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

Astronomy Education Review

2012, AER, 11, 010107-1, 10.3847/AER2012010

A Classical Test Theory Analysis of the Light and Spectroscopy Concept Inventory National Study Data Set

Wayne M. Schlingman
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
Colin S. Wallace
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
Alexander L. Rudolph
Department of Physics and Astronomy, California State Polytechnic University, Pomona, California, 91768
Gina Brissenden
Center for Astronomy Education (CAE), Steward Observatory, University of Arizona, Tucson, Arizona 85721
Received: 03/19/12, Accepted: 05/14/12, Published: 06/22/12

Abstract

This paper is the first in a series of investigations into the data from the recent national study using the Light and Spectroscopy Concept Inventory (LSCI). In this paper, we use classical test theory to form a framework of results that will be used to evaluate individual item difficulties, item discriminations, and the overall reliability of the LSCI. We perform an analysis of individual students' normalized gains, providing further insight into the prior results from this data set. This investigation allows us to better understand the efficacy of measuring student achievement using the LSCI. Future papers will discuss our investigation of the data from the recent national study using item response theory (IRT).

1. INTRODUCTION

In this paper, we perform a classical test theory (CTT) analysis on the matched data from the national study of teaching and learning in introductory general education astronomy using the Light and Spectroscopy Concept Inventory (LSCI; Prather *et al.* 2009; Rudolph *et al.* 2010). Future papers will discuss our investigation of this same data using item response theory (IRT). The LSCI contains 26 conceptual questions designed to probe the ideas and reasoning abilities of students in college-level, general education, introductory astronomy courses (hereafter Astro 101) over a commonly taught set of topics involving properties of light, the Stefan-Boltzmann law, Wien's law, the Doppler shift, and spectroscopy (Bardar *et al.* 2007). Nearly 4000 participants from 69 classes at thirty-one 2-year and 4-year institutions from across the United States comprise the LSCI national dataset (Prather *et al.* 2009, Rudolph *et al.* 2010). In previous publications, we describe how we used this national dataset to investigate the relationship between interactive teaching and classes' learning gains (Prather *et al.* 2009), and how interactive instruction and students' ascribed (e.g., race) and achieved characteristics (e.g., college grade point average) are related to student learning (Rudolph *et al.* 2010). From the full dataset, we select the students with matched pre-instruction and post-instruction responses to all items on the LSCI ($N = 1881$ students) to better understand how individual students' understandings change as the result of having taken an Astro 101 course.

For the matched dataset, we look at changes in individual students' scores and their individual normalized gains. We compare the ranges of scores and individual gains to the class averages previously reported (Prather *et al.* 2009).

In addition to the analysis of individual participants, we perform an item analysis of the data on students' responses to individual items on the LSCI. We compute the item parameters from both the pre-instruction and post-instruction data separately to determine which concepts show the largest improvement in student reasoning about light and which items may need to be revised. This analysis adds to the research base on the efficacy of what the individual items comprising the LSCI measure and the reliability of the LSCI instrument as a whole.

This paper is organized as follows. Section 2 describes the parameters of individual items (i.e., item difficulty and discrimination) and how they change from pre-instruction to post-instruction. Section 3 describes the reliability measurement of the LSCI. Section 4 displays and discusses the wide range of possible student gains and compares them to the class averages previously reported. Section 5 provides a final discussion of our findings, and Section 6 contains the conclusions from this work.

2. LSCI ITEM PARAMETERS

Classical test theory (CTT) provides a framework for measuring the item parameters and reliability of the LSCI. A CTT analysis allows us to compute the item difficulties and discrimination values for each item on the LSCI. We define item difficulty to be the fraction of students responding *incorrectly* to an item. It ranges from 0.0 to 1.0, with larger values indicating more challenging, difficult items. This is different from the oft-used P-value, which is defined as the fraction of students selecting the *correct* response (Crocker and Algina 1986). P-values are sometimes confusing to interpret as measures of items' difficulties, since easier items have larger P-values. We use our definition of item difficulty for clarity, in order to make harder items have larger difficulty values. The range of conventionally accepted values for item difficulty is between 0.2 and 0.8 (Bardar *et al*. 2007). Item discrimination is defined by the value of the point biserial, which is the correlation between students' scores on an individual item and students' total scores on the LSCI as a whole (Lord and Novick 1968). An item's discrimination ranges from -1.0 to $+1.0$, with a value of zero meaning there is no correlation. A negative point biserial indicates that a student's success on the instrument is anticorrelated with a correct response to the item (an indication of a problem with the item). The greater the value of the item's discrimination, the better an item is at selecting high performing students from low performing students. Conventionally accepted values for item discrimination are typically between 0.3 and 0.7 (Bardar *et al*. 2007, Allen and Yen 1979).

