A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

2011, AER, 10, 010107-1, 10.3847/AER2011030

# A Study of General Education Astronomy Students' Understandings of Cosmology. Part II. Evaluating Four Conceptual Cosmology Surveys: A Classical Test Theory Approach

**Colin S. Wallace**
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
**Edward E. Prather**
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
**Douglas K. Duncan**
Department of Astrophysical and Planetary Sciences, University of Colorado at Boulder, Boulder, Colorado 80309
Received: 09/21/11, Accepted: 11/1/11, Published: 12/15/11

## Abstract

This is the second of five papers detailing our national study of general education astronomy students' conceptual and reasoning difficulties with cosmology. This article begins our quantitative investigation of the data. We describe how we scored students' responses to four conceptual cosmology surveys, and we present evidence for the inter-rater reliability of those scores. We devote the bulk of this article to a classical test theory analysis of the data. We calculate difficulties and discriminations for each item, and we compute Cronbach's $\alpha$ as a measure of the reliability of the surveys. We also discuss the implications this analysis has for the validity of the surveys.

## 1. INTRODUCTION

This is the second paper in a five-paper series describing one of the first large-scale, systematic studies of general education introductory astronomy (hereafter Astro 101) students' conceptual and reasoning difficulties with cosmology. In the first article in this series (Wallace, Prather, and Duncan 2011a; hereafter "Paper 1"), we described how we designed four surveys that can be used to investigate Astro 101 students' conceptual understanding of the expansion and evolution of the universe, the Big Bang, and the evidence for dark matter in spiral galaxies. We followed Wilson's (2005) recommendation for the survey design and interpretation process:

(1) Define the constructs to be measured and create construct maps for each construct;
(2) Design survey items;
(3) Categorize and score the full range of responses; and
(4) Apply psychometric models to the data.

Paper 1 focused on the first two of these steps. Each of the four surveys (called Forms A–D) focuses on a different construct: Form A examines students' abilities to interpret Hubble plots, Form B examines students' models of the expansion of the universe and the Big Bang, Form C examines whether or not students understand how the properties of the universe have changed over time, and Form D examines whether students can reconstruct the chain of reasoning linking the flat rotation curves of spiral galaxies to the existence of dark matter. Paper 1 also presents qualitative evidence for the validity of these surveys.

In this article, we address the third and fourth of Wilson's steps. Specifically, we use classical test theory (CTT) to quantitatively analyze students' responses to the surveys. (See Lord and Novick 1968 for a comprehensive treatment of CTT or Ding and Beichner 2009, Hambleton and Jones 1993, or Wallace and Bailey 2010 for briefer pedagogical overviews.) We have two reasons for conducting and presenting this analysis. First, it helps us to understand the strengths and weaknesses of the surveys, including the reliability of students' scores. Second, it allows us to demonstrate how we refined the surveys over time. We will not use this article to describe common student difficulties, nor will we report any pre-instruction-to-post-instruction gains. These are the subjects of Paper 4 (Wallace, Prather, and Duncan 2011c) and Paper 5 (Wallace, Prather, and Duncan 2011d), respectively, of this series.

Note that students' responses to the surveys are almost entirely free-response. This means that the first step in our analysis was scoring students' written responses. Section 2 describes the scoring rubrics we used to assign numerical values to these responses. Because the scoring process required the judgment of those doing the scoring, we also assessed the inter-rater reliability of the scores. Section 3 contains our inter-rater reliability analysis. Section 4 presents the CTT analysis of our data from the fall 2009, spring 2010, and fall 2010 semesters, respectively. Note that because CTT statistics are highly sample-dependent, one cannot compute them for one population of students and assume that they apply to other populations. CTT statistics must be recalculated each time a survey is administered to a new group of students, or each time the survey undergoes a substantial change. This is why we do not calculate a single set of CTT statistics based on our data for all three semesters. Section 5 summarizes this article and explains how the data presented in this article adds quantitative data to the validity argument begun in Paper 1.

