Item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set

Colin S. Wallace[*]
Department of Physics and Astronomy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

Timothy G. Chambers[†]
Department of Materials Science and Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA

Edward E. Prather[‡]
Center for Astronomy Education (CAE), Department of Astronomy, Steward Observatory, University of Arizona, Tucson, Arizona 85721, USA

(Received 14 May 2017; published 15 June 2018)

[This paper is part of the Focused Collection on Astronomy Education Research.] This paper presents the first item response theory (IRT) analysis of the national data set on introductory, general education, college-level astronomy teaching using the Light and Spectroscopy Concept Inventory (LSCI). We used the difference between students' pre- and postinstruction IRT-estimated abilities as a measure of learning gain. This analysis provides deeper insights than prior publications both into the LSCI as an instrument and into the effectiveness of teaching and learning in introductory astronomy courses. Our IRT analysis supports the classical test theory findings of prior studies using the LSCI with this population. In particular, we found that students in classes that used active learning strategies at least 25% of the time had average IRT-estimated learning gains that were approximately 1 logit larger than students in classes that spent less time on active learning strategies. We also found that instructors who want their classes to achieve an improvement in abilities of average $\Delta\theta = 1$ logit must spend at least 25% of class time on active learning strategies. However, our analysis also powerfully illustrates the lack of insight into student learning that is revealed by looking at a single measure of learning gain, such as average $\Delta\theta$. Educators and researchers should also examine the distributions of students' abilities pre- and postinstruction in order to understand how many students actually achieved an improvement in their abilities and whether or not a majority of students have moved to postabilities significantly greater than the national average.

DOI: 10.1103/PhysRevPhysEducRes.14.010149

I. INTRODUCTION

We report on our item response theory (IRT) analysis of a national data set of 3205 students' matched pre- and postresponses to the Light and Spectroscopy Concept Inventory (LSCI) [1,2]. The LSCI is a twenty-six item multiple-choice assessment instrument designed to measure students' conceptual understandings and reasoning abilities on topics involving the properties of light, the luminosity-area-temperature relationship, Wien's law, the Doppler shift, and spectroscopy. All students in the data set were enrolled in one of sixty-nine different introductory, general education, college-level astronomy classes (hereafter, Astro 101) from across the United States (with one class in Ireland), representing twenty-nine colleges and universities, including associate (2-year) colleges, baccalaureate colleges (4-year primarily bachelor's-granting institutions), master's colleges and universities (4-year primarily master's- and bachelor's-granting universities), and doctorate-granting (research) universities. Class sizes ranged from less than ten to more than 400 students. In a previous publication, students' responses from this national data set were used to investigate the relationship between interactive teaching, classes' learning gains, and class size and institution type [3]. Subsequent studies examined how interactive instruction and students' ascribed (e.g., race) and achieved characteristics (e.g., college grade point

[*] cswallace@email.unc.edu
[†] timchamb@umich.edu
[‡] eprather@as.arizona.edu

average) are related to students' learning [4], and how classical test theory (CTT) statistics and individual students' performances change pre- to postinstruction [5].

IRT has a number of potential advantages over CTT with respect to the analysis of concept inventory data. CTT statistics do not estimate the underlying abilities of students independent of the items to which they responded [6]. In contrast, when the assumptions of IRT hold and the model fits the data, an IRT analysis can estimate students' abilities and item properties independent of one another [7]. IRT models have been used in an increasing number of physics and astronomy education investigations, including analyses of the Force Concept Inventory [8–10], the Mechanics Baseline Test [11], the Conceptual Survey of Electricity and Magnetism [12], the Star Properties Concept Inventory [13], the Newtonian Gravity Concept Inventory [14], and the Astronomy and Space Science Concept Inventory [15].

We performed an IRT analysis of the LSCI national data set in order to move beyond the limitations of CTT, gain further insights into the functioning of the LSCI's items, test the robustness of our earlier analyses of the LSCI national data set, and investigate the capacity of active engagement instruction to evolve individual students' underlying astronomy reasoning abilities. This paper is organized as follows. In Sec. II, we demonstrate that our data set satisfies the assumptions of IRT and that the model fits the data. Section III presents the results of our analysis. Our conclusions are in Sec. IV.

II. DATA ANALYSIS

A. Selecting an IRT model

For our analysis, we initially attempted to fit both a two parameter logistic (2PL) and a three parameter logistic (3PL) model to the data. The 2PL model can be written as

$$P(X_{pi}=1|\theta_p,a_i,b_i)=\frac{\exp[a_i(\theta_p-b_i)]}{1+\exp[a_i(\theta_p-b_i)]}, \qquad (1)$$

where $P(X_{pi}=1)$ represents the probability that a student p of ability θ_p correctly answers an item i with difficulty b_i and discrimination a_i. The 3PL model is similar to the 2PL model, except the former includes a third item parameter c_i, which is called the guessing parameter. This guessing parameter takes into account the fact that there may be some items for which students with extremely low abilities θ_p still have a nonzero probability of giving the correct response. The 3PL model can be written as

$$P(X_{pi}=1|\theta_p,a_i,b_i,c_i)=c_i+(1-c_i)\frac{\exp[a_i(\theta_p-b_i)]}{1+\exp[a_i(\theta_p-b_i)]}. \qquad (2)$$

Readers looking for a pedagogical treatment of these IRT models should consult Embretson and Reise [16], Hambleton and Jones [6], Harris [17], or Wallace and Bailey [13], and references therein.

We used the IRTPRO software [18] to estimate item parameters and student abilities. We selected the MML estimation procedure for estimating item parameters and the EAP estimation procedure for estimating students' abilities; see Baker and Kim [19] for details on these estimation procedures. It is very important to note that the logit scale was anchored such that the mean ability of the postinstruction scores is 0 logits.

We first tried to fit the 2PL and 3PL models to both the pre- and postinstruction responses of all 3205 students to all twenty-six of the LSCI's items. However, when we calibrated the items to the preinstruction responses, we got quite different results from when we calibrated the items to the postinstruction responses. Furthermore, all our attempts to fit 2PL and 3PL models to the preinstruction data consistently yielded poor goodness-of-fit statistics. This result makes sense. We previously found that the average preinstruction scores for classes were clustered in the very narrow range of 24% ± 2% [3]. When we look at individual students' preinstruction scores, we find that 57% of students score at or below 25% correct, which is the most probable score one would expect to receive if one is purely guessing [5]. This strongly suggests that, preinstruction, many students possess very little of the latent trait measured by the LSCI, which severely limits the utility of the preinstruction data for producing accurate estimates of the item parameters. Consequently, we used only students' postinstruction responses to estimate the item parameters. We then used these established item parameter values when we estimated students' preinstruction abilities.

We found that the χ^2 goodness-of-fit statistics for many individual items on the LSCI were significantly better for the 3PL model than the 2PL model. While adding another free parameter (c_i) will almost always improve model fit, the degree to which the fit improved was greater than one would expect from simply adding a free parameter. This can be seen by calculating the root mean square error of approximation (RMSEA) [20] for both the 2PL and 3PL models. The RMSEA is a statistic that measures the goodness-of-fit of a model relative to the number of parameters in the model such that merely adding a new parameter cannot reduce the value of the RMSEA unless the new parameter actually models a real feature present in the data (e.g., a lower asymptote in the probability of a correct response *á la* the 3PL model). The 2PL model's RMSEA is 0.07 while the 3PL model's RMSEA is 0.06. This suggests that guessing was a significant factor in many students' responses on many items. While some items ended up with values of c_i near zero (suggesting that these items had many powerful distractors), other items saw as many as 40% of low-ability students answer correctly.

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.

FIG. 1. Item 21 from the LSCI.

Items with high guessing parameters tended to be those with only one or two frequently chosen distractors. In a previous publication, we found that these one or two distractors, plus the correct answer, tend to dominate the answer choices actually selected by students, which implies that these distractors are well matched to students' conceptual and reasoning difficulties [5]. Because guessing appears to be an important component of students' responses, we abandoned the 2PL model. All results reported in the rest of this paper were obtained using the 3PL model.

Before we proceeded with testing for potential violations of IRT's fundamental assumptions, we dropped two of the LSCI's items from our analysis: Items 21 and 25. Item 21 (Fig. 1) had extremely poor goodness-of-fit statistics (e.g., $\chi^2 \approx 180$ with 22 degrees of freedom), regardless of the model used. We found no clear relationship between student ability and success on item 21. We already suspected item 21 might be inappropriate for this population based on our previous CTT analysis, which revealed that it had an extremely low discrimination value, which actually decreased from 0.14 to 0.12 pre- to postinstruction [5]. Item 21 requires students to understand that a hot, diffuse cloud of gas produces a bright line emission spectrum and that a dense hot object does not, which distinguishes choices "a" from "c." While 75% of students postinstruction select either choice a or choice c, over half of those students selected a, suggesting that many students do not understand the distinction between a "dense" and a "diffuse" object, even though they recognize that a "bright line emission spectrum" must come from a hot object [5]. This item fails to probe the latent trait of interest since students' responses are dominated by their knowledge of the definitions of these words.

Item 25 had the largest difficulty parameter of any item on the LSCI ($b_{25} = 21$). Item 25 presents students with graphs of energy output per second as a function of wavelength for four different objects (A–D); students must determine which object, if any, could be the same size as object D. The probability of a student correctly answering this item remains low across all abilities of students in the study population. The reasoning required to correctly answer item 25 challenges many professional astronomers, and we previously found its postinstruction CTT difficulty to be 0.89, with only 11% of students giving the correct answer [5]. Because student success on this item was very weakly correlated with ability, it yielded essentially no useful information about students' abilities, while degrading the overall goodness of fit of the data to the model.

After removing both item 21 and item 25, we examined whether or not we satisfied the two fundamental assumptions of IRT: local independence and unidimensionality. If both of these assumptions hold, then the IRT model possesses the property of parameter invariance, which means that estimates of students' abilities do not depend on the specific items administered and estimates of item parameters do not depend on the abilities of students responding to those items [7].

B. Local independence

An item is locally independent if the probability of correctly answering that item is entirely determined by a student's ability θ_p and not by his or her responses to other items or other sources of unaccounted-for variance [16]. We used Yen's Q3 statistic to look for violations of local independence [21]. For each pair of items, Yen's Q3 statistic is the linear correlation between the items' residuals (i.e., the difference between students' observed and 3PL model-predicted scores). If student ability θ_p is the only latent trait that determines the probability of correctly answering items, then there should be essentially no correlation between the residuals of two different items. Yen and Fitzpatrick recommend flagging item pairs for which the value of $|\mathrm{Q3}| > 0.20$ [22].

