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

Astronomy Education Review

2012, AER, 11, 010103-1, 10.3847/AER2011031

A Study of General Education Astronomy Students' Understandings of Cosmology. Part III. Evaluating Four Conceptual Cosmology Surveys: An Item Response 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: 10/24/11, Accepted: 12/23/11, Published: 02/2/12

Abstract

This is the third of five papers detailing our national study of general education astronomy students' conceptual and reasoning difficulties with cosmology. In this paper, we use item response theory to analyze students' responses to three out of the four conceptual cosmology surveys we developed. The specific item response theory model we use is known as the partial credit model. Since readers may be unfamiliar with the partial credit model, we provide a pedagogical introduction to this model. We use the partial credit model to assess the reliabilities of the four survey forms and to determine the probabilities of students achieving different scores on survey items.

1. INTRODUCTION

This is the third 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 Paper 1 (Wallace, Prather, and Duncan 2011a), we described how we designed four survey forms (denoted A–D) to measure students' conceptual cosmology knowledge. Each survey 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 contains qualitative evidence for the validity of these surveys. Paper 2 (Wallace, Prather, and Duncan 2011b) discusses how we scored students' survey responses and it contains our inter-rater reliability and classical test theory (CTT) analyses of those scores.

This CTT analysis was important for our research because CTT provides statistics that are easy to compute, analyze, and report, and which yield important insights into the functioning of the surveys. But while our CTT analysis informed our interpretations of and revisions to the survey forms, there are still many questions we have that require an analysis beyond CTT to answer. For example, we cannot establish (using CTT alone) whether the low values we obtained for Cronbach's α indicate that our surveys are unreliable, or whether our calculated values of Cronbach's α are dominated by other effects, such as test length (Schmitt 1996) and/or population homogeneity (Thompson 2003). CTT also provides us with no information about how the probability that a student will earn a particular score on a particular item is related to student knowledge/ability.

To help address these issues, we analyzed our data using item response theory (IRT). IRT uses an ensemble of data (students' scores) to simultaneously model the ability of each student and the difficulty of each item. In this paper, we use an IRT model known as the partial credit model (Masters 1982). This paper thus complements the results of the CTT analysis we described in Paper 2. Note that the data to which we applied IRT and CTT analysis methods were created by rigorously analyzing student written responses to the items of the four survey forms and coding these responses with very detailed rubrics (for more, see Papers 1 and 2). Readers who want a pedagogical introduction to IRT should consult Hambleton and Jones (1993), Harris (1989), or Wallace and Bailey (2010). This paper provides a detailed discussion of the results of our IRT analysis.

The purpose of this analysis is to understand the characteristics of the survey forms' items and the reliability of our surveys from an IRT point of view. We will neither elucidate common student difficulties (the subject of Wallace, Prather, and Duncan 2012a, aka "Paper 4") nor will we perform any analysis related to increases in student achievement (the subject of Wallace, Prather, and Duncan 2012b, aka "Paper 5"). Our IRT analysis is thus complimentary to our CTT analysis in Paper 2.

Since members of the astronomy education research community may be unfamiliar with the partial credit model, we give a pedagogical overview. Section 2 describes the partial credit model in detail. Section 3 contains our partial credit model analysis of students' scores. We summarize the results of our partial credit model analysis in Section 4.

2. THE PARTIAL CREDIT MODEL

The partial credit model is a member of the Rasch family of IRT models. The Rasch model (Rasch 1980/1960) is the simplest member of this family. It may be written as

$$\ln\left[\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right] = \theta_p - b_i, \tag{1}$$

where $P(X_{pi}=1)$ represents the probability that a student p correctly answers an item i and $P(X_{pi}=0)$ represents the probability that a student p incorrectly answers an item i. The natural logarithm of the odds (i.e., the ratio between $P(X_{pi}=1)$ and $P(X_{pi}=0)$) equals the difference between student p's ability "θ_p" and item i's difficulty "b_i." This relationship ensures abilities and difficulties are both measured in log odds units (logits).

The model described by Eq. (1) only applies to dichotomously scored items (i.e., items that are scored either 1 or 0 depending on whether a student gave a correct or incorrect response, respectively). Since items on the four conceptual cosmology survey forms were scored polytomously (see Paper 2), we used the partial credit model. Masters (1982) formulated the partial credit model in order to analyze items for which there are ordered levels of performance and for which partial credit is assigned to a respondent's answer.

The key difference between the partial credit model and the Rasch model is that the partial credit model does not assign a single number to an item to represent its difficulty. Instead, it characterizes an item by multiple *step difficulties*. Each step difficulty b_{ij} determines when a student of ability θ_p is just as probable to have a score j as she is to have the next highest score $j+1$ on item i. Equation (1) is thus altered such that it applies to each set of adjacent scores. For example, imagine that an item has four possible scores: 0, 1, 2, and 3. The partial credit model assumes that Eq. (1) holds between scores 0 and 1, 1 and 2, and 2 and 3

$$\ln\left[\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right] = \theta_p - b_{i0}, \tag{2}$$

$$\ln\left[\frac{P(X_{pi}=2)}{P(X_{pi}=1)}\right] = \theta_p - b_{i1}, \tag{3}$$

and

$$\ln\left[\frac{P(X_{pi}=3)}{P(X_{pi}=2)}\right] = \theta_p - b_{i2}. \tag{4}$$

Thus, when $\theta_p = b_{i0}$, student p is just as likely to earn a score of 0 on item i as she is to earn a score of 1. These equations can be combined and generalized for any item i scored $x = 0,\ldots,m_i$. For the category $x = j$, the partial credit model is (Embretson and Reise 2000; Masters 1982)

$$P_{ix}(\theta_p) = \frac{\exp\left[\sum_{j=0}^{x}(\theta_p - b_{ij})\right]}{\sum_{r=0}^{m_i}\left[\exp\left[\sum_{j=0}^{r}(\theta_p - b_{ij})\right]\right]}, \tag{5}$$

as long as

$$\sum_{j=0}^{0}(\theta_p - b_{ij}) \equiv 0. \tag{6}$$

Sometimes we find the step difficulties b_{ij} are not as useful or easy to talk about as an item i's *Thurstonian thresholds* β_{ij}. The Thurstonian threshold β_{ij} for category j is defined as the ability at which the probability of getting a score less than j equals the probability of getting a score of j or greater (Wilson 2005).