## 2. SCORING RUBRICS

We constructed detailed scoring rubrics each semester for each item on each survey. These rubrics allowed us to score and categorize the full range of students' responses—which is essentially Wilson's (2005) third step in survey design and interpretation. We constructed our rubrics only when we had all the pre-instruction and post-instruction responses for a given semester in hand. We also revised our rubrics after each semester in tandem with our revisions to the surveys. Our rubrics are therefore based on detailed, iterative, and qualitative analyses of actual student responses. For most items, the rubric has two components: an overall score, which is based on whether or not the student gave a correct answer and a complete and correct explanation, and codes for the most common reasoning elements used by students. These rubrics allowed us to perform quantitative analyses on the survey responses as well as examine the data for patterns in the reasoning elements used by students pre-instruction and post-instruction (see Paper 4). To give readers a flavor for the amount of detail we can extract from students' responses, Table 1 shows the scoring rubric for a single item on a single survey (Item 1 on Form B, which asks the following: "Explain, in as much detail as possible, what astronomers mean when they say 'the universe is expanding.' Provide a drawing if possible to help illustrate your thinking."). Table 2 shows a sample of student responses along with their assigned overall scores and reasoning element codes.

## 3. INTER-RATER RELIABILITY

We conducted our inter-rater reliability analysis to assess whether or not multiple science education researchers would be able to use these rubrics to arrive at consistent and reliable overall scores and reasoning elements to a subset of student responses (Otero and Harlow 2009). While one of us (Wallace) scored every student response to every survey for every semester (a total of 2318 surveys pre-instruction and 2041 post-instruction), two other science education researchers collaborated to score a representative sample of 65 responses. These responses were taken from 21 students and 9 items. We compared the scores assigned by Wallace with those of two science education researchers. We agreed on the overall score for 83% of the items. The difference between the overall scores was never greater than one integer. Furthermore, 74% of the reasoning elements coded by Wallace for student responses agreed with the reasoning elements assigned by the other researchers. 76% of the reasoning elements assigned by the other researchers agreed with the reasoning elements assigned by Wallace. We find these high percentages of overlap on the assigned reasoning elements to be particularly impressive considering how detailed and lengthy the list of possible reasoning elements is within the rubrics, and how diverse the responses are in the student data set. These results support the inter-rater reliability of our ability to apply these rubrics to students' responses on all four forms.

The statistic Cohen's $\kappa$ (Cohen 1960) provides another way to ascertain inter-rater reliability. It corrects the overall proportion of scores on which two raters agree by the proportion of scores on which they are expected to agree

**Table 1. The scoring rubric for Item 1 on Form B (fall 2010 version)**

| Score | Response characteristics | Reasoning elements |
|---|---|---|
| 4 | Students in Category 4 meet the same criteria as students in Category 3 except they only discuss increasing distances and/or redshifts in the context of galaxies (and not planets, stars, or other objects smaller than galaxies). Students may be placed in Category 4 without explicitly saying that<br>• the universe has no center<br>• the universe has no edge<br>• the Big Bang refers to the evolution of the universe from a hot, dense state.<br>If the student does discuss one of these elements and makes an incorrect statement, then s/he cannot be placed in category 4. | I—Student says that the size of the universe increases over time.<br>II—Student says that the universe has a center.<br>III—Student says the universe has no center.<br>IV—Student says that the universe has an edge.<br>V—Student says that the universe has no edge.<br>VI—Student talks about redshifts/Doppler shifts.<br>VII—Student says space (time) is growing/stretching.<br>VIII— Student talks about the movement of galaxies and/or their increasing distances. |
| 3 | Students in this category say that<br>• the universe increases in size and/or<br>• the distances between all objects increase.<br>Additionally, they must make at least one of the following claims and not contradict the others:<br>• the universe has no center.<br>• the universe has no edge.<br>• the Big Bang refers to the evolution of the universe from a hot, dense state. | IX—Student talks about the movement of stars and/or their increasing distances.<br>X—Student talks about the movement of planets and/or their increasing distances.<br>XI—Student talks about the movement of objects (something unspecified or not about a star, planet, or galaxy) and/or their increasing distances.<br>XII—Student says that the distances between everything increases.<br>XIII—Student says that farther objects move away faster. |
| 2 | Students in this category say that the universe increases in size and either provide no other information or claim one or more of the following:<br>• the universe has a center.<br>• the universe has an edge.<br>• the Big Bang was an explosion.<br>The distances between all objects or objects smaller than galaxies increases. | XIV—Student talks about the Big Bang.<br>XV—Student talks about an explosion.<br>XVI—Student says that the early universe was once hot, small, and/or dense.<br>XVII—Student says that we learn more about the universe over time.<br>XVIII—Student talks about how we are looking further back in time as we look farther into space. |
| 1 | Students in this category describe the expansion of the universe as something other than the universe getting larger over time. The two most popular answers that fit in this category are:<br>• Expansion refers to learning more about the universe over time.<br>• Expansion refers to the creation of new objects over time. | XIX—Student says new things are created in the universe over time.<br>XX—Student gives irrelevant information.<br>XXI—Student gives some other reason not specified above.<br>XXII—Student has no idea.<br>XXIII—Answer field is blank or the student provided no reason or explanation. |
| 0 | Students in this category write nothing (the answer field is blank), or they provide information that does not answer the question, or they say that they have no idea and provide no further information. | |