Since all CTT statistics are highly sample dependent (Hambleton and Jones 1993, Thompson 2003), we expect the items' difficulties and discriminations to be different when they are calculated using only pre-instruction responses versus when they are calculated using only post-instruction responses. Table 1 shows each item's pre-instruction and post-instruction difficulty and discrimination. As expected, the values of item difficulty and discrimination change pre-instruction to post-instruction.

We bolded item parameters in Table 1 that fall outside of the conventionally accepted range of parameter values. Pre-instruction, all but 3 of the 26 items on the LSCI are flagged, and all of these flagged items have difficulties greater than 0.80, discriminations less than 0.20, or both. These results make sense if we take into account that the vast majority of students come to Astro 101 with little prior explicit instruction on the nature of light and spectroscopy in the context of astronomy and have many commonly held naïve ideas and reasoning difficulties that are elicited by the items on the LSCI (Bardar *et al*. 2007, Rudolph *et al*. 2010, Deming and Hufnagel 2001). The high pre-test difficulty values of these items illustrate that the items are hard for students to reason through. Additionally, the low pre-test discrimination values indicate that these items challenge both high and low scoring students equally. The LSCI is challenging for most Astro 101 students prior to instruction; however, as we shall see in Section 4, many Astro 101 students are able to correctly answer a majority of the items on the LSCI pre-instruction.

The post-instruction item parameters show a significant change from those determined from pre-instruction responses. The majority of items decrease in difficulty and increase in discrimination. This makes sense. As students learn to reason about light and spectroscopy, the LSCI's items become easier, and students who do well on the LSCI overall also are more likely to answer correctly on individual items. These patterns can be seen in Figure 1, which plots the post-instruction difficulty values of the LSCI's items as a function of their pre-instruction difficulties, and in Figure 2, which plots the post-instruction discrimination values of the LSCI's items as a function of their pre-instruction discriminations. Figure 3 combines the information in Figures 1 and 2 into one graph by plotting, for each item, the differences between the post-instruction and pre-instruction

Table 1. **The LSCI's items' difficulties and discriminations, calculated for both pre-instruction and post-instruction student responses. Bolded items are outside the conventionally accepted parameter ranges for item difficulty and/or discrimination**

LSCI item	Pre-difficulty	Pre-discrimination	Post-difficulty	Post-discrimination
1	**0.81**	0.36	0.37	0.49
2	**0.89**	**0.11**	0.76	0.53
3	**0.83**	**0.25**	0.76	**0.25**
4	**0.81**	**0.30**	0.31	0.44
5	0.64	**0.27**	0.27	0.37
6	0.77	**0.23**	0.47	0.46
7	**0.80**	**0.23**	0.55	0.46
8	**0.86**	0.31	0.56	0.53
9	**0.84**	**0.24**	0.63	0.50
10	0.75	**0.24**	0.61	0.37
11	**0.82**	**0.15**	0.63	0.44
12	0.65	0.34	0.39	0.47
13	0.74	0.32	0.48	0.47
14	0.35	**0.26**	**0.20**	0.33
15	0.77	0.39	0.41	0.56
16	0.40	**0.28**	0.40	0.36
17	**0.81**	**0.17**	0.73	0.35
18	0.72	**0.22**	0.61	0.37
19	**0.82**	**0.29**	0.63	0.46
20	0.69	**0.28**	0.48	0.34
21	0.71	**0.14**	0.79	**0.12**
22	**0.88**	**0.20**	0.70	0.51
23	**0.86**	0.41	0.47	0.58
24	0.78	**0.19**	0.51	0.44
25	**0.90**	**0.11**	**0.89**	**0.28**
26	0.75	**0.15**	0.70	**0.27**

discrimination values (which, for almost all items, is positive) versus the differences between the post-instruction and pre-instruction difficulty values (which, for almost all items, is negative, indicating that items become easier pre-instruction to post-instruction). Figures 1–3 together show that the majority of items decrease in difficulty and increase in their discriminatory abilities pre-instruction to post-instruction.