We used the CONSTRUCTMAP software (Kennedy *et al*. 2006) to estimate students' abilities θ_p, items' step difficulties b_{ij}, and items' Thurstonian thresholds β_{ij}. CONSTRUCTMAP anchors the logit scale by setting the mean item difficulty to 0 logits. We used the expected *a posteriori* procedure in our estimates of students' abilities, since it is both computational faster and more accurate than other estimation procedures (Embretson and Reise 2000). Readers who are interested in the details of IRT parameter estimation techniques should consult Baker and Kim (2004).

3. PARTIAL CREDIT MODEL ANALYSIS

As noted in Paper 1, we collected and scored a total of 4359 surveys over the course of this research project. We used the partial credit model to analyze each semester's data. We collected a total of 907 student responses in the fall 2009, 2296 in the spring 2010, and 1156 in the fall 2010. For a given semester, we estimated student and item parameters using both pre- and post-instruction responses. This is acceptable because IRT, unlike CTT, attempts to disentangle item difficulty parameters from students' abilities. We can, therefore, estimate difficulty parameters using students of low abilities, students of high abilities, or both. By using both pre- and post-instruction responses in our estimations, we are using the full range of student abilities present in our data.

3.1. Item Parameters

Tables 1–3 show the step difficulties and Thurstonian thresholds for each item on the fall 2009, spring 2010, and fall 2010 versions of Forms A–C (as discussed in Paper 2, we modified several items between semesters, and we removed some items prior to our analysis because they were deemed conceptually problematic from an astrophysical standpoint). We did not apply the partial credit model to any version of Form D since it exhibits what Yen (1993) calls "item chaining." Item chaining means that each item builds off the previous item such that knowing the answer to one item increases the probability that one correctly answers the next. Because Form D has item chaining, it violates the assumption of IRT that IRT model parameters should explain all correlations between students' responses (Embretson and Reise 2000; Yen 1993).

Our experience with IRT analyses suggests that step difficulty and Thurstonian threshold values < -3 logits are abnormally low and values > 3 logits are abnormally high. Most of the step difficulties and Thurstonian thresholds shown in Tables 1–3 fall within these limits, but there are exceptions. We are not concerned about values < -3 logits since our scoring rubrics' requirements for earning scores greater than 0 are easy to achieve: For most items, students simply had to give a response (even an incorrect one) to get a score of 1. On several items, such as Items 1–4 on Form A, students were simply required to choose a graph to earn a 1. Thus, the vast majority of students provided answers, which is why students of even the lowest abilities have non-zero

Table 1. The step difficulty b_{ij} and Thurstonian threshold β_j parameters for the items on Forms A–C for the fall 2009. All values are in logits

		Step parameters				Thurstonian thresholds			
	Item	b_{i0}	b_{i1}	b_{i2}	b_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i4}
	Item 1	−6.70	−0.40	2.72	…	−6.76	−0.48	2.71	…
	Item 2	−4.96	0.71	7.45	…	−5.01	0.66	8.14	…
	Item 3	−5.00	1.62	3.68	…	−5.05	1.46	3.75	…
	Item 4	−4.69	1.92	3.49	…	−4.74	1.71	3.60	…
Form A	Item 5	−3.16	−1.52	4.85	…	−3.37	−1.42	4.80	…
	Item 1	−0.68	−0.20	2.77	0.56	−1.05	0.08	1.59	1.88
	Item 2	−0.81	−0.94	2.03	…	−1.34	−0.48	2.08	…
	Item 3	−0.87	−1.35	1.78	…	−1.51	−0.77	1.82	…
	Item 4	−2.11	1.06	1.52	…	−2.16	0.74	1.89	…
	Item 5	−1.44	1.37	−0.97	…	−1.52	0.16	0.42	…
Form B	Item 6	−2.05	1.31	…	…	−2.09	1.34		…
	Item 1	−3.62	2.32	−0.92	…	−3.62	0.61	0.80	…
	Item 3	−1.67	1.44	0.50	…	−1.67	1.44	0.49	…
	Item 4	−1.97	1.61	1.56	…	−1.99	1.15	2.06	…
Form C	Item 5	−1.33	2.02	0.06	…	−1.38	0.91	1.25	…

probabilities of getting scores greater than 0. Scores higher than 1 required students to give correct answers and/or correct reasons for those answers. See Paper 2 for an example scoring rubric.

In contrast, the fact that some step difficulties and Thurstonian thresholds are > 3 logits is of considerable interest. When this occurs, it indicates an item that is very difficult for students of even the highest abilities.

Table 2. The step difficulty b_{ij} and Thurstonian threshold β_j parameters for the items on Forms A–C for the spring 2010. All values are in logits