**Table 2. A sample of student responses to Item 1 on Form B and their overall scores and reasoning element codes. The student responses contain their original spellings, grammar, and punctuation**

| Student response | Overall score and reasoning element codes |
|---|---|
| "There was a man named 'hubble' who had a telescope and observed galaxies 'red shifting' or moving away. Now when he observed this he had the best telescope and plotted the rate of speed of expansion. He noticed that although the closer galaxies are red shifting at a higher rate, the most far away galaxies were moving away (Red shifting) at a much higher speed. From this observation astronomers deducted the universe is expanding from all far away galaxies are red shifting At a higher speed" | **Score and Codes:** 4_VI_VIII_XIII<br>**Explanation:** This student talked about redshifts (VI), the movement of galaxies (VIII), and how the farther away galaxies are, the faster they move away from us (XIII). Since the student talked about distances to galaxies increasing and did not make any incorrect statements about expansion, he receives an overall score of 4. |
| "'the universe is expanding' because as time goes on, matter is moving further and further away from other matter. Temperature and density have decreased over time as the universe is expanding and matter has moved away" | **Score and Codes:** 3_XI_XVI<br>**Explanation:** This student talks about otherwise unspecified pieces of matter moving away from one another (XI) and talks about how the universe used to be hotter and denser (XVI). While none of these pieces of information are necessarily wrong, she only gets an overall score of 3 because she does not specify whether expansion only affects the distances between galaxies or whether the distances between smaller objects are also affected. |
| "There was a big bang which exploded out everything in the universe. The leading edge of this bang has been expanding ever since." | **Score and Codes:** 2_XIV_XV_XI_IV<br>**Explanation:** This student talks about the Big Bang (XIV) as an explosion (XV). He mentions the movement of objects (XI) and an edge (IV). These reasoning elements give him an overall score of 2. |
| "I don't think that it is acculuy expanding in a physical sense, but instead or knowledge of the univers and the areas that we have discovered is expanding with an increase in technology and invesments in sciens" | **Score and Codes:** 1_XVII<br>**Explanation:** This student denies that the universe is physically growing. Instead, he says that expansion refers to our increase in knowledge over time (XVII), placing him in Category 1. |

by chance. For our data, Cohen's $\kappa = 0.755$. Fleiss, Levin, and Paik (2003) say such a value for $\kappa$ corresponds to an "excellent" agreement, while Landis and Koch (1977) call it a "substantial" agreement.

One weakness of Cohen's $\kappa$ is that it equally weights all deviations between raters' scores (Cohen 1968; Fleiss, Levin, and Paik 2003). For example, an item for which one rater assigned a student a score of 3 and another assigned a score of 1 receives the same weight in the calculation of Cohen's $\kappa$ as an item for which the first rater assigned a student a score of 3 and the other assigned a score of 2. Cohen (1968) introduced a modification of his eponymous statistic (called Cohen's weighted $\kappa$ or $\kappa_w$) in order to account for this weakness. For our data, $\kappa_w = 0.823$. This again corresponds to what Fleiss, Levin, and Paik (2003) call an "excellent" agreement. Landis and Koch (1977) consider this to be an "almost perfect" agreement. Regardless of whether one adopts the characterizations of Fleiss, Levin, and Paik (2003) or Landis and Koch (1977), and regardless of whether we look at $\kappa$ or $\kappa_w$, our inter-rater reliability appears to be quite strong.