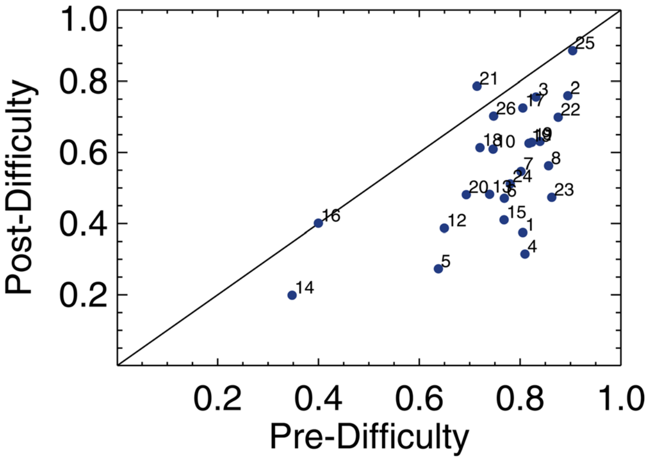


Figure 1. The pre-instruction difficulty and post-instruction difficulty for each item of the LSCI are plotted along the horizontal and vertical axes, respectively. Items that become easier to answer after instruction will lie below the diagonal line

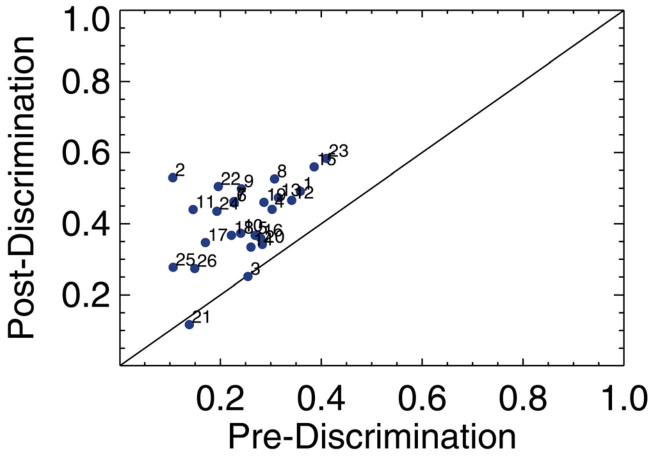


Figure 2. The pre-instruction discrimination and post-instruction discrimination for each item of the LSCI are plotted along the horizontal and vertical axes respectively. Items that increase in discrimination after instruction will lie above the diagonal line

Nevertheless, a few items are flagged as having post-instruction parameters outside of their conventionally accepted ranges of values. These flagged items are Items 3, 14, 21, 25, and 26. Below we provide a brief discussion of each of these items and will provide a more detailed discussion of the particularly problematic of these items in Section 5.

Items 3 and 26 are each flagged because of their somewhat low post-instruction discrimination values. However, each item shows increased performance from pre-instruction to post-instruction, as they each decrease in difficulty and do not decrease in discrimination. These results are illustrated in Figures 1–3. In addition, Item 26 falls within the conventionally accepted range for difficulty both pre-instruction and post-instruction and substantially increases in discrimination (falling very close to the acceptable range for discrimination post-instruction). Because the overall performance of Item 26 is at or near the acceptable range for both difficulty and discrimination post-instruction, we argue to keep Item 26 as is. Further, the most common incorrect choice selected by students for this item, both pre- and post-, elicits students' inability to differentiate between the size of objects from the total energy output and peak wavelength provided in their light curves. However, since the discrimination value of Item 3 remains constant, and below the conventionally accepted range, pre-instruction to post-instruction, it will be an item we discuss in greater detail in Section 5.