		Step parameters				Thurstonian thresholds			
	Item	b_{i0}	b_{i1}	b_{i2}	b_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i4}
	Item 1	−5.36	0.17	3.84	…	−5.37	0.15	3.86	…
	Item 2	−4.75	1.36	5.37	…	−4.76	1.34	5.38	…
	Item 3	−5.21	3.42	3.11	…	−5.21	2.85	3.68	…
	Item 4	−4.95	3.40	3.46	…	−4.95	2.94	3.93	…
	Item 5	−3.62	−0.95	…	…	−3.68	−0.88	…	…
Form A	Item 6	−2.55	1.20	4.36	…	−2.58	1.18	4.40	…
	Item 1	−1.60	−0.27	3.38	−0.17	−1.80	−0.12	1.60	1.75
	Item 2	−1.24	−0.66	1.72	…	−1.58	−0.41	1.80	…
	Item 3	−1.60	−0.19	…	…	−1.78	0.00	…	…
	Item 4	−2.85	1.19	2.85	…	−2.86	1.66	3.12	…
	Item 5	−2.04	2.20	0.52	…	−2.05	1.17	1.59	…
	Item 6	−0.40	1.31	−0.25	…	−0.62	0.51	0.86	…
Form B	Item 7	−1.78	0.83	…	…	−1.84	0.90	…	…
	Item 1	−2.67	1.89	−0.11	…	−2.68	0.73	1.09	…
	Item 2	−2.06	1.32	1.45	…	−2.09	0.91	1.90	…
	Item 3	−1.58	2.07	−0.30	…	−1.61	0.79	1.07	…
	Item 4	−3.49	2.83	−0.17	…	−3.48	1.22	1.45	…
	Item 5	−2.77	2.31	0.56	…	−2.77	1.23	1.64	…
Form C	Item 6	−1.32	1.69	0.33	…	−1.38	0.83	1.29	…

Table 3. The step difficulty b_{ij} and Thurstonian threshold β_j parameters for the items on Forms A–C for the fall 2010. All values are in logits

		Step parameters				Thurstonian thresholds			
Item	b_{i0}	b_{i1}	b_{i2}	b_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i4}	
	Item 1	−5.00	0.62	3.09	…	−5.01	0.55	3.17	…
	Item 2	−4.09	1.64	8.41	…	−4.09	1.64	8.41	…
	Item 3	−4.48	3.01	2.81	…	−4.48	2.47	3.35	…
	Item 4	−4.30	2.86	2.69	…	−4.30	2.33	3.22	…
	Item 5	−2.55	−0.92	…	…	−2.70	−0.77	…	…
Form A	Item 6	−2.02	−0.19	0.15	…	−2.16	−0.45	0.56	…
	Item 1	−1.46	−0.80	3.99	−1.38	−1.78	−0.52	1.33	1.39
	Item 2	−0.18	−1.70	2.61	…	−1.18	−0.73	2.62	…
	Item 3	−1.17	−0.36	…	…	−1.46	−0.07	…	…
	Item 4	−3.28	1.93	2.78	…	−3.29	1.66	3.06	…
	Item 5	−1.65	2.18	0.59	…	−1.67	1.19	1.62	…
	Item 6	−0.65	1.67	0.39	…	−0.76	0.89	1.34	…
Form B	Item 7	−2.03	0.28	…	…	−2.12	0.37	…	…
	Item 1	−3.02	1.36	0.34	…	−3.03	0.57	1.15	…
	Item 2	−1.79	1.40	2.14	…	−1.83	1.14	2.45	…
	Item 3	−0.98	2.49	−0.73	…	−1.03	0.86	1.04	…
	Item 4	−3.13	3.33	−0.90	…	−3.13	1.16	1.28	…
	Item 5	−2.57	2.56	−0.31	…	−2.58	1.02	1.26	…
Form C	Item 6	−1.71	1.22	0.31	…	−1.77	0.52	1.09	…

Stated another way, values > 3 logits show us where most students have a very small probability of achieving a certain score. To take one example, the spring 2010 version of Item 4 on Form A (which asks students to select, for a universe with a decreasing expansion rate, the corresponding Hubble plot and to defend their selection) has $b_{i2} = 3.46$ and $\beta_{i3} = 3.93$. This means most students are extremely unlikely to be able to earn a score of 3 on this item. Tables 1–3 show several other examples of items for which b_{ij} and/or β_{ij} are > 3 logits.

To reiterate, what do we learn when b_{ij} and/or $\beta_{ij} > 3$ logits? We learn that such items are so difficult that most students will not give complete and correct answers. This, in turn, provides us with information on which cosmological concepts are most difficult for students. See Paper 4 for a description of the most common cosmological concepts for students, as revealed by the surveys we developed.

We now examine these results using category response curves (CRCs). CRCs give us a graphical way of displaying the same type of information encoded in Tables 1–3. Figure 1 shows the CRCs for a single item (the fall 2009 version of Form B's Item 1). This item asked students 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." Each curve in Figure 1 shows how the probability of attaining a certain score (0, 1, 2, 3, or 4) changes as a function of student ability θ_p. The higher the point on each curve, the more likely a student with the corresponding ability will provide a written response that would have been coded with the corresponding score (0, 1, 2, 3, or 4). The CRC for score 0 has its greatest probability at the lowest student abilities. By the time we are sampling students of ability $\theta_p = -0.4$, the CRC for score 1 has the greatest probability. The probability of giving a response coded with a score 2 dominates for students with ability $\theta_p = 0.4$ or greater. The last two CRCs illustrate just how difficult it was for students to give responses we would have coded with the highest two scores (3 and 4) for this item.

How does this compare with the information one gets from CTT? In CTT one only gets a single number (the P-value) to describe an item's difficulty. For example, the pre-instruction P-value for the fall 2009 version of Item 1 on Form B is 0.39. This means that students in this surveyed population averaged a 39% on this item (but since this item was scored polytomously, $P = 0.39$ does not mean that 39% of students got this item correct and received full credit). While this is a useful statistic to compute, it does not provide any information about how well one might expect students of different abilities to perform on this item.