## 4. CTT ANALYSIS

In the fall 2009, we collected 501 pre-instruction and 406 post-instruction survey responses from students enrolled in three Astro 101 courses taught at two separate institutions. In the spring of 2010, we surveyed four Astro 101 courses from three institutions for a total of 1215 Astro 101 students pre-instruction and 1081 students post-instruction. In the fall 2010, the final semester in which we collected data, we surveyed a total of 602 students pre-instruction and 554 students post-instruction; these students were drawn from a total of 14 classes representing 12 institutions. See Paper 1 for more demographic details.

We analyzed and scored these responses using the rubrics described in Section 2. We calculated each item's difficulty (i.e., $P$-value, or the average score on that item divided by its total number of possible points) and discrimination (the linear correlation between the item's scores and its survey form's total scores) for every item on every form for each semester, pre-instruction and post-instruction (Tables 3 and 4, respectively). We also calculated Cronbach's $\alpha$ for each survey form each semester, pre-instruction and post-instruction (Table 5). Cronbach's $\alpha$ is given by the formula

$$\alpha = \frac{N}{N-1}\frac{\sigma_x^2 - \sum_{i=1}^{N}\sigma_{y_i}^2}{\sigma_x^2}, \tag{1}$$

where $N$ is the total number of items on a test, $\sigma_x^2$ is the variance in the total scores of the test-taking population, and $\sum_{i=1}^{N}\sigma_{y_i}^2$ is the sum of the variances of the test-taking population's scores on individual items $y_i$ (Lord and Novick 1968). Cronbach's $\alpha$ provides an estimate of a test-score reliability, where reliability is conceptualized as the internal consistency of observed responses. Cronbach's $\alpha$ is close to one when the covariances among items are high, and it is close to zero when the covariances among items are low.

Before providing a description of the item difficulties and discriminations, a few clarifications need to be made. Note that Item 2 from Form C for fall 2009 and Item 3 for Form D for fall 2009 and spring 2010 were removed prior to our analysis because these items were deemed conceptually problematic from an astrophysical standpoint. We also modified the survey forms between semesters as a result of our quantitative analysis of the data and in response to issues raised during think-aloud interviews of Astro 101 students (see Paper 1). Our major changes were the following:

- In the fall 2009 version of Item 5 on Form A, students had to choose a graph. In the spring 2010 and fall 2010 versions, they had to draw their own graph.
- After the fall 2009, we added an additional item (Item 6) to Form A to further probe students' abilities to reason about the age of the universe using Hubble plots.
- After the fall 2009, a new item was inserted in Form B at the location of Item 3 that examined students'

