Investigating students' seriousness during selected conceptual inventory surveys

David P. Waters
Department of Basic Sciences, St. Louis College of Pharmacy, St. Louis, Missouri 63110, USA

Dragos Amarie
Department of Physics and Astronomy, Georgia Southern University, Statesboro, Georgia 30460, USA

Rebecca A. Booth
Calgary Board of Education, Calgary, Alberta T2R 0L4, Canada

Christopher Conover and Eleanor C. Sayre*
Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA

(Received 15 May 2019; published 26 August 2019)

Conceptual inventory surveys are routinely used in education research to identify student learning needs and assess instructional practices. Students might not fully engage with these instruments because of the low stakes attached to them. This paper explores tests that can be used to estimate the percentage of students in a population who might not have taken such surveys seriously. These three seriousness tests are the pattern recognition test, the easy questions test, and the uncommon answers test. These three tests are applied to sets of students who were assessed either by the Force Concept Inventory, the Conceptual Survey of Electricity and Magnetism, or the Brief Electricity and Magnetism Assessment. The results of our investigation are compared to computer simulated populations of random answers.

DOI: 10.1103/PhysRevPhysEducRes.15.020118

I. INTRODUCTION

Conceptual inventories (CIs) came out of our necessity to quantify students' understanding of concepts and their progress in class by monitoring learning gains [1]. The physics education research that followed has driven modern teaching with a focus on developing novel methods to stimulate students' understanding, and has also redefined our learning goals [2]. Halloun and Hestenes raised the concern that traditional instruction marginally affects students' understanding while their common sense beliefs usually contradict the laws of physics [3,4]. Their Force Concept Inventory (FCI) survey arrives as a first tool to measure students' mastery of force concepts widely taught in the first semester of physics [5]. Since then, CIs have gained widespread use in physics and astronomy education [6–8], as well as many other disciplines of STEM [9–22].

Since CIs became more useful to instructors, they started to be used as research-based assessment instruments (RBAIs) in education research [23]. RBAIs are multiple choice but carefully designed surveys to provide insight into students' attitudes and understanding. Over time, RBAIs have undergone different rounds of scrutiny and validation [24]. When RBAI data are collected regularly, they could be valuable measuring tools by providing standardized comparisons among institutions, instructors, and teaching methods, and over multiple implementations of the same course. They also allow us to track trends and investigate correlations over time [25,26]. The physics education research that has followed from the use of RBAIs has driven physics instructors toward developing and implementing novel methods for increasing student understanding as well as toward redefining student learning goals [2]. PhysPort, an online resource for instructors interested in implementing research-based physics teaching practices in their classrooms, currently provides 92 RBAIs with diverse foci, including content knowledge, problem solving, scientific reasoning, lab skills, beliefs and attitudes, and interactive teaching [27].

Among the RBAIs available on PhysPort are the Force Concept Inventory (FCI), the Brief Electricity and Magnetism Assessment (BEMA), and the Conceptual Survey of Electricity and Magnetism (CSEM). The FCI is a 30-question RBAI used to measure student mastery of the mechanics concepts widely taught in a first-semester introductory physics course [5]. The FCI is among the most popular RBAIs, with extensive research on its efficacy and

*Sayre: esayre@ksu.edu

effects on instruction as well as many translations into different languages and formats. In particular, the FCI has been investigated by Hestenes *et al.*, who interviewed students and instructors to confirm that surveyed individuals correctly understood the wording and the pictographs [5,28], whereas Stewart *et al.* confirm that test scores are not particularly context dependent [29]. Version H of the CSEM, published by Maloney, O'Kuma, Hieggelke, and Van Heuvelen in 2001 [30], is a 32-question RBAI used to measure student conceptual understanding of electricity and magnetism at an introductory undergraduate level. The BEMA is a 31-question RBAI also designed to assess conceptual understanding of electromagnetism.

A. Main concerns with RBAIs

From the early days of RBAIs [31], researchers and instructors have raised concerns about whether students might not make a serious attempt at answering the questions on a conceptual-inventory RBAI, such as the FCI, CSEM, or BEMA correctly [23,32–35]. In order for instructors and researchers to appropriately evaluate RBAI data, it is useful to know what proportion of students in a population are taking that RBAI seriously. We define serious students as those who chose answers with consideration, including educated and/or thoughtful guesses, throughout their entire assessment.

Stewart *et al.* [36,37] study the effect of guessing on both the FCI and CSEM tests. They show that gains are invariant to linear transformation and therefore unaffected by guessing and as such their linear models can correct the test results to account for guessing. Yasuda *et al.* show that while question 5 scores on the FCI are marginally affected by erroneous reasoning, questions 6, 7, and 16 are more prone to guessing. These questions return a high percentage of false positives as students seem to reach the right answer while using erroneous conceptual reasoning [38,39].