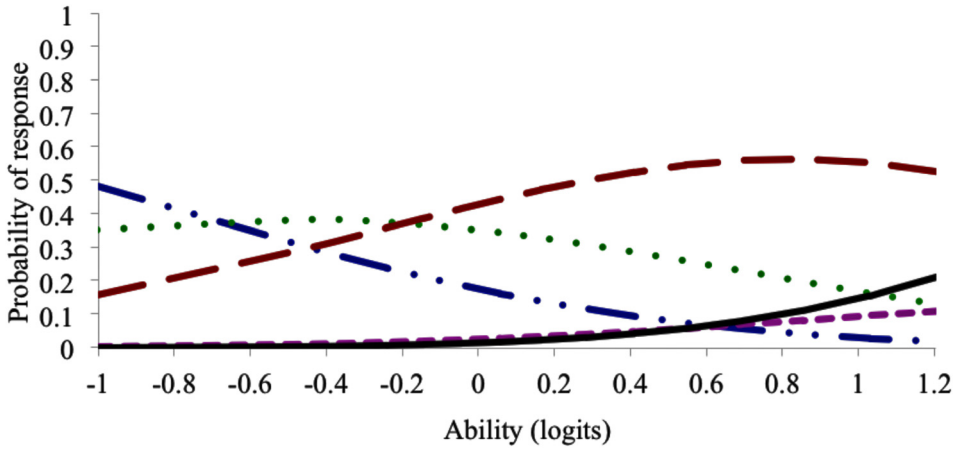

Figure 1. The CRCs for Form B, Item 1 from the fall 2009. Each curve shows the probability of a particular score as a function of ability. The blue dashed and dotted line correspond to a score of 0, the green dotted line to a score of 1, the red dashed line to a score of 2, the purple short dashed line to a score of 3, and the solid black line to a score of 4. The *x*-axis shows the range of estimated student abilities as measured by the fall 2009 version of Form B.

In contrast, IRT (in general) and the partial credit model (in particular) provide a much more expansive view of student performance on a given item. The CRCs allow one to predict the probability of receiving a particular response from a student of any given ability. For example, Figure 1 shows that students with an ability of $\theta_p = -1$ have virtually no chance of earning scores of 3 or 4 on Form B's Item 1. Those students have almost a 50% chance of getting a 0 on this item, although they also have an approximately 35% probability of earning a 1 and an approximately 15% probability of earning a 2.

To take another example, look at Figure 2, which shows the CRCs for Item 3 on the fall 2009 version of Form C. This item asked students the following: "Has the temperature of the universe changed over time, or has it always been about the same? Explain your reasoning and provide a drawing if possible to help illustrate your thinking." For this item, there were only four possible scores a student could have earned (0, 1, 2, or 3). This item is difficult for students with a wide range of abilities, as evidenced by the high probability of earning a score of 1, which indicates an incorrect response. However, students of abilities $\theta_p > 1$ were able to not only give a correct answer but also support that answer with reasoning that was scientifically correct. This result was not seen in the analysis of the next item.

At the more difficult end of the spectrum is Item 3 from the fall 2009 version of Form A. This item asks students "Which graph or graphs [from a bank of graphs we provided] show a universe that is expanding at a faster and faster rate over time? Explain your reasons for making your selection(s)." The CRCs in Figure 3 show that a score of 1 was always the most probable score for students over the entire observed range of abilities. While a

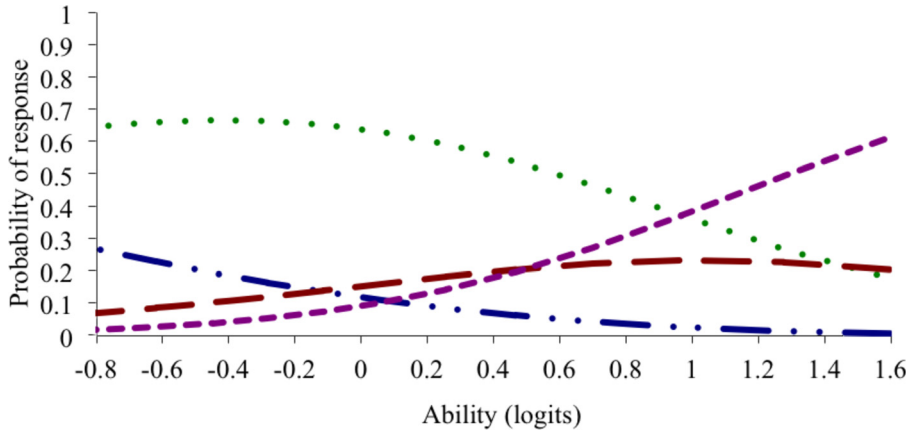

Figure 2. The CRCs for Form C, Item 3 from the fall 2009. Each curve shows the probability of a particular score as a function of ability. The blue dashed and dotted line correspond to a score of 0, the green dotted line to a score of 1, the red dashed line to a score of 2, and the purple short dashed line to a score of 3. The *x*-axis shows the range of estimated student abilities as measured by the fall 2009 version of Form C.

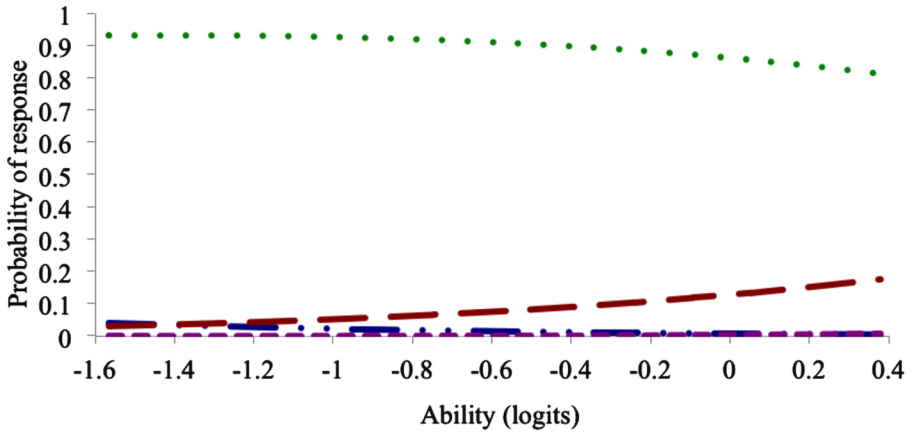

Figure 3. The CRCs for Form A, Item 3 from the fall 2009. Each curve shows the probability of a particular score as a function of ability. The blue dashed and dotted line corresponds to a score of 0, the green dotted line to a score of 1, the red dashed line to a score of 2, and the purple short dashed line to a score of 3. The x-axis shows the range of estimated student abilities as measured by the fall 2009 version of Form A.

small number of students of higher abilities had a non-negligible probability of earning a 2, the CRC shows that students of any ability were extremely unlikely to earn a score of 3. This result stems from the fact that students had to both pick the correct graph and defend their choice with a very sophisticated explanation using conceptually complex chains of reasoning. Overall, the CRCs shown in Figures 1–3 provide graphical illustrations of the kind of information contained in Tables 1–3, which show that we have created items with a wide range of difficulties in order to measure a wide range of student abilities.