Item 14 is flagged for having a low post-instruction difficulty (0.20, which means 80% of students responded correctly). However, even though almost all students respond correctly to this item, we find that the

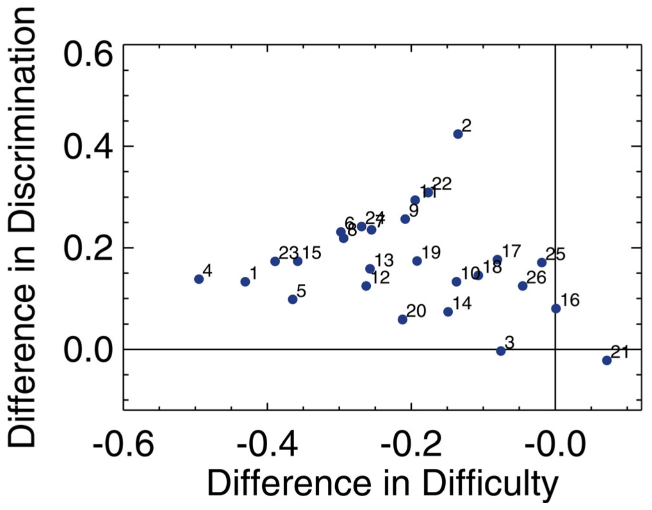


Figure 3. The difference in difficulty, post- minus pre-, for each item of the LSCI, is plotted on the horizontal axis. Items that become easier post-instruction will be farther to the left along the horizontal axis. The difference in discrimination, post-minus pre-, for each item, is plotted on the vertical axis. Items with an increased discrimination value post-instruction will be farther up the vertical axis. The items showing improvement due to instruction are shown above the horizontal line and to the left of the vertical line

post-instruction discrimination value (0.33) is actually quite good. Additionally, this item is well matched to the *pre-instruction* knowledge and reasoning abilities of many Astro 101 students with 65% of students responding correctly. For these reasons, we argue to keep Item 14 as is.

Item 21 is the second most challenging item on the LSCI with a difficulty of 0.79 post-instruction. However, it is flagged because its discrimination value is low both pre-instruction and post-instruction (0.14 and 0.12, respectively). Item 21 also stands out in Figures 1–3 since its difficulty increases and its discrimination decreases pre-instruction to post-instruction. The fact that fewer students responded correctly post-instruction compared to pre-instruction indicates that there might be an underlying problem with the item. We will discuss this further in Section 5.

Item 25 is the most difficult item on the entire instrument both pre-instruction and post-instruction. Only approximately 10% of students respond correctly to this item. However, its discrimination value improves from 0.11 to 0.28 pre-instruction to post-instruction. This increase means the top performing students are correctly reasoning about the concepts probed by Item 25 after instruction. We will further discuss this item in Section 5.

3. LSCI RELIABILITY

The final CTT parameter we calculate for the LSCI is a measure of the overall reliability of the instrument itself: Cronbach's α. Cronbach's α provides a lower limit on the internal consistency of observed responses. Cronbach's α is given by

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

where N is the total number of items on a test, σ_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). The value of Cronbach's α is close to one when the covariances among items are high, and it is close to zero when the covariances among items are low. Conventionally, any value of α over 0.75 is considered good (George and Mallery 2003). The Cronbach's α for the LSCI pre-instruction is $\alpha = 0.37$. This value is low because Cronbach's α is sensitive to the homogeneity of the test-taking population, and pre-instruction, as a population, Astro 101 students have many of the same naïve ideas and reasoning difficulties related to the ideas probed by the LSCI (Thompson 2003). The Cronbach's α for the LSCI post-instruction is $\alpha = 0.78$. This high value of Cronbach's α post-instruction provides strong evidence for the reliability of the LSCI. To check for any variations in consistency, we computed α 26 additional times for a "25 item LSCI," each time removing one of the items from the actual LSCI. We find very little variation in the 26 recomputed values of α for the "25 item LSCIs" (the standard deviation of these αs is 0.004), providing further evidence that the items on the LSCI are internally consistent.

4. STUDENT GAINS ON THE LSCI

The analysis provided above, along with our prior work investigating changes in the averaged normalized gains achieved by different classrooms across the country (Prather *et al*. 2009), provide significant evidence that the LSCI can effectively measure changes in students' conceptual knowledge of light and their abilities to reason about such topics. We now look at the pre-instruction versus post-instruction gains of individual students from our national dataset.

