18	sA231n	.195	.044*
All questions	All statements	.343	<.001*

Table 7 shows that there is a significant correlation between galaxy question scores and ABI scores ($p < .001$), which suggests that student responses to the ABI partially reflect students' preconceived notions about galaxies. This result, combined with the result of Table 6,

holds the promise that the ABI could be administered as a pre-instructional evaluation of student beliefs on black holes and galaxies.

Attendance Questions

As discussed in the introduction, at the end of each class, NFC asks a misconception-based question (for attendance credit) about a topic to be lectured in the subsequent class day. A select set of these "attendance questions" serves as yet another pre-instruction probe for checking the consistency of student recollections of their past misconceptions. Appendix D presents the attendance questions under consideration, their specific categories, their associated ABI statements, the number of student responses to each attendance question, and the fraction of student responses to each category of answers. Included with the set of attendance questions in Appendix D is Fleiss' kappa (κ) for each question. A brief discussion of the meaning of Fleiss' κ follows shortly. Occasionally two ABI statements would overlap with one attendance question. In total, we made 12 comparisons between attendance questions and their associated ABI statements. For each attendance question, we determined categories that best capture the correctness of the responses, as described in Table 8.

Table 8
Attendance question response categories

Category	*Representation*
1	a typical correct response by a student
2	a typical incorrect response that is associated with one or more common misconceptions, such as those already present in the ABI
3	all other "incorrect" responses

We checked the consistency of the categories by performing an analysis of inter-rater reliability (Elder Jr., Pavalko, & Clipp, 1993, pp. 42–44) on four of the nine attendance questions selected at random. For a particular attendance question, one of us scored a sample of written responses, then presented the same sample to an independent rater who did the same scoring without seeing the results of the first rater. For each analysis, a subset of 25 responses was selected at random. The agreement between the two raters, represented by Fleiss' κ, is given by

$$\kappa = \frac{\text{overall agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

where $\kappa > .60$ represents at least substantial agreement (Landis & Koch, 1977). For our analysis, κ ranged from 0.61–1.00.

Three categories were established for each attendance question. For the purpose of assessing the concurrent validity of the ABI, we now reduce the number of attendance question categories down to two, in which we coded score "1" for the correct answer and score "0" for an incorrect answer. We then correlated the scores on the attendance questions with the scores on the associated ABI statements. Because the attendance questions serve as a pre-instructional instrument, we recoded the data in the ABI in the manner presented in Table 2. Specifically, we coded the data to reflect whether or not students reported that they believed a misconception prior to instruction in college. Table 9 presents the correlations between attendance questions and their associated ABI statements, using the method described earlier. The first column,

designated "Question," identifies the question number from the list of attendance questions in Appendix D.

Table 9
Correlations between attendance questions and their associated ABI statements

Question	*ABI Statement*	*r*	*p*
1	sA2	.122	.219
2	sA112	.207	.041*
2	sA148	.277	.006*
3	sA50n	.118	.236
4	sA161	-.067	.508
5	sA46n	.066	.508
6	sA172	.342	.001*
7	sA189	.371	.001*
7	sA190n	.189	.086
8	sA248	-.051	.636
9	sA150	.265	.009*
9	sA222n	.289	.004*
All questions	All statements	.343	.001*

Table 9 shows that there is a significant correlation between attendance question scores and ABI scores ($p = .001$), which suggests that student responses to the ABI partially reflect the reports that students give on the attendance questions. This result not only provides further support for the validity of the ABI, but also holds the promise that the ABI could be administered as a pre-instructional evaluation of student beliefs in a college-level setting.

DISCUSSION

Comparison of Instruments

We have compared the ABI data to data from the course prelims, the final exam, a pretest on black holes and galaxies, and written responses to attendance questions. Our findings regarding the concurrent validity of the ABI in relation to these instruments are summarized in Table 10, which additionally reports the sample size n, the mean score M, and the standard deviation σ for these instruments, except for the attendance questions, which were administered weekly throughout the semester and whose responses were coded into categories. We remind the reader that for the BHGP, one test for each of the two topics was conducted, giving a total of five tests in four instruments.