Wang *et al.* implement item response theory [40] to build a three-parameter item response model and use it to analyze student performance on FCI surveys [41]. They show that a student's proficiency is in linear correlation to a student's raw FCI score. They find that low proficiency students have less than a 5% chance of guessing the correct answer on questions 23 and 26, and a 34% chance of guessing correctly on question 16. They predict that questions 1 and 6 are the easiest, whereas questions 25 and 26 are the most difficult. As anticipated, each of the 30 questions in the FCI has a different guessing chance and difficulty level, which supports our present work hypothesizing that when students take the survey seriously, there is a better chance that they will select the correct answer for those questions [41]. Hake *et al.* considered that motivational factors can persuade students to take the RBAIs seriously. Without much evidence at the time, he made the remark that surveyed students did take the [FCI] pretest seriously [30]. Later, Henderson showed that about 2.8% of surveyed students may not take an RBAI seriously [31]. Henderson was concerned about whether students take the FCI seriously when it is not graded. To identify those students, answer patterns were examined for lack of seriousness from five different angles. By comparison, Pollock *et al.* ran a longitudinal study of students' conceptual understanding using the BEMA survey, and requested that students report how hard they tried. Three levels were identified: take it very seriously, take it seriously, and do not take it seriously. This study shows that over 50% of students took the RBAI very seriously, and only 3% indicated that they did not take it seriously [42].

We have developed a set of seriousness tests and applied them to the FCI, CSEM, and BEMA. It was our goal to develop seriousness tests that could give instructors and researchers an estimate for the proportion of students who did not take an RBAI seriously. Notably, it was not our goal to develop seriousness tests that could identify individual students, and we recommend that the seriousness tests described in this paper not be used in that manner. In subsequent sections we will describe how these seriousness tests were developed as well as those tests' effectiveness in accurately categorizing students as either taking an RBAI seriously or not.

II. DATA SOURCES FOR THE FCI, CSEM, AND BEMA

Data for this paper were obtained from PhysPort's collection of student data. After administering an RBAI, instructors can use the PhysPort Data Explorer to analyze the data from their students. Once the instructors have uploaded their students' responses, the data are stored in a database in PhysPort. We were able to use the data from this database to run our seriousness tests on both the pre- and post-test data for the FCI, CSEM, and BEMA. The database is larger than any data set that has been tested previously, with 64 076 assessment results for the FCI, 15 032 assessment results for the CSEM, and 8708 assessment results for the BEMA. Table I presents the average and the standard deviation for each RBAI.

Along with the RBAI results from PhysPort, we created 20 000 simulated RBAI results each for the FCI, CSEM, and BEMA. Our simulated students guessed randomly on all questions. We generated this simulated data in order to

TABLE I. Means and standard deviations (SD) for each RBAI.

	RBAI	Pre	Post	Overall
Mean (%)	FCI	40.5	57.3	47.8
	CSEM	27.8	43.8	38.9
	BEMA	22.8	46.1	38.7
SD (%)	FCI	20.5	22.1	22.8
	CSEM	12.9	18.6	18.6
	BEMA	10.1	18.1	19.4

model the responses we might expect from nonserious students. Because we could be certain that each simulated individual in the random dataset was a random guesser, the seriousness tests needed to flag a significant fraction of this population in order to be considered successful. We did not expect our seriousness tests to identify every member of the simulated population as nonserious, however, because a seriousness test that achieves this would likely lead to misidentifying serious students as nonserious. It should also be noted that real students are almost never able to behave in a truly random manner on an RBAI, even when they are being nonserious. Their results might show tendencies toward certain answer choices, patterns on the answer sheet, or other trends. This means that students might exist who do not take an RBAI seriously, who also are not well represented in the simulated population.

III. THE SERIOUSNESS TESTS

We developed three seriousness tests that can be applied to FCI, CSEM, and BEMA responses in order to estimate the percent of students in a sample who did not take that RBAI seriously: the pattern recognition test (PRT), the uncommon answers test (UAT), and the easy questions test (EQT). These seriousness tests are not designed, however, to identify individual students who did not take an RBAI seriously. In developing these tests, we made the assumption, based on the previous work from Henderson as well as from Pollock *et al.*, that the majority of students take RBAIs seriously. As such, we expect the portion of the real population that a successful seriousness test identifies to be small.