From the prior analysis of overall class performance using the national LSCI data set, we found that all 69 different classes had average pre-test scores of approximately 25% (24 ± 2%). This score is the score one might expect a class would earn if all students in the class were simply guessing on all the items (assuming all items are multiple-choice questions with four answer choices). The actual guessing score computed for the items of the LSCI is 23%. We computed this guessing score by averaging the probabilities of guessing the correct answers on each of the LSCI's items. From the matched dataset analyzed in this section, we demonstrate that, in-fact, the 25% pre-test score of classes is not due to all students simply guessing.

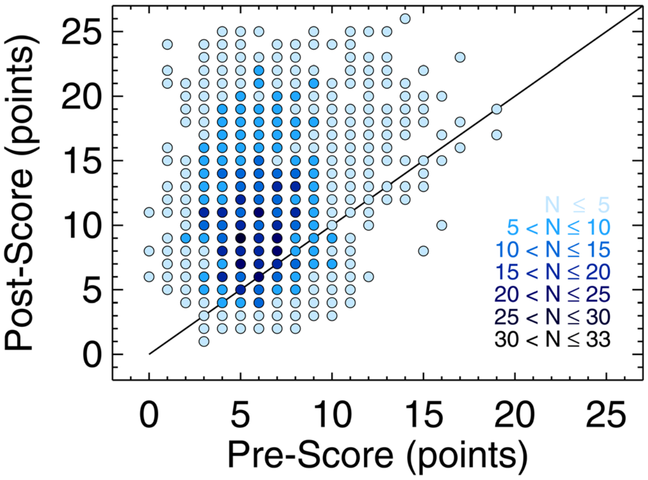


Figure 4. The post-instruction score versus pre-instruction score for the 1881 students with matched data (blue circles). The different shades of blue represent the number of students that have the pre-instruction post-instruction score combination at a given point. The line provided identifies the location of equal pre-instruction and post-instruction scores

Figure 4 shows the post-test versus pre-test scores (out of 26) for individual students' responses. The different shades of blue indicate the total number of students who have the scores corresponding with each data point. From Figure 4, we see that the pre-test scores for individual students range from 0 (students responding incorrectly to every item), to around 20 (students responding correctly to 80% of the items). Pre-test, we find that a significant number of students (43%) outperform the class average pre-test score of 25%, and a full 60% perform better than or equal to the true LSCI guessing score of 23%. In addition, it is worth noting that only approximately one-third of the LSCI's items have a nearly equal distribution of responses across all answer choices for that item (consistent with guessing). In contrast, the majority of items have one or two answer choices that dominate the distribution of answer choices selected by students. This suggests that students either know the correct answer or have naïve ideas and reasoning difficulties that are well matched to the concepts probed by the items of the LSCI.

The black line in Figure 4 identifies the location of equal pre-test and post-test scores. Eighty-two percent of students lie above this line, illustrating that the vast majority of students across the nation are performing better after instruction. To further describe students' performance, in comparison to class performance, we now will analyze the normalized gain score for each student. Figure 5 shows the range of students' normalized gain scores versus pre-test scores for the matched-data set. As expected, while we do see some cases where students are performing worse after instruction and have negative gains, the overwhelming majority of students improve as the result of instruction. While the range in normalized gain scores for classes ($-0.07 < \langle g \rangle < 0.5$) helped us to understand the difference in performance between classes in our national study, there is clearly a much wider