Table 10
Correlations between the ABI and each of the four instruments. Summary statistics for the prelims are reported in the text

Instrument	*n*	*M*	σ	*r*	*p*
Prelims				.239	.012*
Prelim 1	167	66.0	16.0		
Prelim 2	165	58.8	16.4		
Prelim 3	161	61.2	17.6		
Final exam	161	60.1	12.9	.204	.034*
Black Holes and Galaxies Pretest					
Black holes	148	46.0	17.4	.348	<.001*
Galaxies	148	43.4	17.2	.343	<.001*
Attendance questions				.343	.001*

All four instruments provided support for the validity of the ABI based on statistical significance ($p < .05$). The prelims and final exams were administered after instruction, while the BHGP and each attendance question were administered prior to instruction on the relevant topics. Table 10 shows a significant correlation between data from the conventional exams (prelims and the final exam) and the data from the ABI. A significant correlation also exists between the data from free responses and the data from the ABI. However, there is only a weak correlation in the data between the BHGP and the ABI. We are thus confident that the validity of the ABI is supported through significant correlations with conventional tests given after instruction and free-response questions asked prior to instruction. We briefly discuss possible influences on the correlations reported in Table 10.

One influence on the nature of our observed correlations is the scoring of the data. The prelims, the final exam, and the BHGP are multiple-choice exams with all questions containing five options. The attendance questions, however, are free-response questions. The written responses must be categorized and rated. The response format of the ABI more closely mimics a multiple-choice exam than an essay-based exam, since a student completes the ABI by filling in a bubble sheet much like on a multiple-choice exam. One would thus initially expect that correlations may be stronger for multiple-choice exams than for the attendance questions. However, the BHGP is also "distractor driven" (Sadler, 1998), in that the questions were intentionally designed to prompt memory recall of misconceptions that students may harbor. We find that all four instruments produce significant correlations with the ABI, which supports our belief that tests of the concurrent validity of the ABI are not particularly sensitive to the nature of the questions on the instruments listed in Table 10.

An additional influence on the nature of the correlations is that a student who answers a question incorrectly on a prelim could learn the correct related science after taking a prelim and before taking the ABI. For instance, a student has anywhere between a few weeks and a few months (depending on the prelim) to learn the correct science before taking the ABI. The data would have been coded "0" on the prelim but coded "1" on the ABI, which would have caused a reduction in the observed correlation between the prelims and the ABI. Such a delay in administering the ABI would also make the observed correlation less significant. That we observed a significant correlation between the prelims and the ABI ($p = .012$) in our analysis already is promising, to say the least.

Our results motivate us to consider additional plausible influences on the reliability of one's own recollections. In particular, we considered two influences: (i) the student's childhood interest in astronomy, prior to instruction in the course; and (ii) the student's level of stress in the semester. We designed two surveys for the Fall 2013 semester to assess childhood interest in astronomy and stress. We emphasize that the purpose of these surveys is not to test the validity of the ABI, but rather to examine childhood interest and current stress level as *possible influences* on how students report their beliefs on the ABI at the end of one semester of college-level astronomy.

Childhood Interest in Astronomy

The Childhood Interest in Astronomy Survey (CIAS, see Appendix E) asks eight questions of the students to indicate their childhood interest in various activities related to astronomy. A total of 80 students taking AST 109 in the Fall 2013 semester completed the survey, which we posted online and made available only for the first two weeks of the course. For each question, students were asked to report their interest level for each activity, with the options for each question being: "never," "occasionally," and "very often." For example, in Question 1 of the CIAS, if a student has read astronomy books very often during one's own free time, the student responded to Question 1 with "very often." The responses were coded on a kind of Likert metric with "never" scored as 0, "occasionally" scored as 1, and "very often" scored as 2. The codes indicate relative measures of childhood interest. Total scores for each student fell within the range of 0 to 11, with a maximum possible score of 16. The mean and standard deviation of scores on the CIAS were 4.57 and 2.56, respectively, which suggests that the majority of students who took the CIAS had little more than an occasional interest in activities related to astronomy as a child.

One concern about the CIAS is that telescopes, planetaria, and observatories may not have been available to some students, who may have been genuinely interested in astronomy as children. For example, one's interest in astronomy may be influenced if a student has access to these instruments while still a child. We found that for the CIAS, $\alpha = .72$, which is considered adequate (Schmitt, 1996). The internal consistency of the responses would decrease if any one of the eight questions was removed from the survey, except for Question 3, which asks about how often the student used binoculars or a telescope, in which case α increases only marginally to .73. Our results support the hypothesis that we can use data from all eight questions without sacrificing the reliability of the data.