3.2. Wright Maps

A Wright map provides a convenient way to summarize all information about students' abilities and items' difficulties for all items on a single survey form. Figures 4–12 display the Wright maps for each semester's version of Forms A–C. Each Wright map has two components. The left part is a histogram of students' abilities (or "proficiencies" in the nomenclature of CONSTRUCTMAP, which generated these graphs). The right part shows multiple dots for each item. Each blue square represents the ability above which a student is more likely to have a score of 1 or greater on that item. Each green circle represents the ability above which a student is more likely

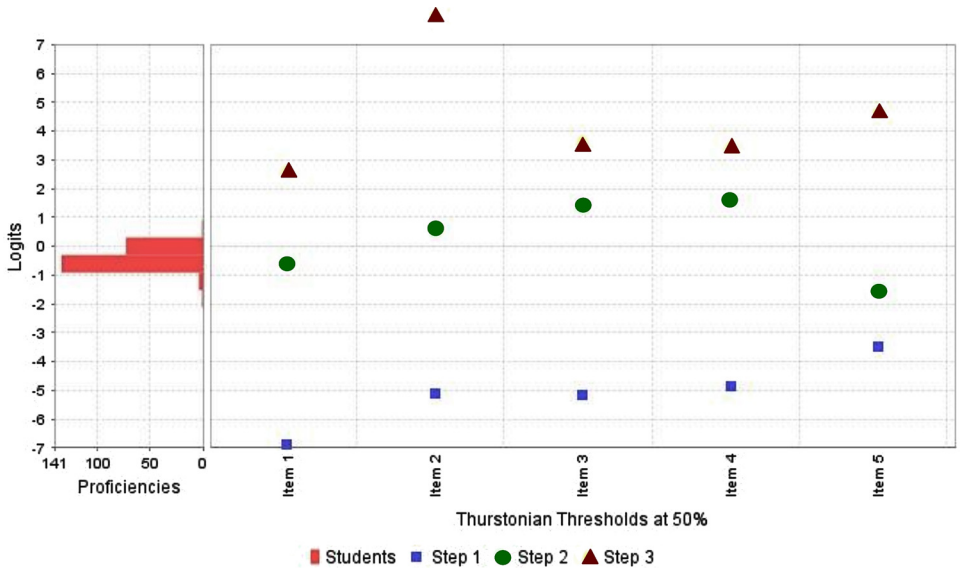

Figure 4. Form A's Wright map for the fall 2009. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

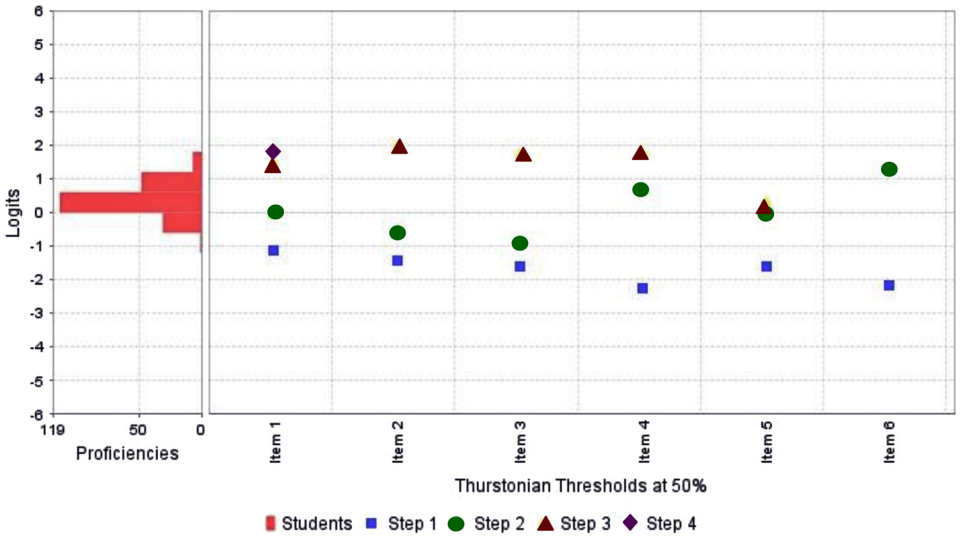

Figure 5. Form B's Wright map for the fall 2009. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

to have a score of 2 or greater on that item. Each red triangle represents the ability above which a student is more likely to have a score of 3 or greater on that item. Each purple diamond represents the ability above which a student is more likely to have a score of 4 on that item. In other words, the dots represent the logit locations of the Thurstonian thresholds (see Tables 1–3) for each item.