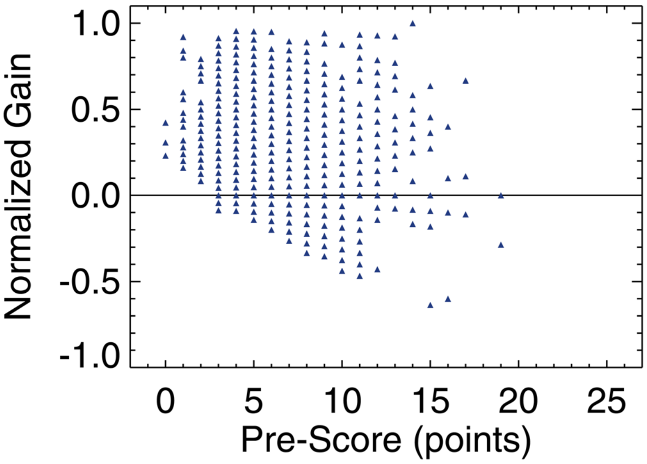
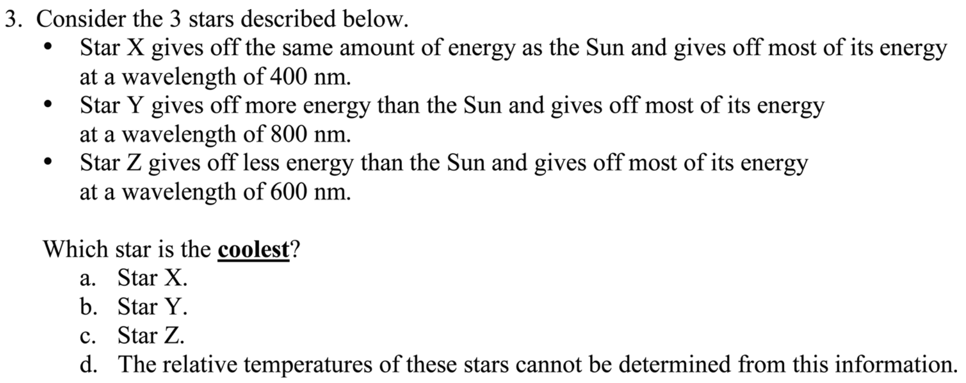


Figure 5. The normalized gain versus pre-instruction score for the 1881 students with matched data (blue triangles). The horizontal line is the zero normalized gain line. Any point below this line corresponds to students that did worse post-instruction than they did pre-instruction



3. Consider the 3 stars described below.
    - Star X gives off the same amount of energy as the Sun and gives off most of its energy at a wavelength of 400 nm.
    - Star Y gives off more energy than the Sun and gives off most of its energy at a wavelength of 800 nm.
    - Star Z gives off less energy than the Sun and gives off most of its energy at a wavelength of 600 nm.

    Which star is the **coolest**?
    a. Star X.
    b. Star Y.
    c. Star Z.
    d. The relative temperatures of these stars cannot be determined from this information.

Figure 6. Item 3 from the LSCI

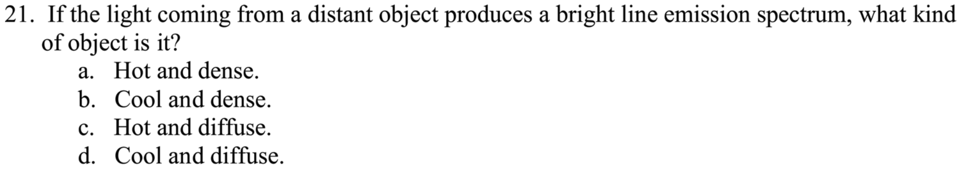
range of normalized gain scores achieved by individual students. This is a powerful indication that some students clearly outperform their class average regardless of the instructor, class-size, type of institution, or their pre-test score. In fact, one-fifth of students outperform the highest-class normalized gain of $\langle g \rangle = 0.5$, with over 120 students even making it into the "high" normalized gain region $\langle g \rangle \geq 0.7$ (Hake 1998).

5. DISCUSSION

We return to the discussion of the three specifically problematic items from Section 3—Items 3, 21, and 25. Each item was originally flagged due to having post-instruction difficulty and discrimination values outside of the conventionally accepted ranges. Each of these items targets a particular known student naïve idea or reasoning difficulty and is therefore potentially valuable to the LSCI regardless of being flagged for low performance post-instruction. It is worth noting, as stated earlier, that the reliability of the LSCI, as measured by Cronbach's α, is essentially unaffected by removing any one of these questions.