**Table 3. The pre-instruction difficulties (*P*-values) and discriminations of the items on Forms A–D**

| | | Form A | | Form B | | Form C | | Form D | |
|---|---|---|---|---|---|---|---|---|---|
| | Item | *P*-val. | Discrm. | *P*-val. | Discrm. | *P*-val. | Discrm. | *P*-val. | Discrm. |
| Fall 2009 | Item 1 | 0.45 | 0.66 | 0.39 | 0.61 | 0.55 | 0.61 | 0.56 | 0.15 |
| | Item 2 | 0.39 | 0.63 | 0.55 | 0.58 | | | 0.87 | 0.60 |
| | Item 3 | 0.34 | 0.58 | 0.64 | 0.64 | 0.42 | 0.77 | | |
| | Item 4 | 0.34 | 0.45 | 0.44 | 0.54 | 0.41 | 0.22 | 0.55 | 0.83 |
| | Item 5 | 0.53 | 0.56 | 0.69 | 0.39 | 0.46 | 0.70 | | |
| | Item 6 | | | 0.62 | 0.32 | | | | |
| Spring 2010 | Item 1 | 0.48 | 0.70 | 0.38 | 0.49 | 0.45 | 0.63 | 0.53 | 0.22 |
| | Item 2 | 0.39 | 0.63 | 0.54 | 0.47 | 0.43 | 0.12 | 0.86 | 0.71 |
| | Item 3 | 0.34 | 0.42 | 0.80 | 0.53 | 0.48 | 0.61 | | |
| | Item 4 | 0.34 | 0.39 | 0.37 | 0.32 | 0.47 | 0.58 | 0.53 | 0.78 |
| | Item 5 | 0.77 | 0.63 | 0.44 | 0.51 | 0.49 | 0.53 | | |
| | Item 6 | 0.33 | 0.61 | 0.58 | 0.56 | 0.60 | 0.50 | | |
| | Item 7 | | | 0.64 | 0.35 | | | | |
| Fall 2010 | Item 1 | 0.48 | 0.59 | 0.48 | 0.57 | 0.48 | 0.47 | 0.57 | 0.64 |
| | Item 2 | 0.38 | 0.48 | 0.58 | 0.46 | 0.42 | 0.23 | 0.54 | 0.44 |
| | Item 3 | 0.34 | 0.24 | 0.73 | 0.50 | 0.45 | 0.71 | 0.76 | 0.73 |
| | Item 4 | 0.34 | 0.28 | 0.37 | 0.37 | 0.46 | 0.45 | 0.68 | 0.72 |
| | Item 5 | 0.80 | 0.75 | 0.40 | 0.35 | 0.49 | 0.56 | 0.57 | 0.65 |
| | Item 6 | 0.62 | 0.75 | 0.43 | 0.58 | 0.61 | 0.56 | 0.34 | 0.51 |
| | Item 7 | | | 0.71 | 0.44 | | | 0.41 | 0.61 |

**Table 4. The post-instruction difficulties (*P*-values) and discriminations of the items on Forms A-D**

| | | Form A | | Form B | | Form C | | Form D | |
|---|---|---|---|---|---|---|---|---|---|
| | **Item** | ***P*-val.** | **Discrm.** | ***P*-val.** | **Discrm.** | ***P*-val.** | **Discrm.** | ***P*-val.** | **Discrm.** |
| Fall 2009 | Item 1 | 0.51 | 0.59 | 0.64 | 0.61 | 0.81 | 0.63 | 0.84 | 0.66 |
| | Item 2 | 0.40 | 0.63 | 0.79 | 0.46 | | | 0.90 | 0.69 |
| | Item 3 | 0.39 | 0.66 | 0.78 | 0.39 | 0.80 | 0.69 | | |
| | Item 4 | 0.37 | 0.67 | 0.70 | 0.45 | 0.62 | 0.69 | 0.73 | 0.80 |
| | Item 5 | 0.47 | 0.61 | 0.84 | 0.60 | 0.69 | 0.75 | | |
| | Item 6 | | | 0.71 | 0.42 | | | | |
| Spring 2010 | Item 1 | 0.53 | 0.56 | 0.64 | 0.38 | 0.86 | 0.55 | 0.75 | 0.59 |
| | Item 2 | 0.44 | 0.64 | 0.79 | 0.38 | 0.67 | 0.51 | 0.88 | 0.54 |
| | Item 3 | 0.37 | 0.61 | 0.82 | 0.50 | 0.78 | 0.63 | | |
| | Item 4 | 0.37 | 0.61 | 0.49 | 0.53 | 0.63 | 0.54 | 0.65 | 0.85 |
| | Item 5 | 0.88 | 0.52 | 0.57 | 0.66 | 0.56 | 0.49 | | |
| | Item 6 | 0.47 | 0.56 | 0.65 | 0.61 | 0.57 | 0.34 | | |
| | Item 7 | | | 0.76 | 0.50 | | | | |
| Fall 2010 | Item 1 | 0.46 | 0.46 | 0.69 | 0.39 | 0.76 | 0.61 | 0.66 | 0.43 |
| | Item 2 | 0.41 | 0.41 | 0.72 | 0.48 | 0.51 | 0.29 | 0.76 | 0.59 |
| | Item 3 | 0.39 | 0.56 | 0.91 | 0.33 | 0.66 | 0.64 | 0.87 | 0.65 |
| | Item 4 | 0.40 | 0.53 | 0.51 | 0.40 | 0.59 | 0.52 | 0.79 | 0.67 |
| | Item 5 | 0.83 | 0.59 | 0.60 | 0.68 | 0.60 | 0.55 | 0.72 | 0.79 |
| | Item 6 | 0.66 | 0.70 | 0.62 | 0.72 | 0.63 | 0.42 | 0.56 | 0.61 |
| | Item 7 | | | 0.83 | 0.50 | | | 0.71 | 0.72 |