There are two pieces of information one can derive by looking at a Wright map. First, notice that for all versions of all forms, the Thurstonian thresholds generally "span the space" of students' abilities, by which we mean the logit range of the items' Thurstonian thresholds is greater than and overlaps with the range of observed student abilities. If the logit values of the items' Thurstonian thresholds were not greater than or did not overlap with the range of observed student abilities, then that would call into question the reliabilities of the surveys. The fact that there is not a large offset between the logit values covered by the items' Thurstonian thresholds and the locations of students' abilities is evidence that the survey forms are reliable (e.g., they can provide accurate measures of students' abilities since the vast majority of students have abilities that fall between the lowest and highest

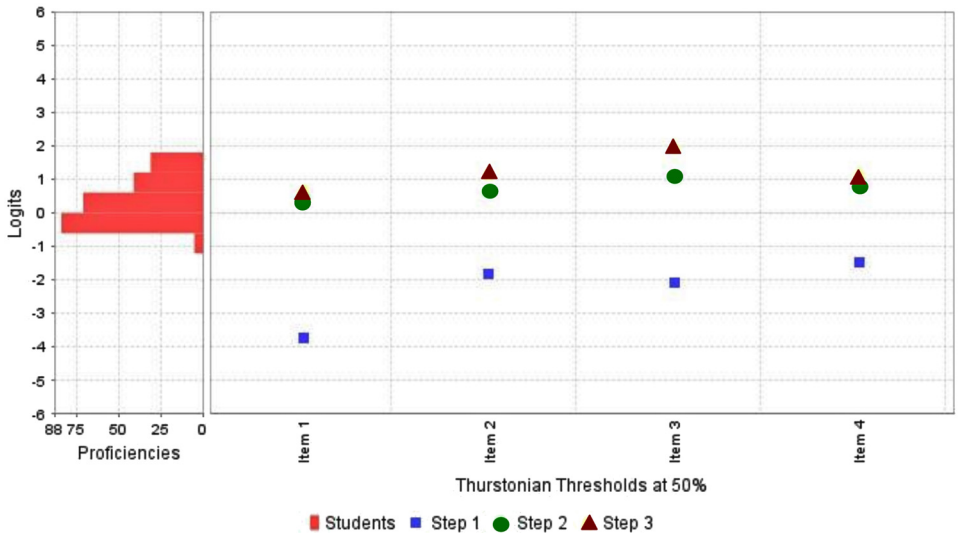

Figure 6. Form C's Wright map for the fall 2009. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

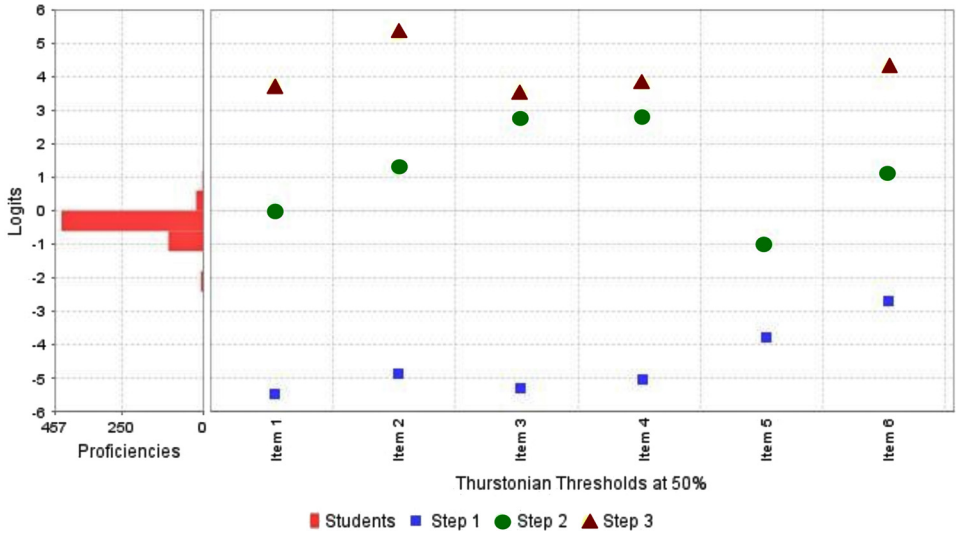

Figure 7. Form A's Wright map for the spring 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

Thurstonian thresholds). This is an important piece of information to extract from this analysis, especially because our CTT estimate of reliability, Cronbach's α, was affected by the brevity of each survey form and by the homogeneity of student responses (Schmitt 1996; Thompson 2003).

Second, one can tell from these Wright maps which scores most students were likely to receive on the survey forms' items. For example, consider Figure 5. The blue squares lie below the majority of student abilities, indicating that few students were likely to earn scores of 0 on any of Form B's items in the fall 2009. The green circles are also below many student abilities, although not as many as the blue squares. This means that many students had reasonable probabilities of scoring at least a 1 on each item on Form B. The red triangles and purple diamonds lie near the top of the histogram of abilities, indicating that only the highest ability students were likely to earn scores of 3 or 4 on Form B's items. For a contrasting case, look at Figure 4. In Figure 4, both the blue squares and red triangles for all items lie well outside of the range of student abilities. This means that all students who took Form A were unlikely to get scores of 0 or 3 on this item. Much more likely were scores of 1 or 2, whose transition point is shown by the green dots. Overall, the Wright Maps in Figures 4–12 reemphasize

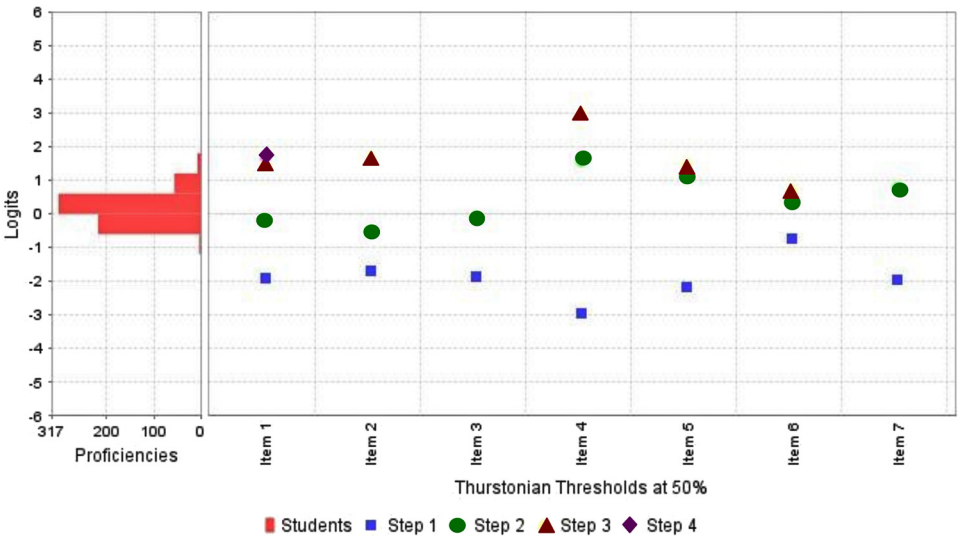

Figure 8. Form B's Wright map for the spring 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