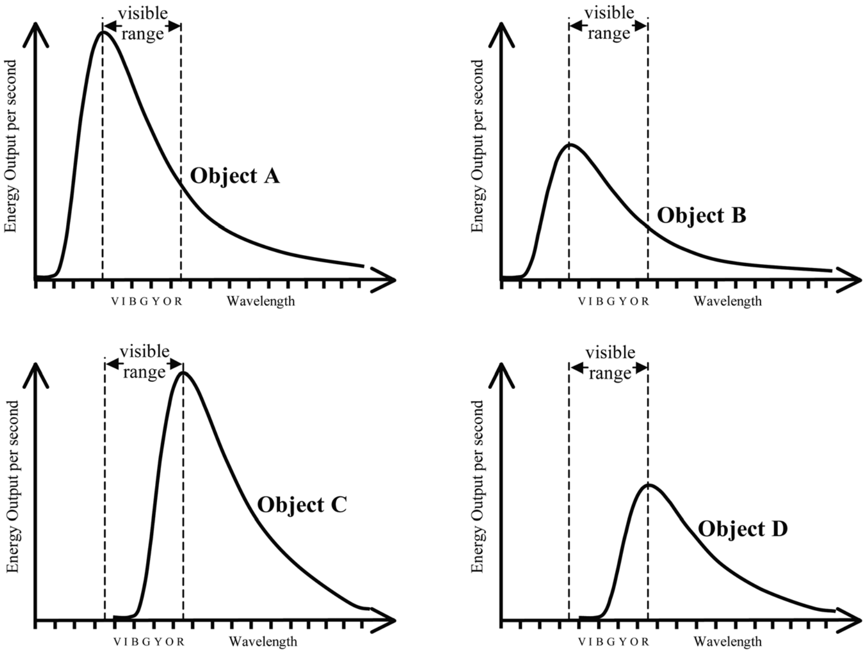
We flagged Item 3 because of its low post-instruction discrimination value of 0.25. While the difficulty of the item goes down after instruction, the discrimination does not change from pre-instruction to post-instruction. Only Items 3 and 21 do not show improvement in their discrimination after instruction. Figure 6 shows Item 3. After instruction, only 24% of students are selecting Star Y, the object with the longest peak wavelength. Nearly 40% of students provide an answer consistent with the reasoning that the object giving off the least amount of energy will be coolest independent of the wavelength of the object's peak energy output (Star Z). This reasoning is consistent with a student who allocates a proportionality mental resource (often referred to as "Ohm's Primitive"; diSessa 1983, diSessa 2000) to answer the question: "more energy given off means a greater temperature." It is important to note that this is exactly the reasoning difficulty this question was designed to elicit. Since this item targets a known reasoning difficulty and continues to offer an enticing distracter even for those students that perform well on the LSCI as a whole, we advocate that it remain in the item bank of the LSCI. There are intentional similarities in the information presented and task offered to students between Item 3 and Item 24. In both cases, students are provided with information about total energy output and peak wavelength of objects, and asked to reason about comparisons in temperature. However, in one case (Item 24), the information is provided in the more familiar graphical representation that is common to traditional classroom instruction on this topic. It is interesting how much better Item 24 performs in comparison to Item 3, especially in terms of item discrimination. Because Item 3 has an unchanging discrimination from pre- to post- we think that it should be studied for future revision. However, this item may simply be indicating a challenge to how we traditionally go about teaching this set of ideas.

Item 21 was flagged for its very low post-instruction discrimination value of 0.12. Item 21 is shown in Figure 7. Post-instruction, 75% students choose either option "a" or option "c," indicating they know the object must be



21. If the light coming from a distant object produces a bright line emission spectrum, what kind of object is it?
    a. Hot and dense.
    b. Cool and dense.
    c. Hot and diffuse.
    d. Cool and diffuse.

Figure 7. Item 21 from the LSCI



25. Which, if any, of the objects could be approximately the same size as object D?
    a. Object A.
    b. Object B.
    c. Object C.
    d. They could all be the same size.
    e. None of the above.

Figure 8. Item 25 from the LSCI

hot. However, 55% of students incorrectly select option "a" on the post-test, suggesting that they do not understand whether the object should be "dense" or "diffuse." Because of the very low discrimination value for this question, it is clear that students struggle to correctly connect the words "dense" and "diffuse" with the processes that create emission spectra. During the original design and validity testing of the LSCI, professors consistently chose the correct response for this question (Bardar *et al.* 2007). However, the low post-instruction discrimination of this item suggests that the vocabulary used in the response choices of this item may be causing unnecessary confusion for students, and so we recommend this item be studied for revision in the future.