understanding of the edge of our observable universe and what exists beyond it. The fall 2009 Item 3 was removed because students could give a “correct” response without providing any scientifically correct reasoning. A revised version of the old Item 3 from the fall 2009 was written and became Item 4 on Form B for spring 2010 and fall 2010. We revised the item to become a multiple-choice question whose choices best reflect the dominant reasoning provided by students in their fall 2009 responses regarding the Big Bang.

- As the result of the changes discussed in the previous bullet, Items 4, 5, and 6 of Form B from the fall 2009 became Items 5, 6 and 7 for Form B for the spring 2010 and fall 2010.
- Items 3, 4, and 5 on the fall 2009 version of Form C became Items 1, 2, and 3 on the spring 2010 and fall 2010 versions.
- Item 1 from fall 2009 of Form C grew to become Items 4, 5, and 6 in the spring 2010. The new items focused on observing light from a specific event (the explosion of a supernova) at a great distance in an expanding

**Table 5. Cronbach’s $\alpha$ (pre-instruction and post-instruction) for Forms A–D**

| | Form | Pre-instruction | Post-instruction |
|---|---|---|---|
| Fall 2009 | Form A | 0.44 | 0.60 |
| | Form B | 0.41 | 0.37 |
| | Form C | 0.40 | 0.62 |
| | Form D | −0.04 | 0.53 |
| Spring 2010 | Form A | 0.59 | 0.59 |
| | Form B | 0.38 | 0.50 |
| | Form C | 0.42 | 0.41 |
| | Form D | 0.42 | 0.16 |
| Fall 2010 | Form A | 0.52 | 0.52 |
| | Form B | 0.41 | 0.51 |
| | Form C | 0.44 | 0.42 |
| | Form D | 0.72 | 0.75 |

universe. These new items allowed us to probe subtly different aspects of student reasoning, which were conceptually lumped together in the fall 2009 version of Item 1.

- For the spring 2010, Item 4 on Form D was converted to a multiple-choice question to best reflect the dominant reasoning provided by students in the fall 2009 responses.
- Between the spring 2010 and fall 2010 semesters, we expanded Form D from three items to seven. The new items ask students to think about the rotation curve for a solar system, the distribution of matter in a solar system, and how these are different for spiral galaxies. Items 1, 2, and 4 from the spring 2010 version became Items 2, 4, and 6, respectively, in the fall 2010 version.

While we made other slight improvements to the wordings of items between semesters, the above list constitutes the major changes we believed would improve our ability to assess students' understanding of cosmology.

With these modifications in mind, there are a few points worth noting about the *P*-values and discriminations shown in Tables 3 and 4. Typically, item *P*-values should be as close to 0.50 as possible in order to maximize test reliability (Ding and Beichner 2009), although many studies accept *P*-values as low as 0.10 or 0.20 and as high as 0.80 or 0.90 (Bardar *et al*. 2007; Ding *et al*. 2006; Maloney *et al*. 2001). Likewise, convention suggests that item discriminations should be greater than or equal to 0.20 (Ding and Beichner 2009). As Tables 3 and 4 show, nearly all the *P*-values and discriminations fall within these conventionally accepted limits. Overall, this analysis indicates that both pre-instruction and post-instruction the survey items were quite successful at challenging students' understandings and discriminating the responses of students.