To examine if childhood interest in astronomy influences a student's endorsement of misconceptions after college instruction in astronomy, we compared the childhood interest in astronomy scores for all students with their scores for the entire ABI (that is, the mean of all 215 statements). We again recoded the ABI data into "1" if the student endorsed a scientifically-correct statement or disbelieved an inaccurate statement, and "0" if the student disbelieved a scientifically-correct statement or endorsed an inaccurate statement. The correlation between childhood interest in astronomy and mean ABI score is .265 ($p = .046$), which is statistically significant, because $p < .05$. The fact that the correlation, .265, is positive, is consistent with the notion that higher childhood interest in astronomy is associated with believing more of the correct science and fewer misconceptions.

While we observe a statistically-significant correlation between childhood interest in astronomy, as measured in the CIAS, and overall score on the ABI, we cannot determine if increased childhood interest in astronomy *causes* students to believe fewer misconceptions. To affirm the extent of causality, we would need to measure, in tandem, both the beliefs and interests in astronomy, of subjects still in their childhood, and we do not have these paired data. We thus cannot firmly conclude anything about the nature of causality between these two instruments. However, our result is at least consistent with the hypothesis that the more

interested children are in astronomy, the less likely they are to endorse misconceptions after college-level instruction in astronomy.

Stress

The results of the above section are consistent with our expectations that students who take an early-in-life interest in astronomy are more likely to recall their memories (which may or may not include misinformation) of what they learned compared to those students who were not interested in the subject material. On the other hand, memory recall is a reconstructive process, so one might expect that high levels of stress in one's life may interfere with the reconstructive process, thus causing inconsistencies in one's own past recall of their own misconceptions (Kuhlmann, Piel, & Wolf, 2005). In response to these considerations, we developed an instrument to evaluate the stress experienced by students at the time of completing the ABI. One such instrument, originally designed by Cohen, Kamarck, and Mermelstein (1983), asks students to report on their own measure of stress in general. For each question in the survey, subjects indicate the extent to which a particular facet represents them (e.g., how often one has felt overwhelmed). Scores for each question range from 0 to 4, where 0 means the subject has never experienced the facet, and 4 means the subject has very often experienced the facet. A related survey designed to assess one's own stress would be most appropriate to administer simultaneously with the administration of the ABI, which we did.

From the original instrument, we created a shortened survey, consisting of four particular questions that were determined by Geoffrey L. Thorpe as being most inter-correlated and comprising a sufficiently large proportion of the variance. (It would be unfair to the students to ask too many questions about their stress level on the same day as the ABI administration and only a few days before their final exams.) The set of these four questions comprises the Stress Survey (see Appendix F). A total of 108 students completed the Stress Survey. In Questions 1 and 4, higher scores indicate higher stress. In Questions 2 and 3, the metric is reversed, so that lower scores indicate higher stress. We found that for the Stress Survey, $\alpha = .79$. To quantify the responses on a kind of Likert metric, we considered the direction of the scores for all four questions, and ultimately we preserved the codes as originally presented to the students. The total stress score per student is $8 + Q1 + Q4 - Q2 - Q3$, where Q1, Q2, Q3, and Q4 are the scores for each of questions 1-4, respectively. (Note that the number of questions has nothing to do with the metric of the response codes.) Scores fell within the range of 0 to 16. The mean and standard deviation of scores on the Stress Survey were 6.63 and 3.47, respectively, indicating that at the time of administration of the ABI (only days before the final exam), the majority of students who took the Stress Survey reported moderate stress in their lives.

To examine if stress influences a student's endorsement of misconceptions after college instruction in astronomy, we compared the stress scores for all students with the *entire* ABI (that is, the mean of all 215 statements). The correlation between stress and mean ABI score is .049 ($p = .651$), which is not statistically significant. The point of the stress survey was to evaluate whether or not a student's stress level, as measured at the end of the course, has a significant influence on how students report their beliefs on the ABI. Our result does not support this hypothesis, and thus we conclude that there is no relationship between a student's stress level and their reports on the ABI.