We found that the following pairs of items had values of $|\mathrm{Q3}| > 0.20$: Items 7 and 8, items 18 and 19, and items 2 and 22. Before discussing how we dealt with these violations of local independence, we must stress that just because these items have high Q3 values does not mean they are "bad" items. To the contrary, Schlingman *et al.*'s CTT analysis suggests that all of these items possess favorable psychometric properties [5]. If the flagged items are not bad, then why do they have high Q3 values? Take, for example, items 18 and 19 (Fig. 2). Item 18 asks students to identify which of four spectra corresponds to an object at rest, while item 19 asks students to identify which spectrum corresponds to the object moving the slowest toward the observer. This item pair had a high Q3 value ($\mathrm{Q3} = 0.51$) because the probability of correctly answering item 19 is not independent of the probability of correctly answering item 18. This pair of items exhibits what Yen calls "item chaining," which means that one item builds off of the previous item such that knowing the answer to one item increases one's probability of correctly answering the other [23]. Someone who gives the correct answer to item 18 has a much higher probability of giving the correct answer to item 19, regardless of his or her ability level.

The other item pairs with high Q3 values also exhibit item chaining. Items 7 and 8 require students to determine which pictorial representation of the Bohr model of the atom corresponds to the formation of an absorption line and

Use the four spectra shown to the right for objects A-D, to answer the next **two** questions. **Note that one of the spectra is from an object at rest (not moving) and the remaining spectra come from objects that are all moving toward the observer.** *Assume that the left end of each spectrum corresponds to shorter wavelengths (blue light) and that the right end of each spectrum corresponds with longer wavelengths (red light).*

18. Which of the four objects A-D is at rest?
    a. Object A.
    b. Object B.
    c. Object C.
    d. Object D.

19. *Of the three objects that **are** moving*, which is moving with the slowest speed?
    a. Object A.
    b. Object B.
    c. Object C.
    d. Object D.
    e. They are all moving the same speed, the speed of light.

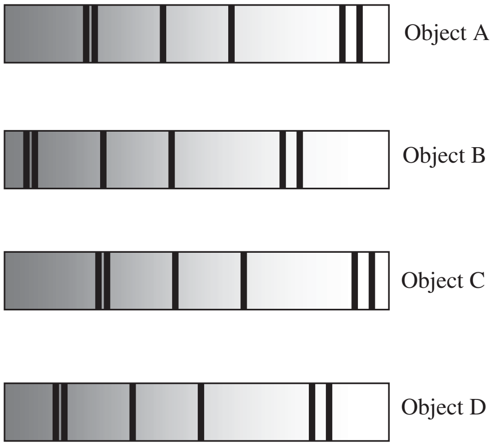

FIG. 2. Items 18 and 19 from the LSCI.

an emission line, respectively. Items 2 and 22 ask students to reason about whether they can infer information about the color and temperature of a star, respectively, given its absorption line spectrum. The high Q3 values for these pairs of items make sense given the overlapping nature of their content.

We also found that item 23 had high Q3 values with several items. Item 23 asks students to compare the energy, frequency, wavelength, and speed of radio waves and visible light. The specific reasons why item 23 exhibits local dependence with multiple items are not clear. However, to correctly answer item 23, students must synthesize their knowledge of how different types of light compare in terms of energy, wavelength, frequency, and the speed at which they travel through a vacuum. These ideas are so fundamental that students must frequently invoke them when reasoning about other items on the LSCI.

There are two possible solutions for how to deal with locally dependent item pairs. One solution is to drop one item from each offending pair from the data set. However, dropping items from the test reduces the amount of available information that can be used to estimate students' abilities. We therefore took an alternative approach and combined each high-Q3 pair into a single polytomous item. We tried several versions of the test with different pairwise combinations in an attempt to find a set of items that were all locally independent. After several trials, we were able to resolve the problem for all items except for item 23; regardless of the changes made to the rest of the test, this item was always found to be locally dependent on other items on the test. We were therefore forced to remove item 23 from the test.

We ended up with three polytomous items (items 7 and 8 combined, items 18 and 19 combined, and items 2 and 22 combined). These three items were calibrated using the two-parameter graded response model [24]. The graded response model can be written as

$$P(X_{pi} \geq j|\theta_p, a_i, b_{ij}) = \frac{\exp[a_i(\theta_p - b_{ij})]}{1 + \exp[a_i(\theta_p - b_{ij})]}, \quad (3)$$

where the student's ability θ_p and item's discrimination parameter a_i have the same meaning as in Eqs. (1) and (2). Unlike the 2PL and 3PL models, the graded response model does not assign each item a single number b_i to represent that item's difficulty. Instead, each polytomous item is assigned multiple threshold parameters b_{ij}. A given threshold parameter b_{ij} represents the ability a student must have in order to have a 50% probability of responding at or above the jth threshold for a given item i. For each of these polytomous items we created (items 7 and 8 combined, items 18 and 19 combined, and items 2 and 22 combined), a student can receive one of three possible scores: 0, 1, or 2. As an example, consider a student responding to items 7 and 8. That student will receive a score of 0 if he incorrectly responds to both items 7 and 8, a score of 1 if he correctly responds to one item but not the other, and a score of 2 if he correctly responds to both items. Therefore, b_{i1} represents the ability a student needs in order to have an equal probability of scoring a 0 or 1, and b_{i2} represents the ability a student needs in order to have an equal probability of scoring a 1 or 2. See Embretson and Reise for a pedagogical treatment of the graded response model [16].

Items 21, 23, and 25 were dropped from the instrument. We maintained all other items in their original form and calibrated them using the 3PL model. We will uses this twenty-item reduced version of the LSCI for all of the analyses subsequently described in this paper. Table I contains the matrix of Q3 values for every item pair on this reduced version of the LSCI. Table I shows that this version of the LSCI satisfies the assumption of local independence.

C. Unidimensionality

A test such as the LSCI is considered to be unidimensional if a single latent trait (aka ability θ_p) can fully explain a student's performance on the test given the parameters describing the items on that test (e.g., a_i, b_i, and c_i). In other words, a test is unidimensional if it measures students' abilities on a single construct. Local independence is a necessary but not sufficient condition for unidimensionality, so we conducted two additional tests to determine whether or not the assumption of unidimensionality holds.

For the first test, we fit the data with a two-latent-trait model and compared the results to those we obtained from the single-latent-trait model. The two-dimensional model did not yield a set of goodness-of-fit statistics that were better overall than those obtained by the unidimensional model. Specifically, neither the average of the items' χ^2 values nor the RMSEA were smaller for the two-dimensional model compared to the unidimensional model. This suggests that a single latent trait is adequate to explain students' response patterns to the reduced version of the LSCI.

We then performed Bejar's test for unidimensionality [25]. Bejar reasons as follows: Imagine that a researcher suspects a test contains subsets of items that each probe their own unique construct. The researcher could estimate item difficulties b_i using the data for every item on the test. The researcher could also estimate the item difficulties for the items on each subtest by using the data on those subtest items only. If the test is truly unidimensional, then a plot of the subtest-based item difficulty estimates versus the whole-test-based item difficulty estimates should show a series of points that fall near a line of slope one and intercept zero. This is because the probability of correctly answering an item should not depend on which items are included on the test if the test is unidimensional. Significant departures from this line are thus considered evidence that unidimensionality is violated.

For Bejar's test, we place items into three mutually exclusive groups, which represented our hypothesis about which items might possibly form subtests that probe different constructs. One group included items that probe students' understandings of Wien's law and the luminosity-area-temperature relationship (items 3, 6, 9, 12, 16, 20, 24, and 26), another included items that probe students'

TABLE I. Yen's Q3 statistic for each pair of items. Two items are considered locally independent if $|Q3| > 0.20$.

	Item 1	Item 3	Item 4	Item 5	Item 6	Item 7 and 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18 and 19	Item 20	Item 2 and 22	Item 24	Item 26
Item 1	1.00	−0.04	−0.12	−0.07	−0.10	−0.07	−0.08	−0.03	−0.03	−0.15	−0.10	−0.08	0.19	−0.07	−0.04	−0.09	−0.05	−0.05	−0.09	−0.06
Item 3		1.00	−0.02	−0.01	0.00	−0.02	0.02	0.07	−0.03	−0.02	0.00	−0.03	−0.01	−0.01	−0.02	0.01	0.04	−0.07	−0.02	0.03
Item 4			1.00	−0.05	−0.05	−0.10	−0.09	−0.01	−0.03	−0.05	−0.03	−0.06	−0.09	−0.01	0.02	−0.03	−0.02	−0.03	−0.06	−0.05
Item 5				1.00	−0.03	−0.04	−0.03	−0.04	−0.02	0.03	−0.02	0.07	−0.10	0.00	−0.04	−0.03	0.01	−0.09	−0.01	0.01
Item 6					1.00	−0.08	−0.07	−0.05	−0.04	−0.04	−0.07	−0.01	−0.10	0.06	−0.03	−0.05	−0.01	−0.04	−0.06	0.04
Item 7 and 8						1.00	−0.04	0.00	−0.02	−0.07	0.12	−0.09	−0.15	−0.08	−0.04	−0.05	−0.02	−0.03	−0.07	−0.01
Item 9							1.00	−0.04	−0.06	0.16	−0.10	−0.04	−0.15	−0.09	−0.10	−0.05	−0.05	−0.12	0.14	−0.05
Item 10								1.00	−0.05	−0.05	−0.02	0.00	−0.01	−0.05	−0.01	−0.02	0.08	−0.05	−0.05	−0.02
Item 11									1.00	−0.06	−0.06	−0.05	−0.06	−0.05	−0.01	−0.01	−0.04	0.03	−0.04	−0.06
Item 12										1.00	−0.07	0.02	−0.12	−0.04	−0.07	−0.03	−0.03	−0.18	0.16	−0.05
Item 13											1.00	−0.01	−0.04	−0.02	0.03	−0.03	0.01	−0.12	−0.08	−0.03
Item 14												1.00	0.05	0.00	0.00	−0.05	−0.03	−0.13	−0.05	−0.06
Item 15													1.00	−0.01	−0.01	−0.07	0.01	−0.14	−0.18	−0.03
Item 16														1.00	−0.02	0.01	−0.04	−0.08	−0.07	0.01
Item 17															1.00	0.02	−0.01	0.09	−0.08	−0.07
Item 18 and 19																1.00	0.02	−0.04	−0.05	−0.02
Item 20																	1.00	−0.10	−0.05	−0.03
Item 2 and 22																		1.00	−0.07	0.00
Item 24																			1.00	−0.03
Item 26																				1.00

understandings of spectroscopy (items 4, 7 and 8, 11, 17, 18 and 19, and 2 and 22), and the third included items that probed students' understandings of the properties of light (items 1, 5, 10, 13, 14, and 15). In all cases, the difficulty of each item fell within two standard errors of the target line of slope one and intercept zero. We conclude that the results of Bejar's test are consistent with the assumption of unidimensionality.