The last item we will discuss is Item 25, shown in Figure 8. This item was flagged primarily for its high post-instruction difficulty (0.89). Only 11% of students are able to answer the item correctly after instruction, which is essentially the same as we saw from the pre-test. This is the most difficult item on the instrument and is known to challenge even professional astronomers. The post-instruction discrimination value for this item is very near 0.3, indicating that the highest achieving students do tend to get this question right. Fifty-one percent of the 120 students with normalized gain scores greater than or equal to 0.7, considered "high" scores (Hake 1998), answer this question correctly. Approximately 45% of students post-instruction choose the object with the curve that is the same height as Object D. These students are using "energy output," or "height on a graph," to define the size of an object. This item was specifically designed to elicit this well-known reasoning difficulty commonly held by this population. It is important to have items that are challenging even for the highest achieving students; we believe that Item 25 is just such an item and should be kept as is.

6. CONCLUSIONS

Overall, this analysis has shown that the LSCI is challenging for students pre-instruction, and that it is well-matched to the understandings of these students after instruction. The items of the LSCI are well matched to assess students' conceptual and reasoning abilities over a commonly taught set of topics involving properties of light, the Stefan-Boltzmann law, Wien's law, the Doppler shift, and spectroscopy.

From this CTT analysis, we have shown that the difficulty of items on the LSCI decreases and the discrimination of items increases from pre-instruction to post-instruction, with only a few exceptions. Furthermore, we have shown that the reliability of the LSCI, as measured by Cronbach's α, is well above the conventionally accepted values. These CTT results show that the LSCI is very capable of measuring changes in student's understanding resulting from a course in introductory astronomy. Further, we have shown that the class averaged normalized gain scores reported in the previous study were hiding the incredible range of performance of individual students both pre-instruction and post-instruction. We were surprised to find that there were so many students getting between 50% and 80% of the items correct pre-instruction, and even more surprised to discover that so many students achieved normalized gains above 0.7. The CTT analysis of the individual items has shown that most of the items of the LSCI are measuring a wide range of students' conceptual understandings.

This CTT analysis has provided new insights into the performance of individual items of the LSCI and the performances of individual students who participated in our study. This allows us to better determine the effectiveness of the LSCI as an introductory astronomy concept inventory and as a tool for measuring changes in a students' understandings. However, this analysis is still dependent on the scores of the specific students who took the LSCI pre-instruction and post-instruction. Our continued work on this data set uses an item response theory analysis, which allows us to characterize the item parameters independent of the population and characterize the inherent reasoning abilities of students independent of their score on the LSCI (Wallace and Bailey 2010).

Acknowledgments

Once again we would also like to extend a sincere thank you to the instructors and students who participated in the LSCI study. This material is based in part upon work supported by the National Science Foundation under Grant No. 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

Allen, M. J., and Yen, W. M. 1979, *Introduction to Measurement Theory*, Long Grove, Illinois: Waveland Press, Inc.

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.

Crocker, L. M., and Algina, J. 1986, *Introduction to Classical and Modern Test Theory*, New York, NY: Holt, Rinehart, and Winston.

Deming, G., and Hufnagel, B. 2001, Who's taking ASTRO 101?, *Physics Teacher*, 39(6), 368.

diSessa, A. 1983, "Phenomenology and Evolution of Intuition," in *Mental Models*, ed. A. L. Stevens, Hillsdale, NJ: Lawrence Erlbaum Associates, 25.

diSessa, A. 2000, *Changing Minds: Computers, Learning, and Literacy*, Boston, MA: MIT Press, 90.

George, D., and Mallery, P. 2003, *SPSS for Windows Step by Step: A Simple Guide and Reference*, 4th ed., 11.0 Update, Boston, MA: Allyn & Bacon/Pearson,

Hake, R. R. 1998, "Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, 66, 64.

Hambleton, R. K., and Jones, R. J. 1993, "Comparison of Classical Test Theory and Their Applications to Test Development," *Education Measurement: Issues & Practices*, 12, 253.

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

Prather, E. E., Rudolph, A. L., Brissenden, G., and Schlingman, W. M. 2009, “A National Study Assessing the Teaching and Learning of Introductory Astronomy. Part I. The Effect of Interactive Instruction,” *American Journal of Physics*, 77, 320.

Rudolph, A. L., Prather, E. E., Brissenden, G., Consiglio, D., and Gonzaga, V. 2010, “A National Study Assessing the Teaching and Learning of Introductory Astronomy. Part II: The Connection between Student Demographics and Learning,” *Astronomy Education Review*, 9, 010107.

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.