A. The pattern recognition test

The pattern recognition test (PRT) is based on the premise that students who do not take a RBAI seriously might choose instead to leave certain patterns throughout their answers. Since computers are not good at picking up on these patterns, we came up with patterns based on what we thought would be likely to be found from nonserious test takers. The patterns that we searched for in the RBAIs were as follows:

TABLE II. Nine questions on each RBAI where a small percentage of the population chose uncommon answers.

RBAI	FCI								
Question	4	6	12	15	16	22	24	27	29
Answers	b,c,d	c,d,e	a,d,e	d,e	d,e	c,e	b,d	d,e	c,e
RBAI	CSEM								
Question	1	3	4	7	8	12	13	18	24
Answers	a,e	a,e	a,e	d,e	a,e	c,e	c,d	a,b	a,e
RBAI	BEMA								
Question	1	2	3	4	5	10	14	21	25
Answers	c,d,g	c,d,g	e,h,i	b,d,j	d,j	h,i	d,e,f	d,j	a,h

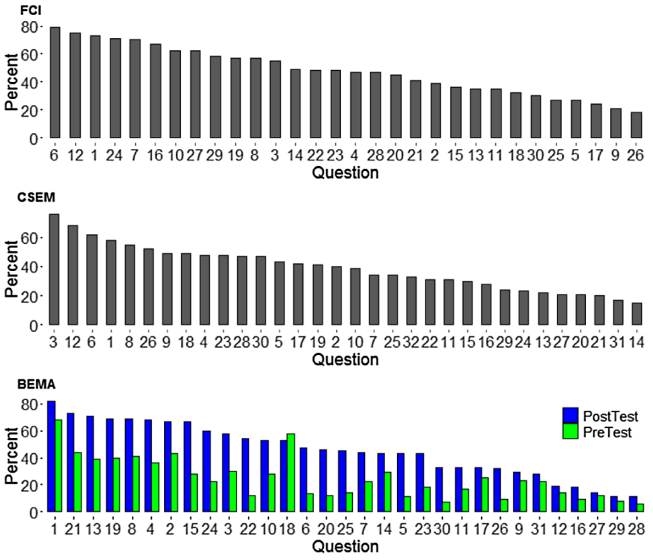

FIG. 1. Score distribution for each question of each RBAI. The pre- and post-test scores are combined for the FCI and CSEM.

(i) more than 50% zeros or blank answers
(ii) more than 50% one letter
(iii) 8 of the same letter in a row
(iv) 3 instances of ABCD
(v) 2 instances of ABCDE
(vi) 1 instance of ABCDEDCBA

When these patterns are present in a response, it is likely that the test taker was not taking the RBAI seriously for a significant portion of the test. The correct answers on none of the RBAI evaluated follow any of these patterns. It should also be noted that nonserious students might sometimes produce response patterns outside of those listed

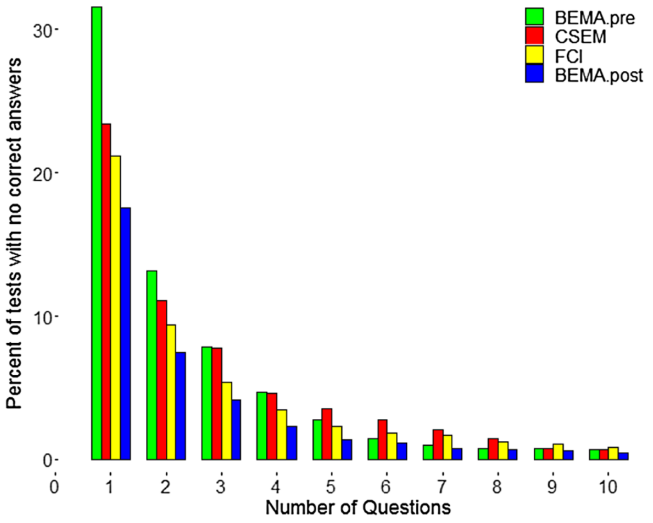

FIG. 2. Justification for the number of easy test questions: Percentage of real test takers who answered none of the questions correctly for an increasing number of easy questions as measured by the easy test. The order of the easy test questions was determined by the data in Fig. 1.

TABLE III. Four questions on each RBAI with the highest percentage of students choosing the right answer.

RBAI			FCI	
Question	1	6	12	24
RBAI			CSEM	
Question	1	3	6	12
RBAI			BEMA	
Pre-test Question	1	2	18	21
Post-test Question	1	13	19	21

above. We limited the patterns that we sought for, however, to avoid misidentifying serious students as nonserious.