III. RESULTS

A. Item parameters and model fit

Table II shows the 3PL-estimated discriminations, difficulties, and guessing parameters of the dichotomous items on the reduced LSCI. The standard errors of these parameters are also shown. In order to assess how well the 3PL model fit the data for each item, we grouped students into ability bins 0.1 logits wide, except in a few cases where we had to increase the bin width in order to ensure there were at least five correct responses per bin. A minimum of five correct responses per bin is generally considered sufficient to accurately estimate the average ability of a bin [26]. Some bins became extremely wide when we attempted to meet this criteria, so we occasionally kept bin width at 0.1 logits and ignored all bins that failed to have at least five correct responses. This is why some items have fewer degrees of freedom than others. We compared each observed score to the expected score predicted by the 3PL model and calculated a χ^2 statistic for each item. Table II also reports the χ^2 values, the degrees of freedom, and the reduced χ^2 values (χ_r^2) for each item.

Table III contains the item parameters, their standard errors, the χ^2 values, the degrees of freedom, and the reduced χ^2 values for the three polytomous items. We calculated the χ^2 values using the same procedure described above, except we found the expected score of each bin by taking the weighted average of the probability of receiving a score of 1 and a score of 2 [i.e., $P(X_{pi} = 1) + 2P(X_{pi} = 2)$].

With a few exceptions, the χ_r^2 values are close to unity, suggesting the IRT models adequately fit the data. As an additional check on model fit, we plotted the model-predicted score on each item as a function of ability θ_p; these plots are reproduced in the Supplemental Material [27]. In each plot, the black curve [called the item characteristic curve (ICC)] represents the model-predicted score while the red triangles represent the average scores of students in each bin. Note that some of the ICCs diverge

TABLE II. The discrimination (a_i), difficulty (b_i), and guessing parameters (c_i) of the seventeen dichotomous items from the reduced LSCI, along with their standard errors (SE). The χ^2, degrees of freedom (d.o.f.), and reduced χ^2 (χ_r^2) values are also shown.

Item	a_i	a_i's SE	b_i	b_i's SE	c_i	c_i's SE	χ^2	d.o.f.	χ_r^2
1	1.30	0.07	−0.75	0.05	0.00	0.00	31.66	15	2.11
3	1.71	0.33	2.09	0.14	0.20	0.01	16.40	18	0.91
4	1.81	0.22	−0.56	0.17	0.23	0.09	27.65	17	1.63
5	1.57	0.23	−0.28	0.21	0.44	0.07	11.40	18	0.63
6	1.83	0.21	0.19	0.10	0.24	0.04	20.05	18	1.11
9	2.22	0.24	0.70	0.05	0.16	0.02	23.42	17	1.38
10	1.33	0.21	1.22	0.10	0.23	0.03	21.61	19	1.14
11	1.81	0.22	0.89	0.07	0.21	0.02	26.95	19	1.42
12	2.68	0.36	0.18	0.08	0.36	0.03	18.99	17	1.12
13	1.35	0.17	0.32	0.13	0.18	0.05	22.29	19	1.17
14	1.34	0.23	−0.73	0.39	0.40	0.14	16.83	18	0.94
15	1.42	0.07	−0.32	0.04	0.00	0.00	37.48	14	2.68
16	1.36	0.24	0.35	0.18	0.36	0.06	11.44	18	0.64
17	1.26	0.22	1.88	0.13	0.15	0.02	25.94	18	1.44
20	1.14	0.23	0.99	0.15	0.31	0.05	17.76	19	0.93
24	2.57	0.33	0.53	0.06	0.31	0.02	15.87	18	0.88
26	2.33	0.34	1.39	0.07	0.23	0.01	30.07	19	1.58

TABLE III. The discrimination (a_i) and thresholds (b_{i1} and b_{i2}) of the three polytomous items from the reduced LSCI, along with their standard errors (SE). The χ^2, d.o.f., and reduced χ^2 (χ_r^2) values are also shown.

Item	a_i	a_i's SE	b_{i1}	b_{i1}'s SE	b_{i2}	b_{i2}'s SE	χ^2	d.o.f.	χ_r^2
7 and 8	1.06	0.05	−0.38	0.05	0.33	0.05	36.65	36	1.02
18 and 19	0.74	0.05	0.01	0.05	1.31	0.09	40.98	37	1.11
2 and 22	1.52	0.07	0.05	0.03	0.84	0.05	43.59	29	1.50

from the data points at the low end of the ability spectrum. This suggests that the abilities of low-ability students may be overestimated, which means the learning gains reported in Secs. III C and III D may actually slightly underestimate the gains achieved by some students. Further elaboration on this point is beyond the scope of this paper, although it may be a worthwhile topic for a future investigation. However, the overall close fit between the observed and predicted response patterns in many cases provides further evidence that the IRT models are appropriate for modeling student ability.

B. Item interpretations

Before discussing the estimated student abilities and the learning gains achieved by different classes in our data set, we must comment on what the item parameters in Tables II and III tell us about the LSCI as an instrument. First, consider the fact that the five items with the largest values of the discrimination parameter a_i (items 12, 24, 26, 9, and 6, ranked from largest to smallest a_i) all come from the group of items that probe students' abilities to reason about and apply Wien's law and the luminosity-area-temperature relationship. Furthermore, all five of these items require students to interpret a graph, such as star properties plotted on a graph of luminosity versus temperature. The remaining items from the Wien and/or luminosity-area-temperature group (items 3, 16, and 20) are entirely word-based questions and have lower discrimination values. This demonstrates that graph-based items assessing Astro 101 students' understandings of Wien's law and/or the luminosity-area-temperature relationship are especially effective at discriminating between high- and low-ability students.

The plots in the Supplemental Material [27] show that a student must have an ability greater than 0 logits in order to have at least a 50% probability of correctly answering any of the Wien and/or luminosity-area-temperature items, with the exception of items 6 and 12. This is significant because the average postinstruction ability of students in the data set was set at 0 logits. That means 50% or more of students have less than a 50% chance of giving the correct answer to six of the eight Wien and/or luminosity-area-temperature items even at the end of their Astro 101 course.

Overall, the Wien and/or luminosity-area-temperature items appear to be challenging for most Astro 101 students. However, they are not so difficult that success on these items is unattainable, which is why they tend to have high values of a_i, indicating that they are effective at discriminating between students of different abilities. We suspect that many of these items might have had higher discrimination values if not for the fact that they also have nonzero guessing parameters. Items 3, 6, 9, and 26 have guessing parameters that are around 0.20 to 0.25, which is consistent with low-ability students randomly guessing the correct answers when there are four to five available choices. Items 20 and 24 have guessing parameters $c_{20} = c_{24} = 0.31$ and items 12 and 16 have $c_{12} = c_{16} = 0.36$. These values of c_i suggest that, after instruction, many low-ability students can eliminate at least one of the distractors before making a guess. For example, only 10% of students selected choice a for item 20 (Fig. 3), while 52% selected b, 24% selected c, and 14% selected d. We conclude that the discriminatory powers of the Wien and/or luminosity-area-temperature items are attenuated because low-ability students have a nonzero probability of correctly guessing the correct answers. This result is consistent with the findings of Wooten *et al.*, which suggest that student performance on multiple-choice questions in many cases overestimates student understanding of a topic [28].

20. The coolest stars emit most of their energy in which portion of the electromagnetic spectrum?

    a. X-ray.
    b. Infrared.
    c. Visible.
    d. Ultraviolet.

FIG. 3. Item 20 from the LSCI.

In contrast to the Wien and/or luminosity-area-temperature items, items probing students' understandings of the properties of light (items 1, 5, 10, 13, 14, and 15) tend to have both lower discrimination values and lower difficulty parameters. If we separate the nine dichotomous items with the largest values of a_i from the nine dichotomous items with the smallest values of a_i, then we find that all of the properties of light items fall in the latter category. Furthermore, many students with below average postinstruction abilities ($\theta_p < 0$ logits) still have a greater than 50% chance of correctly answering items 1, 5, 14, and 15. Items 5 and 14 have extremely high guessing parameters (0.44 and 0.40, respectively). These questions ask students to select a photon (item 5) or an electromagnetic wave (item 14) with the largest energy. We suspect that many students, even those of low ability, can eliminate one or more of the distractors based on what they learned in their Astro 101 classes about the relationships between the energy, wavelength, frequency, and color of light. Items 1 and 15 are interesting because they both have guessing parameters of 0. Both of these questions address the common incorrect idea that more energetic forms of light travel faster. The fact that these items have nonexistent guessing parameters while simultaneously having low difficulties suggest that while many students can readily learn the fact that all forms of light travel at the same speed in a vacuum, low-ability students who never commit this fact to memory are almost certainly going to choose one of the distractors. This implies that the distractors on these items are highly effective. Overall, we are not surprised by the low difficulties and discriminatory capabilities of these items given that they tend to probe what is simply declarative knowledge for many Astro 101 students.

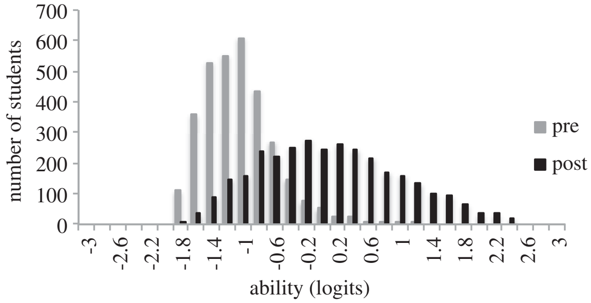

FIG. 4. The distribution of pre- and postinstruction abilities for all 3205 students in the data set.

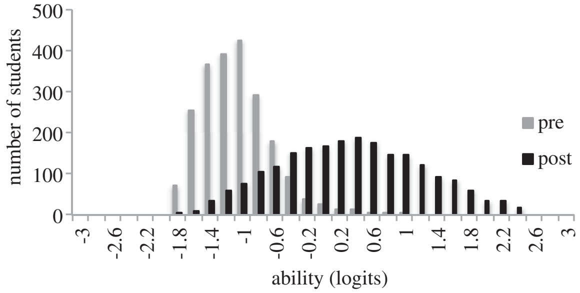

FIG. 6. The distribution of pre- and postinstruction abilities for the 2178 students in high-IAS classes with at least 25 students.

C. Estimated student abilities

Figure 4 shows the distribution of estimated pre- and postinstruction abilities for all 3205 students in the data set. Preinstruction abilities range from -1.9 to 1.0 logits, with an average of -1.1 logits and a standard deviation of 0.45 logits. The postinstruction abilities span a wider range, from -1.9 to 2.4 logits, with an average of 0 logits and a standard deviation of 0.89 logits. Recall that the average postinstruction ability is set at 0 logits, as described in Sec. II.