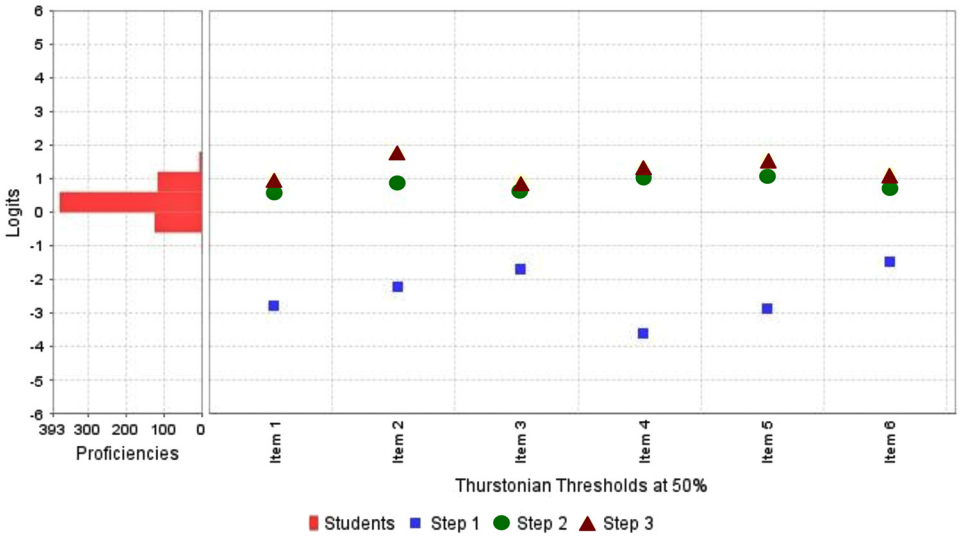

Figure 9. Form C's Wright map for the spring 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

that, for any semester, the constructs covered by Forms B and C (the expansion of the universe and the Big Bang, and the evolution of the universe, respectively) were easier for students to master than Form A's construct (interpreting Hubble plots). Note that only IRT provides this level of information—we could not make similar claims if we only analyzed our data with CTT, as we did in Paper 2.

4. SUMMARY

This paper presented our IRT analysis of students' responses to Forms A–C of the conceptual cosmology surveys. This analysis yielded valuable insights into the surveys, and the conceptual knowledge and reasoning abilities of our students. The analysis of step difficulties and Thurstonian thresholds for each item revealed which

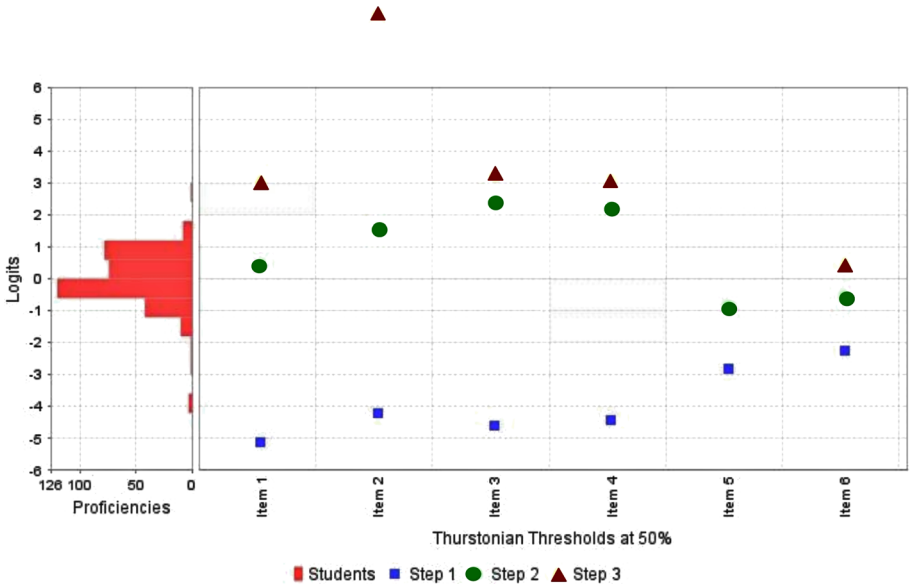

Figure 10. Form A's Wright map for the fall 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