CONCLUSIONS

In this paper we have tested the validity of the ABI by comparing the results we get from it with results from other testing instruments. These include prelims, the final exam, a pretest on black holes and galaxies, and attendance questions. We correlated the results from

one instrument at a time to those of the ABI. We have found that the validity of the ABI is supported through significant correlations with exams given after instruction, a distractor-driven pretest on black holes and galaxies, and free-response questions asked prior to instruction. We are confident that the ABI could be administered as a pre-instructional evaluation of student beliefs in a college-level setting. We have also discovered a statistically significant correlation between childhood interest in astronomy and the extent to which a student believes more of the correct science and fewer misconceptions after college instruction. We also found that there is no relationship between a student's stress level and their reports on the ABI.

In future papers, we will present the theoretical background for principal components analysis, a technique for identifying groups of correlated misconceptions, as the technique applies to our overall project. We will clarify the extent to which semester-to-semester variability in misconception endorsement influences *correlations* between misconceptions, and we will address the concern that the correlations are not significantly affected by the per-semester variability in misconception endorsement. Further papers will explore item response theory and, using it, we will propose an optimal sequence to teach concepts within individual topics in astronomy. 

The authors would like to acknowledge Jessica W. Clark, graduate student at the University of Maine, for volunteering her time as the second rater in our inter-rater analysis.

REFERENCES

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,* 010110.

Bailey, J. M., & Slater, T. F. (2004). A review of astronomy education research. *Astronomy Education Review, 2*(2), 20-45.

Bransford, J. D. (2000). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Brewin, C. R., Andrews, B., & Gotlib, I. H. (1993). Psychopathology and early experience: A reappraisal of retrospective reports. *Psychological Bulletin, 113*(1), 82-98.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24,* 385-396.

Comins, N. F. (2001). *Heavenly errors: Misconceptions about the real nature of the universe.* New York, NY: Columbia University Press.

Comins, N. F. (2013). *Discovering the essential universe (5th ed.)*. New York, NY: W. H. Freeman & Company.

Comins, N. F., & Kaufmann III, W. J. (2012). *Discovering the universe (9th ed.)*. New York, NY: W. H. Freeman.

Elder Jr., G. H., Pavalko, E. K., & Clipp, E. C. (1993). *Working with archival data: Studying lives*. Newbury Park, CA: Sage Publication.

Favia, A., Comins, N. F., Thorpe, G. L., & Batuski, D. J. (2014). A direct examination of college student misconceptions in astronomy: A new instrument. *Journal and Review of Astronomy Education and Outreach, 1*, A21-A39.

Henry, B., Moffitt, T. E., Caspi, A., Langley, J., & Silva, P. A. (1994). On the 'remembrance of things past': A longitudinal evaluation of the retrospective method. *Psychological Assessment, 6*, 92-101.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks, CA: Sage Publications.

Kuhlmann, S., Piel, M., & Wolf, O. T. (2005). Impaired memory retrieval after psychosocial stress in healthy young men. *Journal of Neuroscience, 25*(11), 2977-2982.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lombardi, D., Sinatrab, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27,* 50-62.

Olson, J. M., & Cal, A. V. (1984). Source credibility, attitudes, and the recall of past behaviors. *European Journal of Social Psychology, 14*, 203-210.

Onwuegbuzie, A. J., Daniel, L., & Leech, N. L. (2007). Pearson product-moment correlation coefficient. In N. J. Salkind & K. Rasmussen (Eds*.), Encyclopedia of measurement and statistics* (pp. 751-756). Thousand Oaks, CA: Sage Publications.

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,* 122-136.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265-296.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Slater, T. F., & Adams, J. P. (2002). *Learner-centered astronomy teaching: Strategies for astro 101*. Boston, MA:Addison-Wesley.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. Duschl & R. Hamilton (Eds.), *Philosophy of science, cognitive science, and educational theory and practice* (pp. 147-176). Albany, NY: State University of New York Press.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*, 45-69.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage Publications.