We calculated the difference between post- and preinstruction abilities ($\Delta\theta$) for each student. This difference represents an IRT-estimated learning gain [13]. Figure 5 shows the distribution of these IRT learning gains for all 3205 students. The minimum "gain" was -1.8 logits, the maximum was 4 logits, and the average was 1.1 logits with a standard deviation of 0.93 logits. These data show a range in the shift of abilities, with less than 10% of the assessed population exhibiting a shift ≤ 0 logits, which would be consistent with students moving backward or achieving no improvement in their understanding.

Since this study was carried out in order to examine the effects of active learning on individual students, we want to look at changes in abilities for students who took Astro 101 classes with high and low levels of interactive instruction. The preceding study of Prather *et al.* [3] looked at the level of interactivity of classes in this data set with at least 25 students. They estimated each class's level of interactivity based on instructors' responses to the interactivity assessment instrument [3]. These responses allowed Prather *et al.* to calculate an interactivity assessment score (IAS) for each class. IAS scores ranged from 0% to 49% and represent an estimate of the percentage of class time during which active learning techniques are used. Prather *et al.* found that IAS scores of at least 25% are necessary, but not sufficient, to produce classes with average normalized learning gains above $\langle g \rangle = 0.30$. An average normalized gain of $\langle g \rangle = 0.30$ is significant because Hake [29] found that only interactive physics classes—and not traditionally taught classes—were able to achieve this level of improvement in student performance on the Force Concept Inventory (FCI). Consequently, Hake defined $\langle g \rangle = 0.30$ as the cutoff between "low" and "medium" levels of gain.

In this study, we again look at at students enrolled in Astro 101 classes with at least 25 students. Like the prior study by Prather *et al.*, we divide these classes into two groups: high-IAS (i.e., IAS $\geq$ 25%) and low-IAS (IAS < 25%). Figures 6 and 7 show the pre- and postinstruction ability distributions for students in high- and low-IAS classes, respectively. Students in high-IAS classes have preinstruction abilities that range from -1.9 logits to 0.92 logits with an average of -1.2 logits and a standard deviation of 0.42 logits. Their postinstruction abilities range from -1.9 logits to 2.4 logits with an average of 0.23 logits and a standard deviation of 0.88 logits. In contrast, students in low-IAS classes have pre-instruction abilities that range from -1.9 logits to 0.91 logits with an

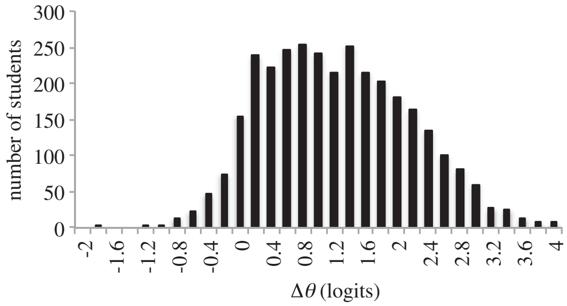

FIG. 5. The distribution of IRT-estimated gains ($\Delta\theta$) for all 3205 students in the data set.

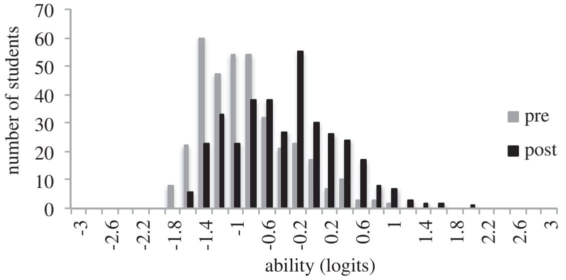

FIG. 7. The distribution of pre- and postinstruction abilities for the 363 students in low-IAS classes with at least 25 students.

average of −0.95 logits and a standard deviation of 0.56 logits. Their postinstruction abilities range from −1.8 logits to 1.9 logits with an average of −0.45 logits and a standard deviation of 0.69 logits. While the distributions of pre- and postinstruction abilities for students in high- and low-IAS classes cover approximately the same range, the postinstruction averages are noticeably different. A one-tailed t test for two independent samples revealed that the difference in the postinstruction means was statistically significant ($p < 0.0001$) and of large effect size (Cohen's $d = 0.80$) [30]. Interestingly, and unexpectedly, a one-tailed t test also revealed that the difference in the preinstruction means was also statistically significant ($p < 0.0001$) and of medium effect size (Cohen's $d = 0.49$). These results show that even though the population of students from high-IAS classes began with a smaller average preinstruction ability, they had a higher average postinstruction ability than their peers in low-IAS classes. A significant number of students in high-IAS classes moved into regions of ability space that were unoccupied in the preinstruction distribution, and they did so at a greater percentage than students in low-IAS classes. This means a significant number of students in high-IAS classes, compared to students in low-IAS classes, acquired astronomical reasoning abilities and knowledge that were not held by most students prior to instruction. This important result is consistent with the CTT findings of Prather *et al.* and with decades of research highlighting the pedagogical effectiveness of interactive instruction [31].

The disparity in student achievement between high- and low-IAS classes is also seen when we examine the distributions of IRT-estimated learning gains ($\Delta\theta$) (see Fig. 8). Students in high-IAS classes have values of $\Delta\theta$ that range from −1.2 logits to 4 logits with an average of 1.4 logits and a standard deviation of 0.90 logits. Students in low-IAS classes have values of $\Delta\theta$ that range from −0.95 logits to 2.6 logits with an average of 0.49 logits and a standard deviation of 0.66 logits. Once again, a one-tailed t test revealed the difference in these averages to be statistically significant ($p < 0.0001$). This difference in averages also corresponds to a very large effect size (Cohen's $d = 1.2$), according to the effect size classification scheme proposed by Sawilowsky [32]. Surprisingly, students in high-IAS classes averaged a pre-post improvement in their abilities that was almost an entire logit greater than the average pre-post ability improvement of students in low-IAS classes. To get a sense of the meaning of a difference of 1 logit, consider the difficulty parameters of the seventeen dichotomous items on the LSCI (Table II). These difficulty parameters range from −0.75 logits to 2.09 logits. A student whose ability increases by 1 logit will have a significantly higher probability of correctly answering many of the LSCI's items. This same reasoning also applies to the three polytomous items (Table III). For example, a student with an ability of 0 logits has a 55% chance of answering item 6 correctly, but a student with an ability 1 logit greater has an 85% probability of answering this item correctly.

Many astronomy and physics education researchers frequently use Hake's average normalized gain $\langle g \rangle$ to make inferences about the amount of learning experienced by populations of students [14,29,33,34], including the earlier study by Prather *et al.* [3]. In addition to calculating $\Delta\theta$ for all 3205 students in the data set, we also calculated their normalized gains g. Figure 9 shows a graph of $\Delta\theta$ versus g for all 3205 students. There is a definite correlation between the two measures ($r = 0.93$). But note that each value of g corresponds to a range of values of $\Delta\theta$ approximately 1 logit or more wide. There are many IRT-estimated gains associated with a single value of g. Two students who have values of $\Delta\theta$ separated by 1 logit have experienced significantly different improvements in their underlying abilities, even if they possess the same normalized gain. This result makes sense when one recalls that IRT estimates a person's ability based on the relative difficulty of the questions she correctly answered, not just the total number of correct answers. This result suggests that while average normalized gains may be good at summarizing the performance of a population of students, g may not be as informative an indicator of the learning gains of individual students.

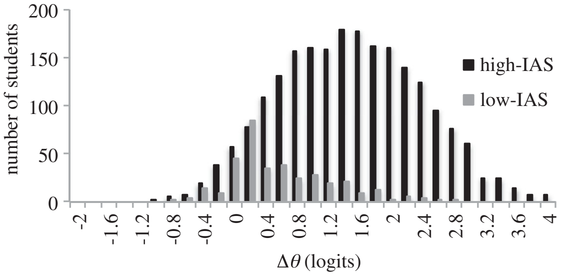

FIG. 8. The distribution of IRT-estimated gains ($\Delta\theta$) for the 2178 students in high-IAS classes and the 363 students in low-IAS classes.

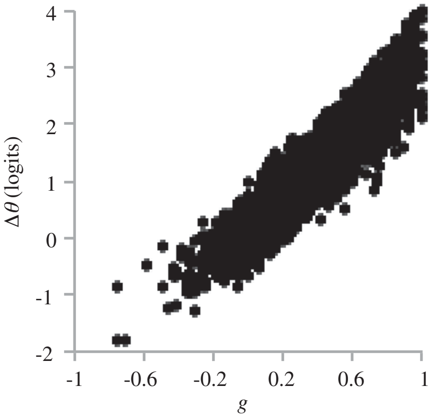

FIG. 9. $\Delta\theta$ versus g for all 3205 students in the LSCI national data set.

Of course, Fig. 9 also shows that there are also multiple values of $\Delta\theta$ for each value of g. Why then do we claim that $\Delta\theta$ more robustly models changes in student understanding than g? Recall that in order to obtain these values of $\Delta\theta$ we had to perform numerous statistical tests to demonstrate that the IRT models we used fit the data and satisfied the underlying assumptions of local independence and unidimensionality. When these conditions are satisfied, IRT models provide estimates of students' abilities that are independent of the specific items they answered [7]. If one wants to argue that g is a more accurate measure than $\Delta\theta$ of learning gains, then one must justify why a raw test score is a better measure than θ of the latent trait of student ability, despite of the amount of statistical rigor required to produce θ values. We believe that the statistical analysis underlying $\Delta\theta$ makes it highly unlikely that g is a superior measure of learning gain.

D. Effectiveness of different Astro 101 classes

Earlier investigations of the LSCI national data set examined the relationships between the average normalized learning gains of classes, the amount of time devoted to active learning, and the quality of an instructor's implementation of those strategies [3,4]. Consequently, we are interested in examining the average IRT-estimated learning gains for classes in this data set, as well as the pre- and postinstruction ability histograms for individual classes, in order to determine how well each class did with regards to evolving students' underlying abilities over the course of the semester.

For each class with at least 25 students in the data set, Table IV shows the type of institution at which the class was taught, the number of enrolled students, the average pre- and postinstruction scores on the LSCI, $\langle g\rangle$, the average pre- and postinstruction abilities, average $\Delta\theta$, and the instructor's IAS. The classes are ordered from largest to smallest average $\Delta\theta$. Figure 10 plots these average $\Delta\theta$ values versus $\langle g\rangle$. There is a large correlation between these two measures ($r = 0.99$), which supports the robustness of the results reported in Prather *et al.* [3], Rudolph *et al.* [4], and Schlingman *et al.* [5]. Figure 11 shows average $\Delta\theta$ versus IAS. As expected, this reveals that spending more time on active learning strategies is

TABLE IV. The institution type, number of enrolled students, average pre- and postinstruction LSCI scores, average normalized gain $\langle g\rangle$, average pre- and postinstruction abilities θ, average IRT learning gain $\Delta\theta$, and IAS for each class in the national data set with at least 25 students.