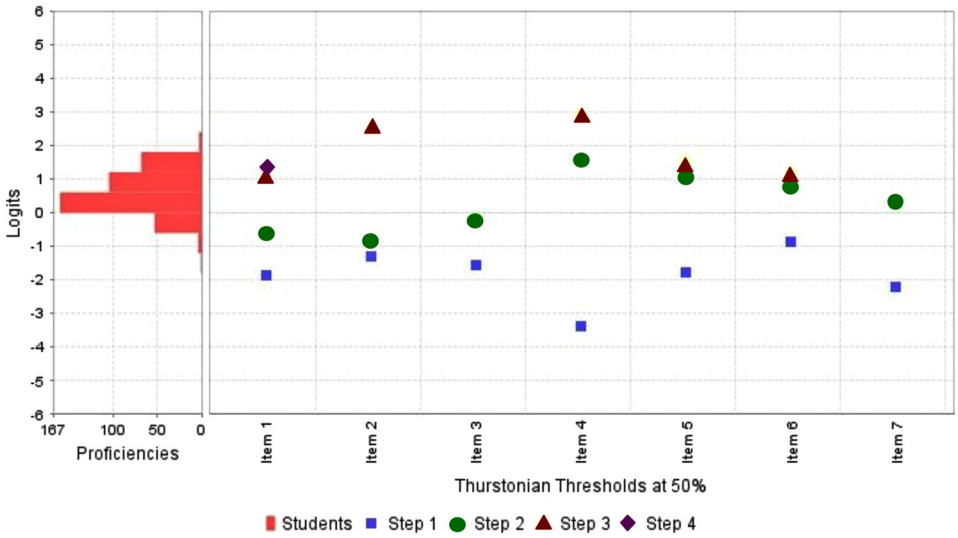

Figure 11. Form B's Wright map for the fall 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

levels of understanding were attainable or well beyond the abilities of our students. The data presented in this paper tell us that interpreting Hubble plots (the subject of Form A) is much more difficult for students than understanding the Big Bang and the expansion and evolution of the universe (the subjects of Forms B and C). We also obtained evidence for the reliabilities of Forms A–C, since the Wright maps presented in this paper show our surveys' items adequately span the space of students' abilities. This adds an important component to the validity evidence we began enumerating in Papers 1 and 2.

The information in Paper 1, Paper 2, and this paper illustrates the level of rigor our analysis has undergone to accurately probe Astro 101 students' conceptual and reasoning difficulties with cosmology. These three papers combined establish the foundation for our research methodology and the reliability and validity of the surveys we used to assess students' understandings of cosmology. The next two papers in this series present findings from this study that answer the fundamental research questions we posed in Paper 1: What are Astro 101 students' common conceptual and reasoning difficulties with cosmology? Do the new cosmology *Lecture-*

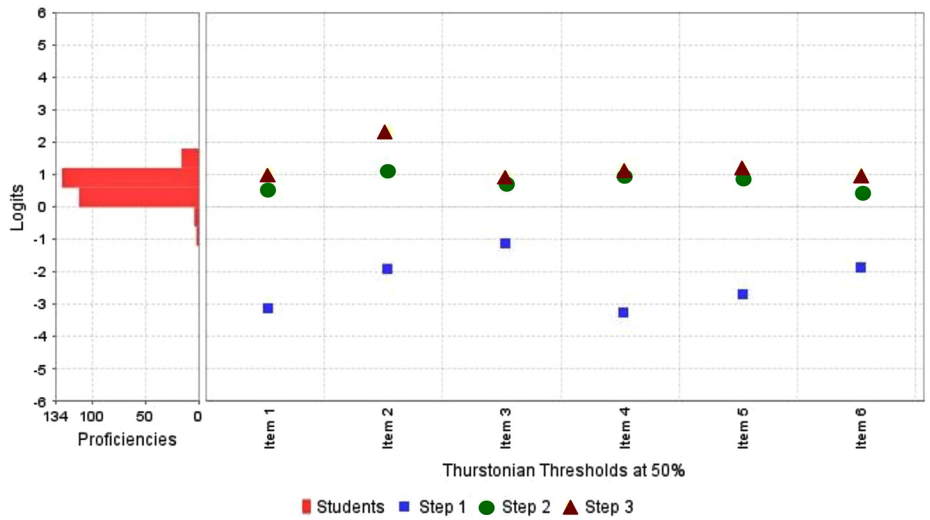

Figure 12. Form C's Wright map for the fall 2010. A histogram of students' abilities (proficiencies) is shown on the left. On the right are the Thurstonian thresholds for each item. Blue corresponds to β_{i0} (i.e., the ability at which one has equal probability of earning a score < 1 and a score ≥ 1), green to β_{i1} (i.e., the ability at which one has equal probability of earning a score < 2 and a score ≥ 2), etc.

Tutorials help students overcome these difficulties and achieve more expert-like understandings of cosmology? Answering the former question is the subject of Paper 4, while Paper 5 addresses the latter.

Acknowledgments

Gina Brissenden provided valuable feedback on earlier versions of this manuscript. This paper benefited from the anonymous reviewer who suggested edits that improved the final version of this paper. 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

Baker, F. B., and Kim, S. 2004, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., New York, NY: Marcel Dekker, Inc.

Embretson, S. E., and Reise, S. P. 2000, *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Harris, D. 1989, "Comparison of 1-, 2-, and 3-Parameter IRT Models," *Educational Measurement: Issues and Practice*, 8, 157.

Kennedy, C. A., Wilson, M., Draney, K., Tutunciyan, S., and Vorp, R. 2006, *ConstructMap Software*, Berkeley, CA: Berkeley Evaluation and Assessment Research (BEAR) Center. Available at: http://bearcenter.berkeley.edu/ConstructMap.

Masters, G. N. 1982, "A Rasch Model for Partial Credit Scoring," *Psychometrika*, 47, 149.

Rasch, G. 1980, *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago, IL: University of Chicago Press. Originally published in 1960, Copenhagen, DK: The Danish Institute for Educational Research.

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. 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. 2011b, "A Study of General Education Astronomy Students' Understandings of Cosmology. Part II. Evaluating Four Conceptual Cosmology Surveys: A Classical Test Theory Approach," *Astronomy Education Review*, 10, 010107.

Wallace, C. S., Prather, E. E., and Duncan, D. 2012a, "A Study of General Education Astronomy Students' Understandings of Cosmology. Part IV. Common Difficulties Students Experience with Cosmology," *Astronomy Education Review*, 11, 010104.

Wallace, C. S., Prather, E. E., and Duncan, D. 2012b, "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.

Yen, W. M. 1993, "Scaling Performance Assessments: Strategies for Managing Local Item Dependence," *Journal of Educational Measurement*, 30, 187.

ÆR
010103-1–010103-13