APPENDIX A

PRELIM QUESTIONS AND ASSOCIATED STATEMENTS FROM THE ASTRONOMY BELIEFS INVENTORY

sA50n: few objects orbit the Sun in circular orbits
Prelim 1 Question 14: Which of the following most accurately describes the shape of Neptune's orbit around the Sun?

sA80: craters are volcanic in origin
Prelim 1 Question 38: Which statement about the surface of the Moon is correct?

sA164n: Mars currently lacks running water on its surface
Prelim 2 Question 2: Which statement about water on Mars is correct?

sA50n: few objects orbit the Sun in circular orbits
Prelim 2 Question 25: What shape is the orbit of Halley's comet?

sA172: Saturn's rings are solid
Prelim 2 Question 30: Saturn's rings are best described by which of the following?

sA76: Jupiter's Great Red Spot is a volcano erupting
Prelim 2 Question 39: Jupiter's Great Red Spot is most accurately called a(n)?

sA46n: Jovian planets (Jupiter, Saturn, Uranus, Neptune) have gaseous or icy surfaces
Prelim 2 Question 41: Which statement most accurately describes descending into Jupiter?

sA150: the Earth is in the middle of the Milky Way galaxy
sA222n: the Sun is far away from the center of the Milky Way galaxy
Prelim 3 Question 7: What is located at the center of the Milky Way galaxy?

APPENDIX B

FINAL EXAM QUESTIONS AND ASSOCIATED INVENTORY STATEMENTS

sA111n: Earth's axis is tilted compared to the ecliptic
Final Exam Question 3: Where is the ecliptic as seen from the Earth?

sA2: there are 12 zodiac constellations
Final Exam Question 9: How many *zodiac* constellations are there?

sA185: the Sunspot cycle is 11 years long
Final Exam Question 10: What is the length of the Sun's entire magnetic cycle?

sA46n: Jovian planets (Jupiter, Saturn, Uranus, Neptune) have gaseous or icy surfaces
Final Exam Question 21: Which of the following best describes the surface of Jupiter?

sA172: Saturn's rings are solid
Final Exam Question 22: Which statement about Saturn's rings is most accurate?

sA171: Saturn is the only planet with rings
Final Exam Question 23: Which of the following planets does not have rings?

sA8: the north star is the brightest star in the sky
Final Exam Question 24: What is the brightest star in the night sky?

sA248: cosmic rays are light rays
Final Exam Question 40: Cosmic rays are best described by which of the following?

sA108n: Venus is similar to earth in size
sA170: Mars is the sister planet to earth in physical properties and dimensions
Final Exam Question 44: Which planet is most similar in physical dimensions and composition to the Earth?

sA208: the Sun will explode as a nova
Final Exam Question 51: The Sun will end its "life" as a(n):

sA50n: few objects orbit the Sun in circular orbits
Final Exam Question 67: Which of the following most accurately describes the shape of Venus's orbit around the Sun?

sA262n: the universe as a whole is changing
Final Exam Question 98: Which of the following best describes the overall motion of the universe?

APPENDIX C

BLACK HOLES AND GALAXIES PRETEST

Question 1: How are black holes created today?

(A) They form spontaneously (that is, without the need for any external matter or energy) in empty space
(B) While black holes once formed, they do not form today
(C) They form deep inside some old stars
(D) They form as the result of collisions between high-energy particles (atoms or molecules) in space
(E) There is no evidence that black holes exist, hence I don't believe they form today.

Question 2: What is the fate of black holes?

(A) They last forever, each with a fixed mass
(B) They slowly lose mass, that is, they evaporate
(C) They continue to gain mass indefinitely
(D) They continue to grow, but one is growing faster than the others. It will eventually swallow the universe
(E) Black holes don't exist, hence the question is irrelevant

Question 3: Black holes consist of:

(A) nothing - they are just holes in space
(B) a uniform "sea of energy"
(C) ultra-condensed concentrations of matter and energy
(D) ultra-condensed concentrations of matter, only
(E) misleading question since black holes don't really exist

Question 4: As detectable from our universe, what shape does a black hole have?

(A) a pinhole in space
(B) a sphere
(C) a wormhole
(D) many different possible shapes, including spherical, disc, and donut shapes
(E) misleading question since black holes don't really exist

Question 5: A spacecraft is put into circular orbit around a distant black hole. Which one of the following outcomes will happen to the spacecraft?

(A) It will be sucked straight into the black hole, like dust into a vacuum cleaner
(B) It will spiral inward, eventually falling into the black hole
(C) It will stop and hover above the black hole
(D) It will orbit around the black hole forever in the same orbit
(E) The question assumes black holes exist, which I believe they don't.

Question 6: What would happen to an asteroid that passed into a black hole?