Class	Institution type	Students	Average prescore	Average postscore	$\langle g\rangle$	Average pre-θ	Average post-θ	Average $\Delta\theta$	IAS
1	Research University	96	5.79	15.26	0.47	−1.18	0.66	1.84	45.9
2	Research University	93	6.06	14.92	0.44	−1.16	0.68	1.84	45.9
3	Research University	63	6.35	14.68	0.42	−1.04	0.57	1.61	45.9
4	4-yr Masters and Bach. Univ.	33	6.94	15.00	0.42	−0.96	0.60	1.56	30.8
5	Research University	84	5.85	13.99	0.40	−1.15	0.48	1.63	45.9
6	Research University	65	5.17	13.58	0.40	−1.19	0.41	1.60	45.9
7	Research University	444	5.41	13.61	0.40	−1.21	0.37	1.58	45.9
8	4-yr Masters and Bach. Univ.	43	6.79	14.37	0.39	−0.98	0.40	1.38	26.1
9	Research University	344	5.21	13.20	0.38	−1.24	0.33	1.57	45.9
10	Research University	61	5.33	13.02	0.37	−1.26	0.27	1.53	45.9
11	Research University	402	5.94	12.27	0.32	−1.12	0.14	1.25	45.9
12	4-yr Masters and Bach. Univ.	36	6.53	12.50	0.31	−1.01	0.13	1.14	30.8
13	4-yr Masters and Bach. Univ.	28	7.89	13.14	0.29	−0.83	0.22	1.05	30.8
14	2-yr College	66	5.58	11.41	0.29	−1.16	−0.02	1.14	48.6
15	Research University	64	6.16	11.45	0.27	−1.04	−0.07	0.97	3.6
16	4-yr Masters and Bach. Univ.	40	5.65	10.43	0.23	−1.19	−0.19	0.99	34.3
17	4-yr Masters and Bach. Univ.	40	6.55	11.05	0.23	−1.06	−0.22	0.84	47.8
18	4-yr Masters and Bach. Univ.	65	6.23	10.38	0.21	−1.08	−0.22	0.86	21.6
19	Research University	47	6.62	10.62	0.21	−1.04	−0.18	0.86	34.5
20	4-yr Masters and Bach. Univ.	41	5.49	9.59	0.20	−1.25	−0.41	0.84	34.3
21	4-yr Bachelors College	33	5.24	8.91	0.18	−1.24	−0.48	0.76	36.9
22	Research University	28	5.93	9.36	0.17	−0.94	−0.36	0.58	2.1
23	4-yr Masters and Bach. Univ.	42	5.17	8.45	0.16	−1.27	−0.54	0.73	34.3
24	4-yr Masters and Bach. Univ.	77	5.21	7.88	0.13	−1.26	−0.59	0.67	47.3
25	2-yr College	27	5.37	7.41	0.10	−1.16	−0.83	0.33	22.1
26	4-yr Masters and Bach. Univ.	62	5.85	7.61	0.09	−1.13	−0.76	0.37	19.3
27	2-yr College	25	6.00	7.72	0.09	−1.19	−0.86	0.33	22.1
28	4-yr Bachelors College	27	5.26	6.74	0.07	−1.21	−1.05	0.16	9.7
29	4-yr Masters and Bach. Univ.	65	10.12	10.14	0.00	−0.28	−0.27	0.01	5.4

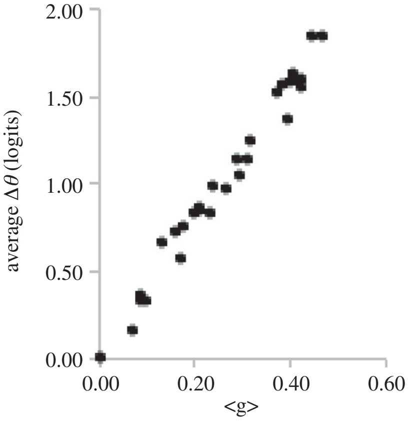

FIG. 10. Average $\Delta\theta$ versus $\langle g\rangle$ for classes with at least 25 students in the LSCI national data set.

necessary to move beyond small learning gains, supporting the validity of the findings of prior CTT studies, which often report measures of learning gain that are two times larger for students taught interactively than students taught traditionally [3,29]. A similar plot of $\langle g\rangle$ versus IAS in Prather *et al.* [3] revealed that only instructors with an IAS greater than or equal to 25% had classes with at least a medium gain ($\langle g\rangle = 0.3$ according to Hake [29]). Comparing Fig. 11 with the $\langle g\rangle$ versus IAS plot from Prather *et al.* [3] suggests that an average $\Delta\theta = 1$ is approximately equivalent to $\langle g\rangle = 0.3$. Figure 11 also suggests that an IAS of 25% is a necessary, though not sufficient, condition to achieve average $\Delta\theta > 1$. This result strongly suggests that simply making a class more interactive is not enough to maximize student learning; the quality of an instructor's ability to create an effective active learning classroom plays a significant role in student learning outcomes.

We created pre- and postinstruction ability histograms for each of the twenty-nine classes with at least 25 students. These histograms are located in the Appendix (Figs. 15–43). We now move to investigate these histograms of student abilities in order to gain deeper insights into the effectiveness of instruction in different classes. We will focus our investigation on the distributions from three classes that represent dramatically different outcomes.

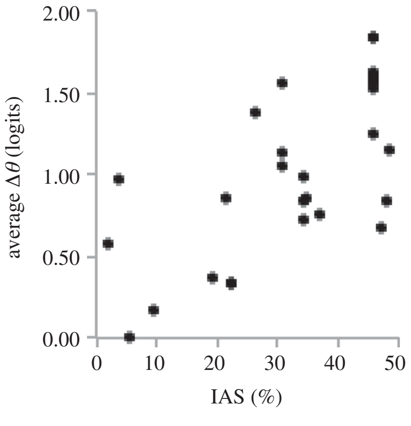

FIG. 11. Average $\Delta\theta$ versus IAS for classes with at least 25 students in the LSCI national data set.

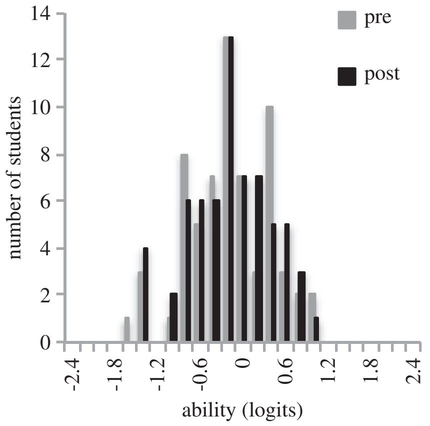

FIG. 12. The pre- and postability histogram for class 29.

Class 29 (Fig. 12) has the lowest average $\Delta\theta$ value (0.01) and the third lowest IAS in the data set. Note that Class 29 also has the highest average pre-instruction score on the LSCI and the highest average preinstruction ability. Despite the apparent advantages this class of students had at the beginning of their Astro 101 course, the histogram graphically illustrates that very few students improved in ability since the pre- and postinstruction distributions almost completely overlap one another.

In contrast, the histogram for class 19 (Fig. 13) shows that the preinstruction and postinstruction distributions of student abilities have far less overlap than what we observe for class 29. It is important to note that there are a considerable number of students with postinstruction abilities that none of the students had prior to instruction. This is powerful and illustrative evidence for the assertion that significant learning did occur in class 19. However, Class 19 does not represent the upper limit of what we observed with respect to student learning. While there is a clear separation between the distributions of the pre- and

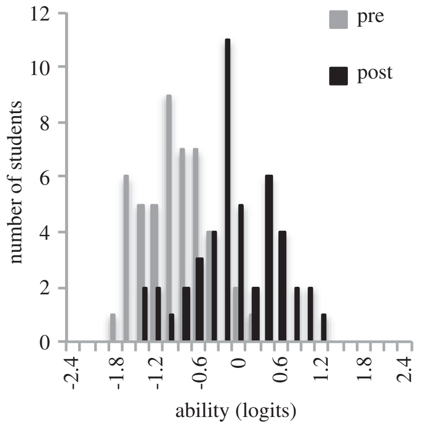

FIG. 13. The pre- and postability histograms for class 19.

postinstruction abilities, there is also a significant amount of overlap in these ability values. This suggests that there may be some students who did not experience any improvement in their abilities as a result of the instruction from class 19. Furthermore, the majority of postinstruction abilities have values less than 0 logits. This means that many students in class 19 still had postinstruction abilities that were below the study postinstruction average. Given the relatively low post instruction abilities of the students in class 19, it is informative to examine data from a class with very little overlap in the pre-post ability distributions, an impressive average $\Delta\theta$, and which has students who have achieved high postinstruction abilities.

The most impressive shift in student abilities was observed with class 1 (Fig. 14). There is astonishingly little overlap between the distributions of students' pre- and postinstruction abilities for these classes. A careful inspection of the pre- and postinstruction distributions for class 1 also reveals that after instruction almost every student has an ability that none of the students had prior to instruction—a truly remarkable teaching and learning accomplishment. Additionally, most students in class 1 have postinstruction abilities greater than 0 logits, meaning they were above the data set's postinstruction average. Some students in class 1 achieved postinstruction abilities of 2.2 logits, which is at the extreme high end of the distributions shown in Fig. 4—and this in a class with one of the lower average pre-instruction abilities. Overall, Class 1 serves as an example for how transformative a single semester introductory astronomy course can be with regards to improving students' conceptual and reasoning abilities on fundamental astrophysical ideas. During our presentations, after sharing the results from class 1 with faculty, most are quick to switch to aspiring for learning outcomes similar to class 1 over class 19.