Note that the items on Form A present the greatest overall challenge to students, both pre-instruction and post-instruction. These items ask students to reason about the expansion rate and age of the universe using Hubble plots. There are many factors that make it difficult for Astro 101 students to achieve the highest scores in our rubric. In order for students' responses to be given the highest scores, the response had to identify the correct graph and include a very sophisticated explanation using conceptually complex chains of reasoning. Additionally, these items present a task that elicits well-documented student difficulties relating to graph interpretation (McDermott, Rosenquist, and van Zee 1987). Finally, the data shown combines all students from all types of courses. Only some of these courses were taught using research-validated curricula explicitly designed to address learning difficulties related to Hubble plots.

In comparing the pre-instruction and post-instruction values for items on Forms B, C, and D, one sees that the difficulties for several items change dramatically. For example, the *P*-value of Item 1 on the spring 2010 version of Form C changes by 0.41 pre-instruction to post-instruction. Many items on Forms B, C, and D have post-instruction *P*-values in the 0.80 or greater range. This indicates that many students improved their understandings of the concepts probed by these items as a result of their instruction on these cosmology topics.

Table 5 shows Cronbach's $\alpha$ pre-instruction and post-instruction for Forms A–D for the fall 2009, spring 2010, and fall 2010 semesters. Almost all these values are lower than the conventionally accepted minimum value of $\alpha = 0.70$ (George and Mallory 2009). Why is Cronbach's $\alpha$ low for each survey form, pre-instruction and post-instruction?

Like many CTT statistics, Cronbach's $\alpha$ depends on the test-taking population and the items to which they respond. For example, homogeneous sets of responses yield lower values of Cronbach's $\alpha$, since Cronbach's $\alpha$ increases with increased values for total score variance of the analyzed data (Thompson 2003). Cronbach's $\alpha$ is also sensitive to the brevity of a test (Schmitt 1996). Tests with fewer items will typically yield smaller values of Cronbach's $\alpha$. One can estimate how the value of $\alpha$ will change for an equivalent test of arbitrary length using the Spearman–Brown prophecy formula (Lord and Novick 1968). The effects of test brevity are quite evident with the values of Cronbach's $\alpha$ for the fall 2009 and spring 2010 versions of Form D. Because the responses of students were often homogeneous and because there were so few items on these survey forms, small changes in students' responses pre-instruction to post-instruction can lead to large changes in the values of Cronbach's $\alpha$.

Because Cronbach's $\alpha$ is sensitive to the homogeneity of the test-taking population and the length of the test, we cannot automatically conclude that our surveys are unreliable when they have values of Cronbach's $\alpha < 0.70$. Cronbach's $\alpha$ is only a lower bound on test score reliability (Lord and Novick 1968). Indeed, the other statistics presented in this article, plus our item response theory (IRT) analysis (Wallace, Prather, and Duncan 2011b, aka "Paper 3") provide evidence for the reliability of Forms A-D.

## 5. SUMMARY

In this article, we used CTT to analyze a total of 4359 student responses to the four conceptual cosmology surveys forms. We collected these responses over a period of three consecutive semesters (fall 2009-fall 2010) from a nationwide sample of Astro 101 students. This analysis is a quantitative counterpart to the qualitative data we described in Paper 1. We looked at how students' responses to the four survey forms were analyzed and coded and how various measures of inter-rater reliability were established. Further, we provided the results of our calculations of the difficulties and discriminations of all the survey items, and Cronbach's $\alpha$ for each semester's version of each survey form.

The analysis presented here illustrates how CTT can be used to revise and refine the items used to probe students' understandings of key topics in cosmology. Our analysis shows that the items represent a wide range of difficulties and discriminations and challenge students' underlying ideas and reasoning difficulties related to understanding cosmology. Because survey design is often an iterative procedure, we want to make our process as transparent as possible so others can judge the validity of the results we report in all the papers in this series.

We can also use this data to begin addressing one of the unaddressed components of our validity argument, which we began in Paper 1: Can we reliably transform students' responses into numerical scores? Our inter-rater reliability analysis in Section 3 suggests that the answer to this question is "Yes." Multiple people can use our scoring rubrics for our survey items to arrive at the same scores.

Our CTT analysis has helped us to better understand which cosmology topics addressed by our research students struggle to understand. Our research goes beyond CTT to include an analysis of this data set using IRT. IRT offers the potential for estimating sample-independent item and student parameters. It also provides alternative ways for evaluating the reliability of the surveys. Our IRT analysis is the subject of Paper 3 of this series.