B. The uncommon answers test

The uncommon answers test (UAT) is based on the idea that students who do not take an RBAI seriously sometimes choose answers that were uncommonly chosen by the larger student population. There are nine questions on each of the RBAIs where two or three answer choices were preferred by most of the population. The common answers were most often the correct answer plus one or more of the incorrect answers. Evidently, these preferred choices are attractive to people who were reading carefully through all questions and were being thoughtful in their responses.

If a student chose an unpopular answer on several of these questions, it is likely that they were guessing rather than applying reasoning throughout the assessment. We identified uncommon answer choices based on how few students have picked those answers in the existing PhysPort data. Table II summarizes the questions and the less frequently chosen answers. We identified 9 questions with uncommon answers for each RBAI. For the FCI, fewer than 7% of the population chose one of the uncommon answers for each identified question. For the CSEM, fewer than 10% of students chose one of the uncommon answers for each identified question. For the BEMA, fewer than 6% of students chose one of the uncommon answers for each identified question. We counted survey takers who choose at least 4 uncommon answers for the FCI or CSEM or at least 3 uncommon answers for the BEMA as possibly nonserious.

C. The easy questions test

The easy questions test (EQT) was based on the idea that students who take a concept-inventory RBAI seriously will get most of the easier questions correct. A student making an effort on such a RBAI might still have one or two of even these questions incorrect, but they are unlikely to be incorrect for all the easy questions. It stands to reason that an answer set in which all the responses to the easy questions are incorrect is more likely to come from a student who did not take that assessment seriously.

We looked at the existing PhysPort data to determine which questions were easiest for students (Fig. 1). For each

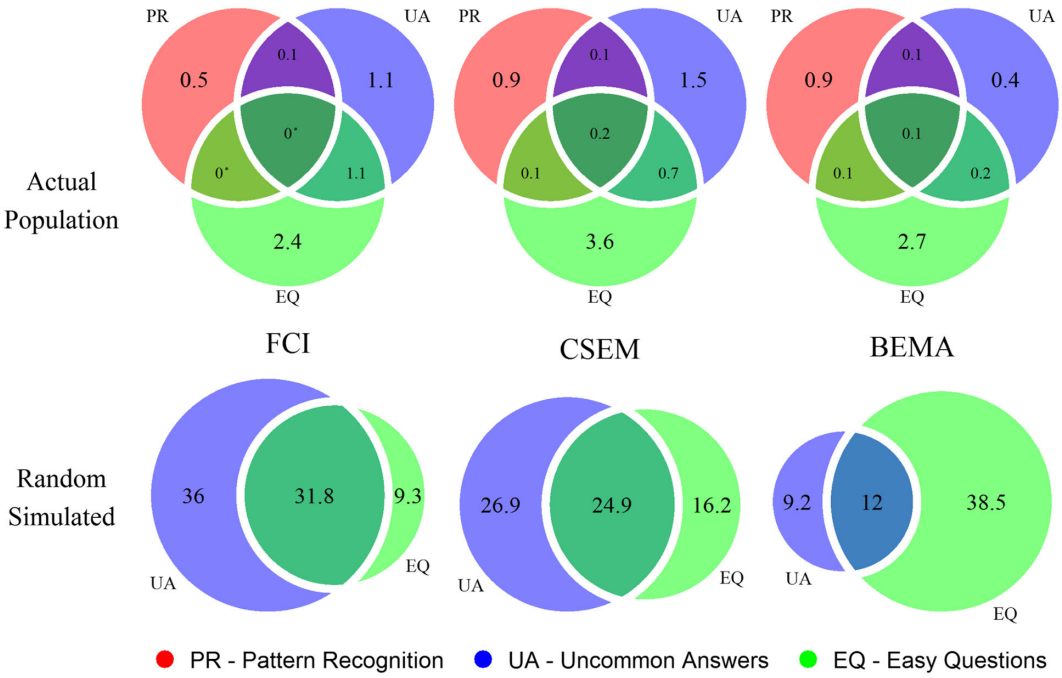

FIG. 3. Venn diagrams showing the percent of nonserious responses caught by each of the three tests for each of the datasets for each RBAI. The percent in each segment is based on the symmetric difference, where each segment includes what is not in the intersection. Note that the PRT results are excluded from the random simulation because almost none of the simulated RBAI results exhibited the patterns sought for by that test. Values with an asterisk round to zero and represent fewer than 32 (FCI), 8 (CSEM), and 5 (BEMA) students.

RBAI, we chose the top four questions which had the highest scores, and calculated the percent of students who got a certain number of those questions correct. The students who answered all four easy questions incorrectly were considered as not having taken the assessment seriously. We note, however, that even a random guesser is likely to choose at least one correct answer in any set of five-choice questions. Overall, this means that the EQT will undercount the number of non-serious test takers.