(A) It would be crushed and would remain in the black hole
(B) It would enter another dimension
(C) It would tunnel through a wormhole back into our universe
(D) It would remain intact in the black hole
(E) The question assumes black holes exist, which I believe they don't.

Question 7: How do astronomers detect black holes?

(A) Astronomers detect black holes by observing their glow
(B) Astronomers detect black holes by seeing matter being ejected from them
(C) Astronomers detect black holes by putting spacecraft in orbit around them
(D) Astronomers detect black holes by their effects on nearby gas and stars
(E) No one has actually found a black hole

Question 8: If we boarded a spaceship to journey into a black hole, what would happen to us?

(A) We would begin to spin faster and faster
(B) We would get stuck inside the black hole
(C) We would have the ability to travel freely in time
(D) We would teleport to another dimension
(E) We would never find a black hole since they don't exist

Question 9: How many galaxies exist in the universe today?

(A) There are no galaxies
(B) There is just our galaxy
(C) There are a few galaxies
(D) There are several thousand
(E) There are billions of galaxies

Question 10: Relative to the Milky Way galaxy, the solar system is located:

(A) above the Milky Way
(B) between two spiral arms
(C) in a spiral arm
(D) in or near its center
(E) far away from Milky Way

Question 11: How many distinct shapes do galaxies have?

(A) All galaxies are shaped like spirals
(B) Galaxies can be either spiral or elliptical
(C) Galaxies can have any shape other than spiral or elliptical
(D) Galaxies can be spiral, elliptical, or irregularly shaped
(E) Galaxies can have any shape

Question 12: Where in the universe is the Milky Way located today?

(A) The Milky Way has no special location in the universe
(B) The Milky Way is located at or near the center of the universe
(C) The Milky Way is the universe
(D) The Milky Way is located outside the universe
(E) The Milky Way is located inside the solar system

Question 13: Where is the Sun located relative to the Milky Way?

(A) The Sun is located above the plane of the Milky Way
(B) The Sun is located between two spiral arms of the Milky Way
(C) The Sun is located at the Milky Way's center
(D) The Sun is located in a spiral arm of the Milky Way
(E) The Sun is located far outside of the Milky Way

Question 14: How are galaxies distributed throughout the universe?

(A) The galaxies are distributed randomly
(B) The galaxies are grouped together by size
(C) The galaxies are in large groups separated by voids
(D) The galaxies are grouped together by type or design
(E) There is only one galaxy in the universe

Question 15: Which statement about observing stars in the Milky Way is most accurate?

(A) Astronomers can observe up to a few hundred stars in the Milky Way
(B) Astronomers can observe those stars not obscured by galactic dust
(C) Astronomers can observe hundreds of thousands of stars in the Milky Way
(D) Astronomers can observe all of the stars in the Milky Way
(E) Astronomers can observe only the brightest stars (O, B, A)

Question 16: Which one statement about galaxy properties is correct?

(A) All galaxies are held together by gravity
(B) All galaxies have the same shape
(C) All galaxies have the same size
(D) All galaxies have the same color
(E) All galaxies rotate at the same speed

Question 17: Which statement most accurately describes the contents of the Milky Way?

(A) The Milky Way lacks gas, but contains dust and stars
(B) The Milky Way lacks dust, but contains gas and stars
(C) The Milky Way lacks gas and dust, but contains stars
(D) The Milky Way contains gas, dust, and stars
(E) The Milky Way lacks stars, gas, and dust altogether

Question 18: Which one of the following is true about the Milky Way?

(A) The Milky Way is still producing both stars and planets
(B) The Milky Way is still producing stars, but not planets
(C) The Milky Way is still producing planets, but not stars
(D) The Milky Way is no longer producing stars or planets
(E) The Milky Way never created its own stars or planets

APPENDIX D

ATTENDANCE QUESTIONS

Attendance Question 1

n = 171, no inter-rater analysis conducted

sA2. there are 12 zodiac constellations

September 5, 2013 (immediately at the start of class, prior to lecture). How many zodiac constellations are there?

1. (36%) 13
2. (34%) 12
3. (35%) Other, vague, unsure

Attendance Question 2

n = 162, $\kappa = 0.74$

sA112. summer is warmer because we are closer to the sun during the summertime

sA148. seasons are caused by speeding up and slowing down of Earth's rotation

September 5, 2013 (at the end of class, as normal). What causes the seasons?