Even though there is a large correlation between average $\Delta\theta$ values and $\langle g \rangle$, Fig. 10 and Table IV also show that it is possible for two classes to have the same value for $\langle g \rangle$ but very different average $\Delta\theta$ values, and vice versa. For example, classes 3 and 8 have similar values of $\langle g \rangle$ (0.42 and 0.39, respectively) but $\Delta\theta$ values that differ by 0.23 logits (1.61 logits versus 1.38 logits, respectively). We also suspect it is possible to have a class with a large average $\Delta\theta$ value but a histogram of pre- and postinstruction ability distributions that is unimpressive in important aspects (e.g., most students are still below average postinstruction average θ). Such findings as these reinforce the value of an IRT analysis for extracting information from larger educational data sets. The above considerations, plus our above analyses of class 1, 19, and 29, suggest that instructors seeking a full understanding of the effectiveness of their classroom instruction should compare a measure of their classes' average improvement (e.g., $\langle g \rangle$ and/or average $\Delta\theta$) with the distribution of students' pre- and postinstruction abilities, and the distribution of individual student gains $\Delta\theta$. By combining these multiple perspectives on individual and classwide abilities and gains, one can obtain a much more robust understanding of the effects of instruction. Even so, the outcomes of one class are much more meaningful when compared to the outcomes of other classes; using a widely validated and applied instrument such as the LSCI allows instructors to understand the efficacy of their teaching in both local and global contexts.

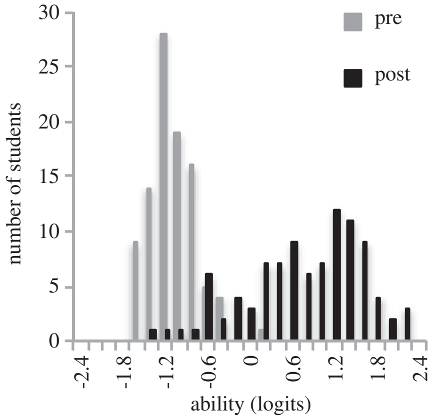

FIG. 14. The pre- and postability histograms for class 1.

IV. SUMMARY AND CONCLUSIONS

We used IRT to analyze the responses of 3205 Astro 101 students from sixty-nine classes (representing all types of colleges and universities) to the LSCI. As part of our analysis, we removed two items from the LSCI: Item 21, due to the fact that it is known to be a problematic item [5], and item 25, since it is so difficult that students' success on it shows only a weak correlation with their underlying abilities. In order to satisfy IRT's assumption of local independence, we removed a third item, item 23, and we combined three pairs of items (items 7 and 8, items 18 and 19, and items 2 and 22) into three polytomous items. After making these modifications, we were able to fit the 3PL model to the remaining seventeen dichotomous items and the graded response model to the three polytomous items, while simultaneously satisfying IRT's assumptions of local independence and unidimensionality. By satisfying these assumptions—and in contrast to classical test theory (CTT)—we achieved parameter invariance, which means our estimates of students' underlying abilities and the parameters of the items to which they responded do not depend upon one another [7].

Our IRT analysis provided new insights into the functioning of many of the LSCI's items. Since the 3PL model contains a "guessing parameter" (c_i), the probability of correctly answering an item with a large value of c_i (e.g., Item 3) is influenced by many low-ability students guessing the correct answer. Items with small values of c_i (e.g., Item 1) must possess particularly powerful distractors that limit the influence of guessing on the probability of students getting the right answer. This kind of analysis is not possible with CTT.

When we look at specific categories of items on the LSCI, we learned that items probing the properties of light are the easiest for students to correctly answer. In contrast, items that require students to reason using Wien's law and/or the luminosity-area-temperature relationship are among the most difficult and discriminating items of the LSCI, especially when these items require students to interpret graphical or pictorial representations.

The results of our IRT analysis also support the robustness of the research results from prior classical test theory analyses of this data set [3–5]. We split all classes with at least 25 students in the data set into two categories: classes in which the instructor used active learning strategies for 25% of class time or more and classes in which the instructor spent less than 25% of class time using active learning strategies. Students in classes that used active learning strategies for 25% of class time or more had higher average postinstruction abilities and larger average IRT-estimated learning gains (average $\Delta\theta$) than students in classes that spent less time on active learning strategies—despite the fact that the higher IAS classes actually began Astro 101 with lower preinstruction abilities. Students in high IAS classes had an average $\Delta\theta$ that was approximately 1 logit greater than their peers in low IAS classes. This difference of 1 logit represents a significant fraction of the range of the LSCI's items' difficulties and threshold parameters, demonstrating that students in high active learning classes have significantly higher probabilities of correctly answering the LSCI's items. This is further supported by the fact that the average $\Delta\theta$ for high IAS classes is more than twice as large as the average $\Delta\theta$ for low IAS classes. When we plot the average $\Delta\theta$ versus the average normalized gain $\langle g\rangle$ for all classes with at least 25 students, we find a high correlation ($r = 0.99$) between these two measures. A plot of $\Delta\theta$ versus the percentage of class time spent on active learning reproduces the equivalent plot from Prather *et al.* in which $\langle g\rangle$ was used as the ordinate variable [3]. We make the empirical inference from the data that instructors who want their classes to achieve an average improvement in abilities of $\Delta\theta = 1$ logit must spend at least 25% of class time on active learning strategies. We believe this result supports the idea that faculty who are adopting active learning methods need to do more then simply add a few Peer Instruction or Think-Pair-Share questions every now and then or have students work on problems together in class every couple of weeks. Instead, using proven active learning strategies needs to become a significant and regular part of their teaching and their formative assessment of learning. However, the wide range in $\Delta\theta$ for high-IAS classes suggests that just using these strategies often is not enough; one's ability to create an effective classroom environment that incorporates active learning strategies is critical.

Our results also imply that for faculty and STEM education researchers to gain a more complete understanding of the learning of individual students and the effectiveness of a particular class requires more than just a measure of a class's average learning gain, such as $\langle g\rangle$ or $\Delta\theta$. We plotted the IRT-estimated gain $\Delta\theta$ versus the normalized gain g for all 3205 students in the data set. Each value of g corresponds to a range of values of $\Delta\theta$. The size of this range is typically at least 1 logit, which, as noted earlier, represents a significant difference in the probability of giving a correct response to any particular item. This result suggests that while $\langle g\rangle$ may be good at summarizing the average improvement of an entire class, g may not adequately assess individual student learning.

In order to evaluate the effectiveness of different Astro 101 classes represented in the data set, we created histograms of the pre- and postinstruction ability distributions for each class with at least 25 students. Such histograms provide information that is not captured by a single number such as $\langle g\rangle$ or average $\Delta\theta$. An examination of a class's histogram can reveal to what extent the pre- and postinstruction distributions overlap one another; the smaller the amount of overlap, the greater the fraction of students in that class who actually experienced a change in their abilities. Additionally, the histograms reveal how many students are still below the average postinstruction ability, even after a semester of instruction. In principle, it is possible for a class to have a large average $\Delta\theta$ and still have a majority of its students with below average abilities postinstruction. Educators and researchers who are interested in evaluating the overall effectiveness of a class should look at all of these pieces of information in order to obtain a more complete understanding of the class.

Item response theory has the power to help researchers and instructors visualize and better understand whether their classes are achieving the kinds of transformative learning experiences they hope to provide for their students. By sharing the results of IRT analyses with faculty, we have seen them become inspired and empowered to engage in course transformation that they believe can substantially improve the learning experiences for their students. IRT analyses of student performance, such as the one described in this paper, may be able to play an important role in motivating instructors to adopt active learning methods that have been developed and are supported by research into astronomy and physics education.

ACKNOWLEDGMENTS

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.

APPENDIX: CLASS HISTOGRAMS

Below are the histograms of pre- and postinstruction ability distributions for all twenty-nine classes in the data set with at least 25 students. The classes are ordered from high to low average $\Delta\theta$.

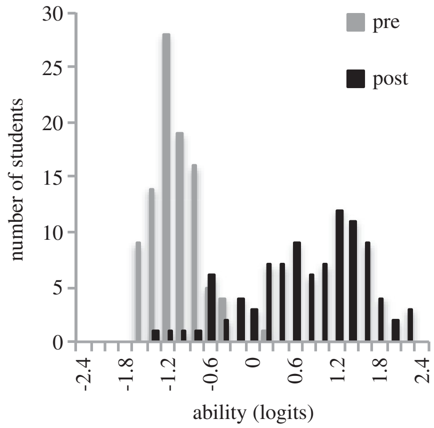

FIG. 15. The pre- and postability histogram for class 1.

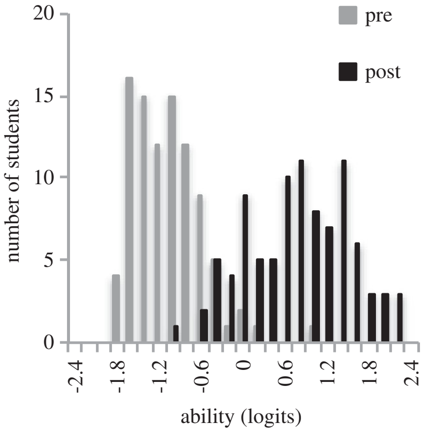

FIG. 16. The pre- and postability histogram for class 2.

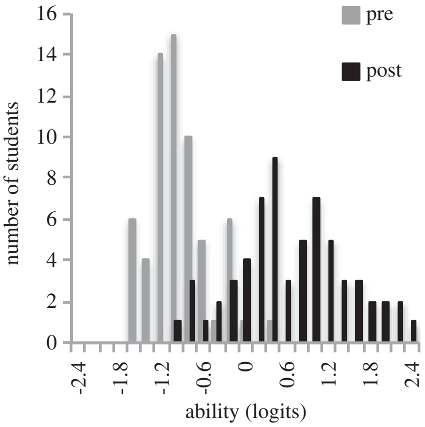

FIG. 17. The pre- and postability histogram for class 3.

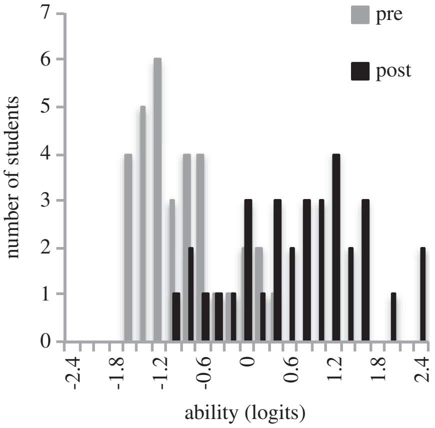

FIG. 18. The pre- and postability histogram for class 4.

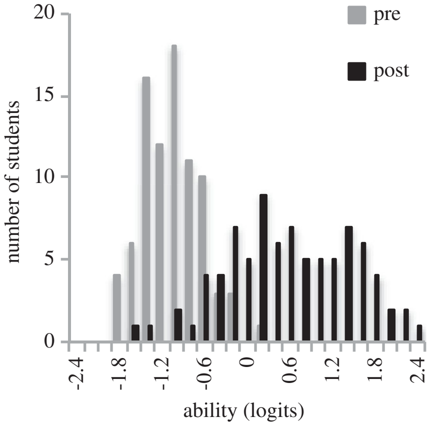

FIG. 19. The pre- and postability histogram for class 5.