Figure 2 shows the percent of students who answered all of the easy test questions incorrectly for an increasing number of easy questions. We can see that when we choose four or more easy questions, the percent of students who get all of the questions wrong stays relatively constant. For this reason, we chose the four easiest questions from each RBAI based on the proportion of correct responses to that question. The questions chosen for the EQT for each RBAI are shown in Table III. For the FCI, an easy question has a score greater than 71%. For the CSEM, an easy question is one with a percent of correct responses above 58%. As shown in Fig. 1 for the BEMA, there is a large discrepancy on the pre- and post-test as to which questions are easy questions. The easy questions for the pre- and post-tests of the BEMA had scores greater than 43% and 69%, respectively.

IV. ANALYSIS AND DISCUSSION

Results of the PRT, UAT, and EQT are shown in Table IV and in Fig. 3. In Fig. 3, each of the segments includes the percent of test takers caught by that test, excluding what is shown in the intersecting segments. As an example, the percent of the actual population from the CSEM caught by the PRT is 1.3%, as shown in Table IV. This comes from combining 0.92% with each of the segments within the entire PRT circle. The percent of the actual population found to be nonserious by each of these tests is very small, ranging from less than 1% up to a few percent.

Applying the PRT to the actual population data identifies between 0.6% and 1.3% as nonserious for each of the different RBAIs. The PRT identified nearly zero nonserious survey takers in the simulated population, however. This is unsurprising because the patterns sought for are nonrandom. Pattern recognition was thus excluded from Fig. 3 for the random simulated results.

For comparison, the UAT and EQT found a high proportion of nonserious responses in the random simulated data. Sixty-eight percent and 52% of the respective simulated populations for the FCI and CSEM had four or more uncommon answers in their responses, and 21% of the BEMA random population had three or more uncommon answers (blue circle). Forty-one percent, 41%, and 51% of the respective simulated populations for the FCI, CSEM, and BEMA were identified by the EQT as nonserious (green circle). In the actual populations, on the other hand, the UAT identified slightly more than 2% of students as nonserious for each RBAI, and the EQT identified between 3% and 4.6% of students as nonserious for each RBAI.

Comparing the uncommon answers chosen for each data set and each RBAI in Fig. 4, we see that the actual population chose fewer uncommon answers than the random simulation. Fewer than 50% of the real students selected any uncommon answers. Conversely, there was an

TABLE IV. Percentage of each population detected as nonserious for each RBAI by each of the three seriousness tests.

Test	PRT	UAT	EQT
FCI	0.63	3.5	2.3
CSEM	1.3	4.6	2.5
BEMA	1.2	3.1	3.2

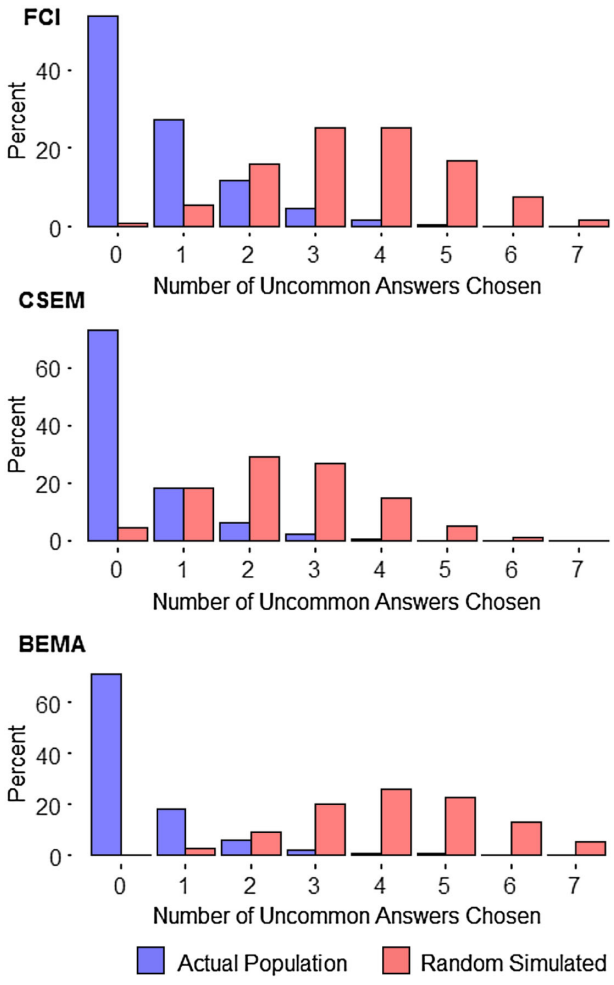

FIG. 4. Distribution of the percentage of assessments that selected a number of uncommon answers for both the real and simulated populations for each RBAI. There were 9 questions for each RBAI where uncommon answers are rare. Nonserious test takers were those who chose at least 4 uncommon answers for the FCI and CSEM and at least 3 uncommon answers for the BEMA.