## Acknowledgments

We are indebted to Sébastien Cormier for his help in scoring students' responses. Gina Brissenden provided valuable feedback on earlier versions of this manuscript. We also thank the anonymous reviewer who suggested several edits that improved this article. This material is based in part upon work supported by the National Science Foundation under Grant Nos. 0833364 and 0715517, a CCLI Phase III Grant for the Collaboration of Astronomy Teaching Scholars (CATS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

Bardar, E. M., Prather, E. E, Brecher, K., and Slater, T. F. 2007, "Development and Validation of the Light and Spectroscopy Concept Inventory," *Astronomy Education Review*, 5, 103.

Cohen, J. 1960, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37.

Cohen, J. 1968, "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin*, 70, 213.

Ding, L., and Beichner, R. 2009, "Approaches to Data Analysis of Multiple-Choice Questions," *Physical Review Special Topics–Physics Education Research*, 5, 020103.

Ding, L., Chabay, R., Sherwood, B., and Beichner, R. 2006, "Evaluating an Electricity and Magnetism Assessment Tool: Brief Electricity and Magnetism Assessment," *Physical Review Special Topics– Physics Education Research*, 2, 010105.

Fleiss, J. L., Levin, B., and Paik, M. C. 2003, *Statistical Methods for Rates and Proportions*, 3rd ed., Hoboken, NJ: John Wiley and Sons, Inc.

George, D. and Mallery, P. 2009, *SPSS for Windows Step by Step: A Simple Guide and Reference*, Boston, MA: Pearson Education, Inc.

Hambleton, R. K., and Jones, R. J. 1993, “Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development,” *Educational Measurement: Issues and Practice*, 12, 253.

Landis, J. R., and Koch, G. G. 1977, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 33, 159.

Lord, F. M., and Novick, M. R. 1968, *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Maloney, D. P., O’Kuma, T. L., Hieggelke, C. J., and van Heuvelen, A. 2001, “Surveying Students’ Conceptual Knowledge of Electricity and Magnetism,” *American Journal of Physics*, 69, S12.

McDermott, L. C., Rosenquist, M. L., and van Zee, E. H., 1987, “Student Difficulties in Connecting Graphs and Physics: Examples From Kinematics,” *American Journal of Physics*, 55, 503.

Otero, V. K., and Harlow, D. B. 2009, “Getting Started in Qualitative Physics Education Research,” in *Reviews in PER Vol. 2: Getting Started in PER*, ed. C. Henderson and K. A. Harper, College Park, MA: American Association of Physics Teachers, 1.

Schmitt, N. 1996, “Uses and Abuses of Coefficient Alpha,” *Psychological Assessment*, 8, 350.

Thompson, B. 2003, “Understanding Reliability and Coefficient alpha, Really,” in *Score Reliability*, ed. B. Thompson, Thousand Oaks, CA: SAGE Publications, 3.

Wallace, C. S., and Bailey, J. M. 2010, “Do Concept Inventories Actually Measure Anything?,” *Astronomy Education Review*, 9, 010116.

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011a, “A Study of General Education Astronomy Students’ Understandings of Cosmology. Part I. Development and Validation of Four Conceptual Cosmology Surveys,” *Astronomy Education Review*, 10, 010106.

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011b, “A Study of General Education Astronomy Students’ Understandings of Cosmology. Part III. Evaluating Four Conceptual Cosmology Surveys: An Item Response Theory Approach,” *Astronomy Education Review* (submitted).

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011c, “A Study of General Education Astronomy Students’ Understandings of Cosmology. Part IV. Common Difficulties Students Experience with Cosmology,” *Astronomy Education Review* (submitted).

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011d, “A Study of General Education Astronomy Students’ Understandings of Cosmology. Part V. The Effects of a New Suite of Cosmology *Lecture-Tutorials* on Students’ Conceptual Knowledge,” (in preparation).

Wilson, M. 2005, *Constructing Measures: An Item Response Modeling Approach*, Mahwah, NJ: Lawrence Erlbaum Associates.

ÆR
010107-1–010107-9