1. (38%) Tilt of the Earth
2. (17%) Changing distance between the Earth and the Sun
3. (46%) All other (less logical) (e.g. "the Earth's rotation")

Attendance Question 3

n = 166, no inter-rater analysis conducted

sA50n. few objects orbit the Sun in circular orbits

September 10, 2013. What is the shape of the Earth's orbit around the sun?

1. (86%) elliptical or oval
2. (13%) circular or sphere
3. (2%) other

Attendance Question 4

n = 154, $\kappa = 0.62$

sA161. Mars is green (from plant life)

September 30, 2013. Have astronomers found life on Mars?

1. (38%) No, (5%) Only as fossils (traces)
2. (18%) Other (e.g., found water/ice, didn't answer question)
3. (40%) Yes

Attendance Question 5

n = 150, no inter-rater analysis conducted

sA46n. Jovian planets (Jupiter, Saturn, Uranus, Neptune) have gaseous or icy surfaces

October 3, 2013. Describe the surface of Jupiter

1. (37%) Gas, liquid (but not water), or "no surface" - but not solid
2. (35%) Solid - not gas (e.g., "cratered")
3. (29%) All other, including mixed surface types (e.g., gas and rocky)

Attendance Question 6

n = 138, κ = 0.71

sA172. Saturn's rings are solid

October 10, 2013. Briefly describe the rings of Saturn

1. (78%) Particles, ice, debris, rocks, asteroids, metals, "like Jupiter's"
2. (14%) Solid continuous matter, inc. gravity-held "material" or "rocky elements"
3. (8%) Gas only, clouds, or liquid

Attendance Question 7

n = 120, no inter-rater analysis conducted

sA189. the Sun is the brightest star in universe

sA190n. some objects in the universe are brighter than the Sun

November 7, 2013. How bright is the sun compared to other stars?

1. (40%) average or about average
2. (37%) dimmer, or (22%) brighter
3. (2%) the Sun is the brightest star

Attendance Question 8

n = 128, κ = 1.00

sA248. cosmic rays are light rays

November 12, 2013. What are cosmic rays?

1. (1%) High-energy particles from space, (2%) High-energy particles (not from space or unspecified)
2. (21%) Light or light rays
3. (77%) All other (inc. vague, unspecified, multiple responses, e.g., "rays coming from stars," "radiation," "shock waves from supernovas")

Attendance Question 9

n = 125, no inter-rater analysis conducted

sA150. the Earth is in the middle of the Milky Way galaxy

sA222n. the Sun is far away from the center of the Milky Way galaxy

November 21, 2013. Where in the Milky Way are we?

1. (2%) Between two spiral arms, (30%) On or in a spiral arm
2. (30%) In or near the center
3. (37%) All other or too vague to tell (e.g., "upper left," "off to the side," "on the edge")

APPENDIX E

THE CHILDHOOD INTEREST IN ASTRONOMY SURVEY

Possible response options for each question: “never,” “occasionally,” “very often”

1. Prior to college, how often did you read astronomy books on your own?
2. Prior to college, how often did you watch astronomy programs on your own?
3. Prior to college, how often did you use binoculars or a telescope to view the night sky?
4. Prior to college, how often did you participate in an astronomy club out of pure interest?
5. Prior to college, how often did you go to planetarium shows out of pure interest?
6. Prior to college, how often did you go to an observatory out of pure interest?
7. Prior to college, how often did you keep up with news stories related to astronomical events?
8. Prior to college, how often did you choose to talk to others about astronomy, out of pure interest?

APPENDIX F

STRESS SURVEY

[As discussed in the Section on Stress, these questions were adopted from Cohen et al. (1983).]

The questions in this scale ask you about your feelings and thoughts during the last month. In each case, you will be asked to indicate how often you felt or thought a certain way. Although some of the questions are similar, there are differences between them, so you should treat each question separately. The best approach is to answer each question fairly quickly. That is, don't try to count up the number of times you felt a particular way; instead, indicate the alternative that seems like a reasonable estimate.

For each question, choose from the following alternatives:

0	1	2	3	4
Never	Almost Never	Sometimes	Fairly Often	Very Often

1. In the last month, how often have you felt that you were unable to control the important things in your life?
2. In the last month, how often have you felt confident about your ability to handle your personal problems?
3. In the last month, how often have you felt that things were going your way?
4. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?