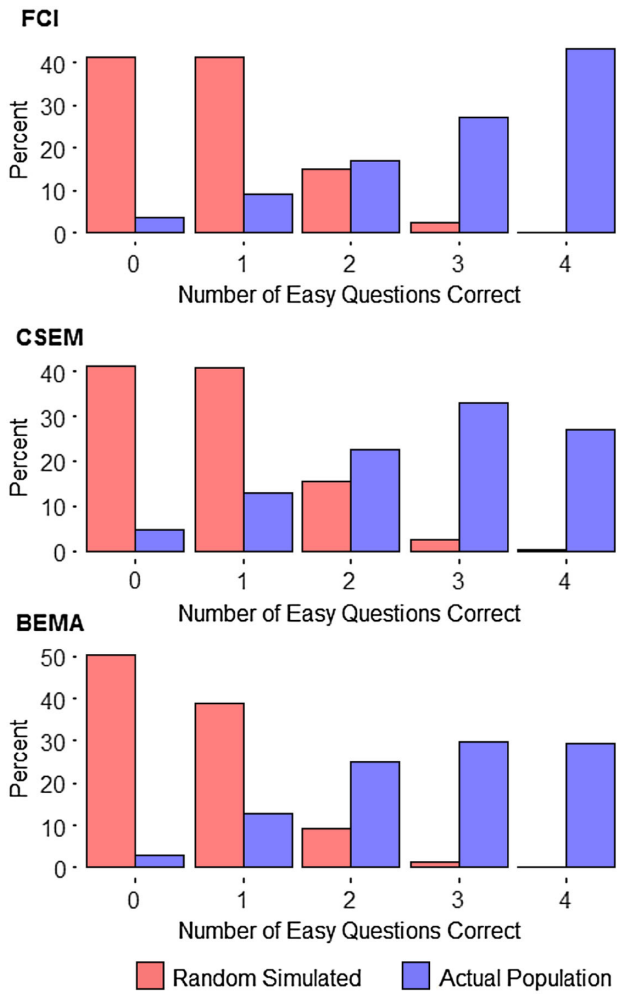

FIG. 5. Distribution of the percentage of assessments that correctly answered a number of easy questions comparing both the actual and simulated populations for each RBAI. There were 4 questions for each RBAI. Nonserious test takers were those who answered all 4 questions incorrectly.

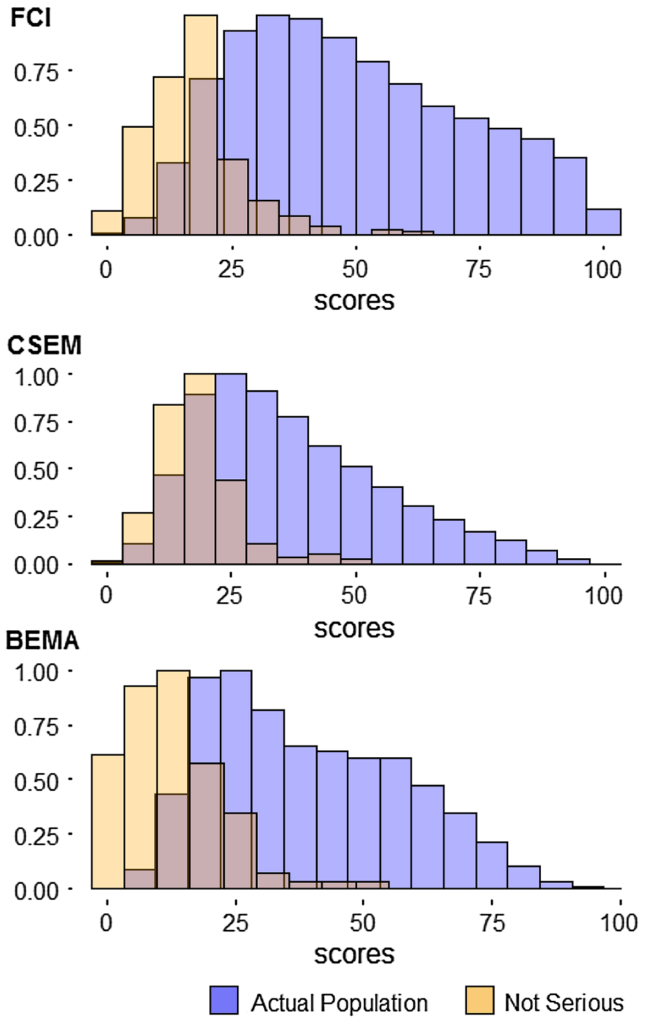

FIG. 6. All vs not-serious scores: Distribution of scores for each RBAI comparing all of the scores in the actual population to the responses caught by the tests and considered a nonserious attempt by the actual population.

average of three or four uncommon answers for the random population. Therefore, testing for the number of uncommon answers helped us differentiate serious students from random guessers. For the easy questions test, we saw in Fig. 5 that the simulated population chose fewer correct easy questions while most of the actual population were able to correctly answer at least two easy questions.