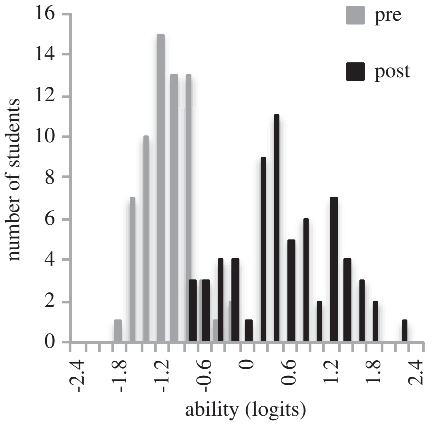

FIG. 20. The pre- and postability histogram for class 6.

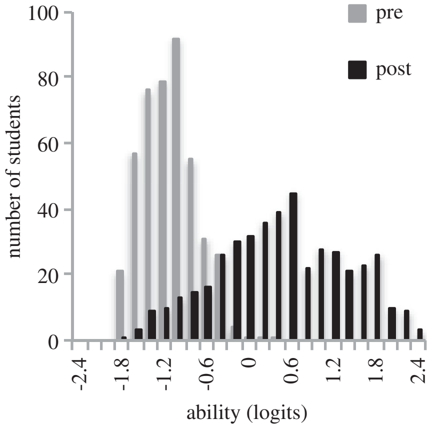

FIG. 21. The pre- and postability histogram for class 7.

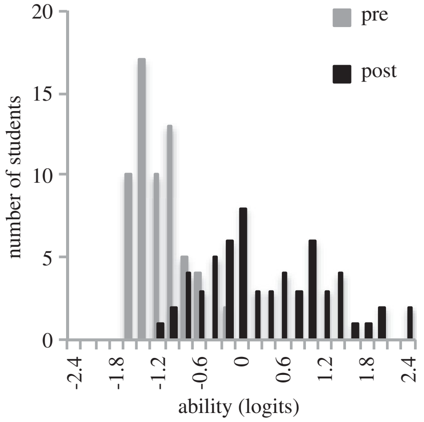

FIG. 24. The pre- and postability histograms for class 10.

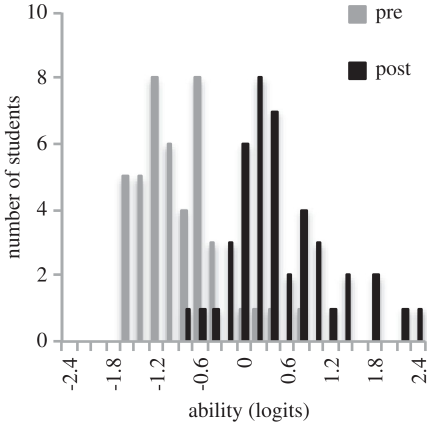

FIG. 22. The pre- and postability histogram for class 8.

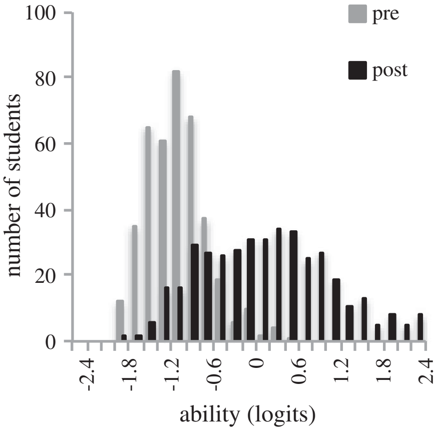

FIG. 25. The pre- and postability histograms for class 11.

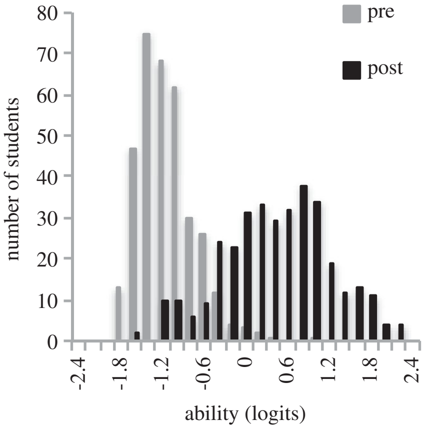

FIG. 23. The pre- and postability histogram for class 9.

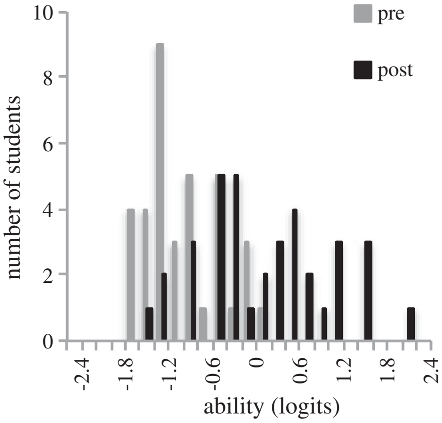

FIG. 26. The pre- and postability histograms for class 12.

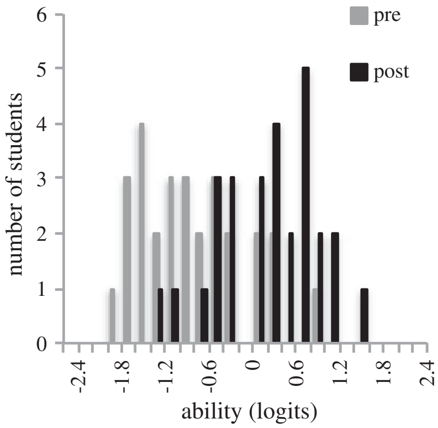

FIG. 27. The pre- and postability histograms for class 13.

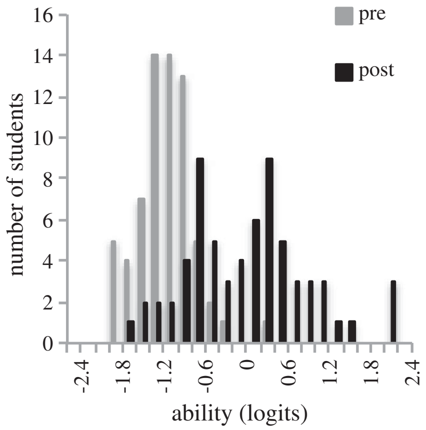

FIG. 28. The pre- and postability histograms for class 14.

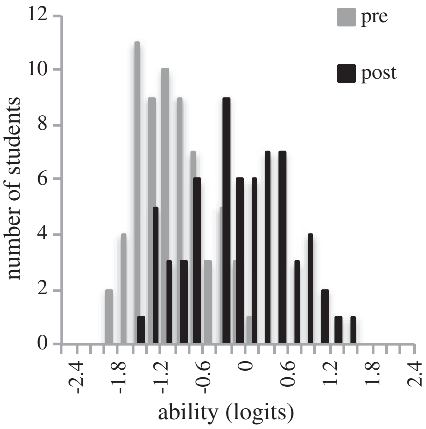

FIG. 29. The pre- and postability histograms for class 15.

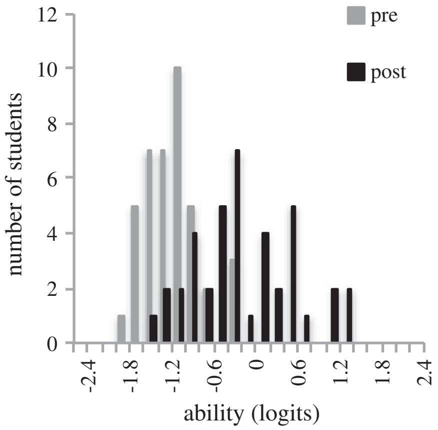

FIG. 30. The pre- and postability histograms for class 16.

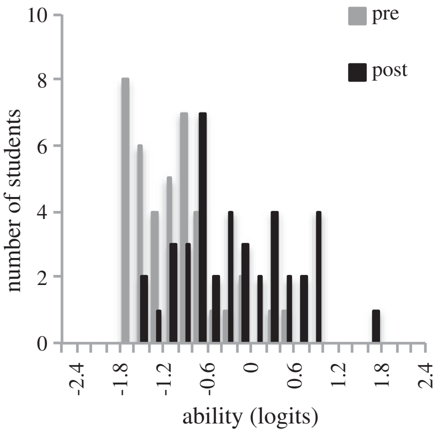

FIG. 31. The pre- and postability histograms for class 17.

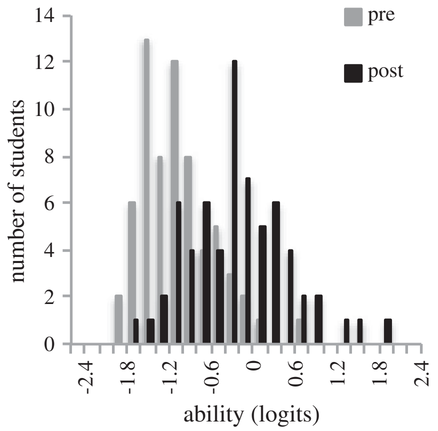

FIG. 32. The pre- and postability histograms for class 18.

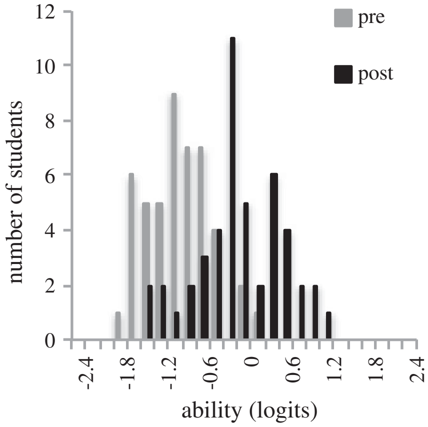

FIG. 33. The pre- and postability histograms for class 19.

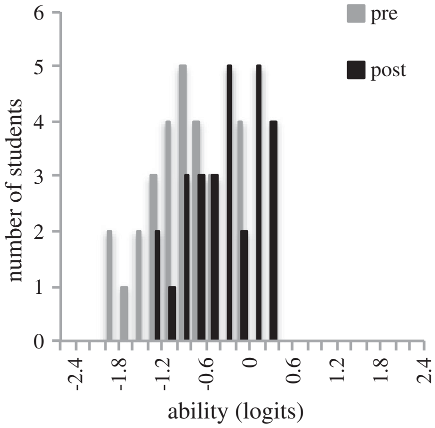

FIG. 36. The pre- and postability histogram for class 22.