A. Combining the results to determine the overall percent of nonserious students

The percent of the actual population identified by each of the three seriousness tests as nonserious is small, ranging from less than 1% up to a few percent. The center segments of Fig. 3 show that very few real test takers were caught by all three tests for any of the RBAIs, with the FCI catching only 0.016% with all 3 tests.

As a final comparison, we looked at the scores of any test taker in the actual population who was identified as nonserious by the PRT as well as both of the other two seriousness tests. When we look at the tests caught by the PRT and tests caught by both the EQT and the UAT, the percent of test takers not taking the assessment seriously was 1.5%–2.2% of the population. The number of test takers in the actual population caught by the nonserious tests was 1120 out of 63 896 assessments (1.8%) for the FCI, 329 out of 14 876 assessments (2.2%) for the CSEM, and 133 out of 8642 assessments (1.5%) for the BEMA. These values could be determined using Fig. 3 by combining the percentages within the entire red (PR) circle with the overlapping segment between UA and EQ. These results were very similar, suggesting that our seriousness tests accurately determine the percent of students who did not take a RBAI seriously. Figure 6 shows graphs of the nonserious scores, identified in the combined manner described vs all of the scores in the actual population.

From this graph, we can see that the nonserious assessment scores are much lower than those from the actual population. This is further evidence that the scores identified by a combination of the PRT and both the UAT and the EQT were most likely truly not serious.

V. CONCLUSION

Our results are in contrast to work mentioned in the introduction by Henderson, who found that about 2.8% of students did not take the FCI seriously [32], and the results from Pollock *et al.* who found that 3% indicated that they did not take the BEMA seriously [42]. We find that fewer students were caught by our seriousness tests, and we conclude that the overall percentage of students who did not take the CIs seriously is only about 1.5%–1.6%. Regardless, our results are in line with previous work that shows that the incidence of nonseriousness in RBAI results is very low.

In addition, our seriousness tests might undercount incidents of nonseriousness. We made deliberate choices to avoid misidentifying serious test takers as nonserious, and those choices could have resulted in misidentifying some nonserious test takers as serious. It is still likely, however, that the methods described will sometimes falsely identify a serious student as nonserious. Because of this, we do not recommend using any of these seriousness tests to identify individual students as serious or nonserious.

We suggest that the three seriousness tests developed here could be used together as described to give reasonable estimates of percents of nonseriousness for FCI, CSEM, and BEMA datasets, and that results similar to those just described indicate a low incidence of nonseriousness in a dataset. In addition, these seriousness tests might be applied to other concept-inventory RBAIs, although some details would need to be worked out for the UAT and EQT for each RBAI.

ACKNOWLEDGMENTS

This work was partially funded using the PhysPort Data Explorer, NSF DUE-1347821/1347728.

[1] D. Sands, M. Parker, H. Hedgeland, S. Jordan, and R. Galloway, Using concept inventories to measure understanding, Higher Educ. Pedagogies **3**, 173 (2018).
[2] J. T. Laverty and M. D. Caballero, Analysis of the most common concept inventories in physics: What are we assessing?, Phys. Rev. Phys. Educ. Res. **14**, 010123 (2018).
[3] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).
[4] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).
[5] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[6] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).
[7] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).
[8] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource letter RBAI-2: Research-based assessment instruments: Beyond physics topics, Am. J. Phys. **87**, 350 (2019).
[9] S. R. Singer, N. R. Nielsen, and H. A. Schweingruber, *Discipline-based education research: understanding and improving learning in undergraduate science and engineering* (National Academies Press, Washington, DC, 2012).
[10] M. W. Klymkowsky and K. Garvin-Doxas, Recognizing student misconceptions through Ed's tools and the biology concept inventory, PLoS Biology **6**, 14 (2008).
[11] C. D'Avanzo, Biology concept inventories: Overview, status, and next steps, BioScience **58**, 1079 (2008).
[12] D. R. Mulford, An inventory for measuring college students' level of misconceptions in first semester chemistry, in Master of Science thesis, Purdue University, 1996.
[13] D. R. Mulford and W. R. Robinson, An inventory for alternate conceptions among first-semester general chemistry students, J. Chem. Educ. **79**, 739 (2002).
[14] K. Goldman, P. Gross, C. Heeren, G. Herman, L. Kaczmarczyk, M. C. Loui, and C. Zilles, Identifying important and difficult concepts in introductory computing courses using a Delphi process, in *Proceedings of the 39th ACM SIGCSE Technical Symposium on Computer Science Education, Portland, OR* (Curran Associates, Inc., Red Hook, NY, 2008).
[15] K. Goldman, P. Gross, C. Heeren, G. L. Herman, L. Kaczmarczyk, M. C. Loui, and C. B. Zilles, Setting the scope of concept inventories for introductory computing subjects, ACM Trans. Computing Educ. **10**, 1 (2010).
[16] B. Vice Bowling, E. E. Acra, L. Wang, M. F. Myers, G. E. Dean, G. C. Markle, C. L. Moskalik, and C. A. Huether, Development and evaluation of a genetics literacy assessment instrument for undergraduates, Genetics **178**, 15 (2008).
[17] M. K. Smith, W. B. Wood, and J. K. Knight, The genetics concept assessment: A new concept inventory for gauging student understanding of genetics, CBE Life Sci. Educ. **7**, 422 (2008).
[18] S. Singer and K. A. Smith, Discipline-based education research: Understanding and improving learning in undergraduate science and engineering, J. Engin. Educ. **102**, 468 (2013).

[19] J. W. Pellegrino, L. V. DiBello, K. James, N. Jorion, and L. Schroeder, Concept inventories as aids for instruction: A validity framework with examples of application, in *Proceedings of the Research in Engineering Education Symposium* (Curran Associates, Inc., Red Hook, NY, 2011).

[20] N. Jorion, B. D. Gane, L. V. DiBello, and J. W. Pellegrino, Developing and validating a concept inventory, in *Proceedings of the 122nd ASEE Annual Conference and Exposition* (Curran Associates, Inc., Red Hook, NY, 2015).

[21] J. Epstein, Development and validation of the calculus concept inventory, in *Proceedings of the 9th International Conference on Mathematics Education in a Global Community* (University of North Carolina, Charlotte, NC, 2007).

[22] J. Epstein, The calculus concept inventory—measurement of the effect of teaching methodology in mathematics, Not. Am. Math. Soc. **60**, 1018 (2013).

[23] A. Madsen, S. McKagan, and E. C. Sayre, Best practices for administering concept inventories, Phys. Teach. **55**, 530 (2017).

[24] S. Bates and R. Galloway, Diagnostic tests for the physical sciences: A brief review, New Directions Teaching Phys. Sci. **6**, 10 (2010).

[25] C. Henderson, C. Turpen, M. Dancy, and T. Chapman, Assessment of teaching effectiveness: Lack of alignment between instructors, institutions, and research recommendations, Phys. Rev. ST Phys. Educ. Res. **10**, 010106 (2014).

[26] A. Madsen, S. B. McKagan, M. S. Martinuk, A. Bell, and E. C. Sayre, Research-based assessment affordances and constraints: Perceptions of physics faculty, Phys. Rev. Phys. Educ. Res. **12**, 010115 (2016).

[27] American Association of Physics Teachers and Kansas State University, PhysPort: Supporting physics teaching with research-based resources, National Science Foundation, 2011–2018. Available: https://www.physport.org/. [Accessed Dec. 2018].

[28] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[29] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).

[30] D. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

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

[32] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).

[33] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[34] R. J. Dufresne, W. J. Leonard, and W. J. Gerace, Making sense of students' answers to multiple-choice questions, Phys. Teach. **40**, 174 (2002).

[35] R. B. Frary, Formula scoring of multiple-choice tests (correction for guessing), Educ. Meas. **7**, 33 (1988).

[36] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).

[37] J. Stewart and G. Stewart, Correcting the normalized gain for guessing, Phys. Teach. **48**, 194 (2010).

[38] J.-i. Yasuda and M.-a. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, Phys. Rev. ST Phys. Educ. Res. **9**, 010113 (2013).

[39] J.-i. Yasuda, N. Mae, M. M. Hull, and M.-a. Taniguchi, Analyzing false positives of four questions in the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010112 (2018).

[40] D. Harris, Comparison of 1-, 2-, and 3-parameter IRT models, Educ. Meas. **8**, 35 (1989).

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

[42] S. J. Pollock, Longitudinal study of student conceptual understanding in electricity and magnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020110 (2009).