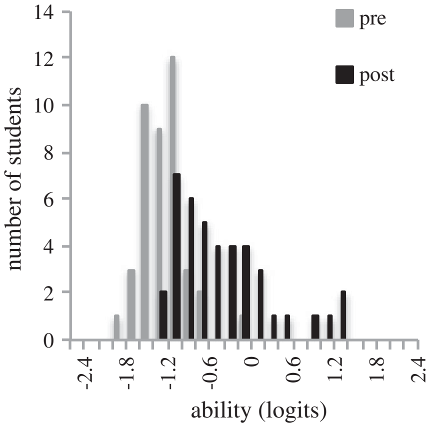

FIG. 34. The pre- and postability histogram for class 20.

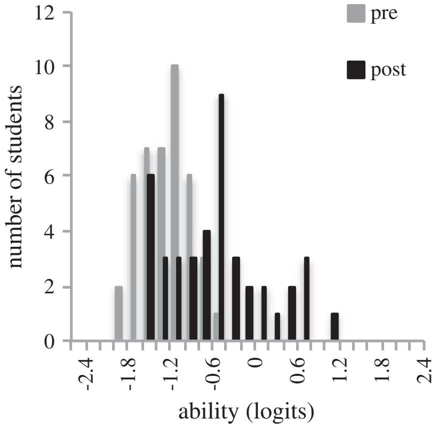

FIG. 37. The pre- and postability histogram for class 23.

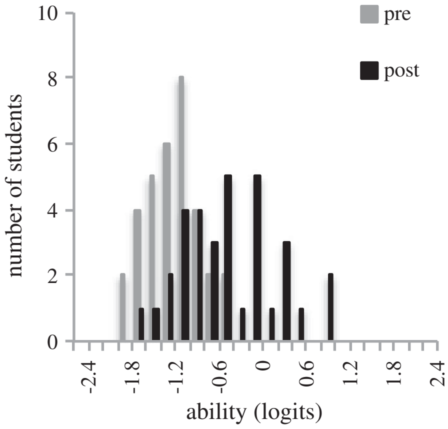

FIG. 35. The pre- and postability histogram for class 21.

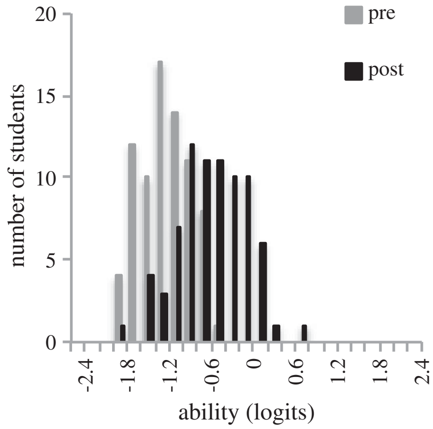

FIG. 38. The pre- and postability histogram for class 24.

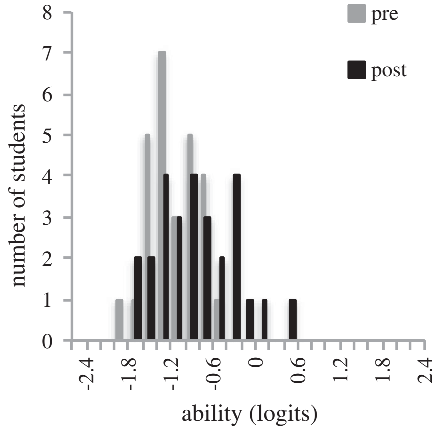

FIG. 39. The pre- and postability histogram for class 25.

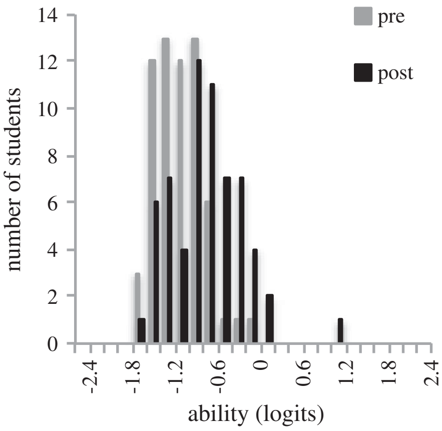

FIG. 40. The pre- and postability histogram for class 26.

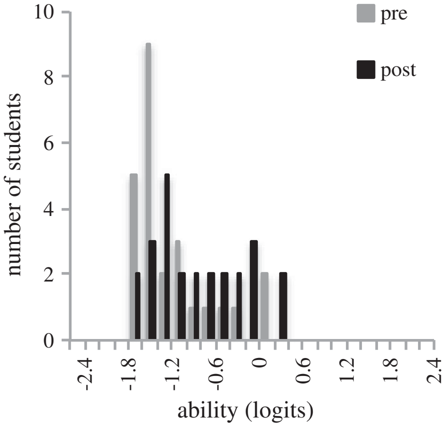

FIG. 41. The pre- and postability histogram for class 27.

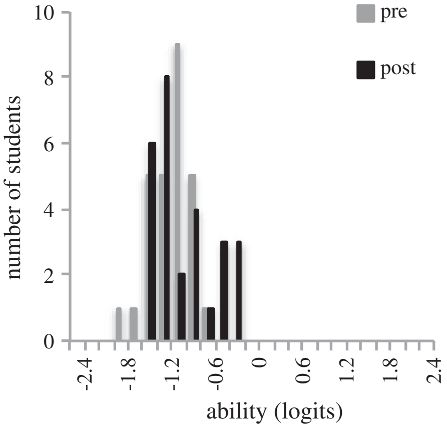

FIG. 42. The pre- and postability histogram for class 28.

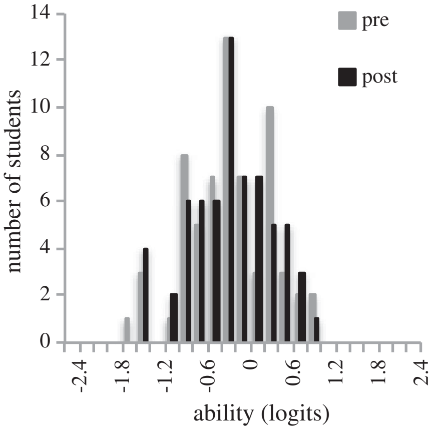

FIG. 43. The pre- and postability histogram for class 29.

[1] E. M. Barddar, E. E. Prather, K. Brecher, and T. F. Slater, Development and validation of the light and spectroscopy concept inventory, Astron. Educ. Rev. **5**, 103 (2007).

[2] The LSCI is available at http://ftp.aip.org/epaps/aer/E-AERSCZ-5-2006020/LSCIspring2006.pdf.

[3] E. E. Prather, A. L. Rudolph, G. Brissenden, and W. M. Schlingman, A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction, Am. J. Phys. **77**, 320 (2009).

[4] A. L. Rudolph, E. E. Prather, G. Brissenden, D. Consiglio, and V. Gonzaga, A national study assessing the teaching and learning of introductory astronomy. Part II. The connection between student demographics and learning, Astron. Educ. Rev. **9**, 010107 (2010).

[5] W. M. Schlingman, E. E. Prather, C. S. Wallace, A. L. Rudolph, and G. Brissenden, A classical test theory analysis of the Light and Spectroscopy Concept Inventory data set, Astron. Educ. Rev. **11**, 010107 (2012).

[6] R. K. Hambleton and R. J. Jones, Comparison of classical test theory and item response theory and their applications to test development, Educ. Meas.: Issues & Pract. **12**, 38 (1993).

[7] A. A. Rupp and B. D. Zumbo, Understanding parameter invariance in unidimensional IRT models, Educ. Psychol. Meas. **66**, 63 (2006).

[8] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[9] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[10] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[11] C. N. Cardamone, J. E. Abbott, S. Rayyan, D. T. Seaton, A. Pawl, and D. E. Pritchard, Item response theory analysis of the mechanics baseline test, AIP Conf. Proc. **1413**, 135 (2012).

[12] M. Planinic, The Rasch model-based analylsis of the Conceptual Survey of Electricity and Magnetism, in *Procedings of GIREP Conference 2006: Modeling in Physics and Physics Education* (University of Amsterdam, Amsterdam, NL, 2006), pp. 133–134.

[13] C. S. Wallace and J. M. Bailey, Do concept inventories actually measure anything?, Astron. Educ. Rev. **9**, 010116 (2010).

[14] K. E. Williamson, Development and calibration of a concept inventory to measure introductory college astronomy and physics students' understanding of Newtonian gravity, Ph.D. thesis, Montana State University, Bozeman, MT, 2013.

[15] P. M. Sadler, H. Coyle, J. L. Miller, N. Cook-Smith, M. Dussault, and R. R. Gould, The astronomy and space science concept inventory: Development and validation of assessment instruments aligned with the K-12 national science standards, Astron. Educ. Rev. **8**, 010111 (2010).

[16] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists* (Lawrence Erlbaum Associates, Mahwah, NJ, 2000).

[17] D. Harris, Comparison of 1-, 2-, and 3-Parameter IRT Models, Educ. Meas.: Issues & Pract. **8**, 35 (1989).

[18] L. Cai, D. Thissen, and S. H. C. du Toit, *IRTPRO for Windows [Computer software]* (Scientific Software International, Lincolnwood, IL, 2011).

[19] F. B. Baker and S. Kim, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed. (Dekker, New York, NY, 2004).

[20] A. Maydeu-Olivares and C. García-Forero, Goodness-of-Fit Testing, *International Encyclopedia of Education*, 3rd ed. (Elsevier, Amsterdam, NL, 2010), pp. 190–196.

[21] W. M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, Appl. Psychol. Meas. **8**, 125 (1984).

[22] W. M. Yen and A. R. Fitzpatrick, Item response theory, in *Educational Measurement*, 4th ed. (American Council on Education/Praeger, Westport, CT, 2006), pp. 111–153.

[23] W. M. Yen, Scaling performance assessments: Strategies for managing local item dependence, J. Educ. Measure. **30**, 187 (1993).

[24] F. Samejima, Estimation of latent ability using a response pattern of graded scores, Psychometrika **34**, 1 (1969).

[25] I. I. Bejar, A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates, J. Educ. Measure. **17**, 283 (1980).

[26] J. M. Utts and R. F. Heckard, *Mind on Statistics*, 2nd ed. (Thompson Brooks/Cole, Belmont, CA, 2004), p. 530.

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.010149 for the item characteristic curves for each item.

[28] M. M. Wooten, A. M. Cool, E. E. Prather, and K. D. Tanner, Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory, Phys. Rev. ST Phys. Educ. Res. **10**, 020103 (2014).

[29] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics course, Am. J. Phys. **66**, 64 (1998).

[30] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).

[31] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A., **111**, 8410 (2014).

[32] S. S. Sawilowsky, New effect size rules of thumb, J. Mod. Appl. Stat. Meth. **8**, 597 (2009).

[33] R. S. Lindell, Enhancing college students' understandings of lunar phases, Ph. D. thesis, University of Nebraska at Lincoln, 2001.

[34] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).