Visualizing patterns in CSEM responses to assess student conceptual understanding

Ryan Tapping,[1] G.P. Lepage,[1] and N.G. Holmes[1]

[1]*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*

The Conceptual Survey of Electricity and Magnetism (CSEM) has been utilized to measure learning gains in electricity and magnetism (E&M) physics courses, where "correct" vs "incorrect" responses are typically used for analysis. However, such comparisons do not necessarily identify specific changes in student reasoning from pre- to post-instruction. To address this issue, we have generated network-like graphs for each question: Responses at pre- and post-test are represented by nodes connected by edges representing the change in student response choice. We demonstrate the visualizations using data from CSEM responses from over 2500 students at Cornell University across 12 semesters of an introductory E&M course. We demonstrate a vector analysis method that can categorize response patterns and quantify the way students change their responses. We show the potential use of these methods for both instructors as well as for answering deeper research questions.

I. INTRODUCTION

Multiple choice tests, such as the Force Concept Inventory (FCI) and the Conceptual Survey of Electricity and Magnetism (CSEM), are commonly used for assessment of the effectiveness of introductory physics courses [1–3]. These assessments have the potential to address and identify student misconceptions before and after instruction. Analysis of such tests is thus very important to understanding and answering possible research questions, as well as for instructor feedback.

There are many methods and statistics used by both instructors and researchers to evaluate teaching and view student performance on such assessments. Simple reports of average scores before and after instruction are useful, but directly comparing semesters is challenging if pre-test performance is different. A common statistic used in education research for accounting for pre-test performance differences is Hake's gain [4], but this statistic has significant limitations depending on comparison groups [5, 6] and even pre-test score [7]. In addition to Hake's gain, researchers also use methods such as classical test theory, item response theory, and factor analysis [8]. However, all of these methods use binary models of correct and incorrect responses to items and miss significant detail about how individual student reasoning changes from pre- to post-test. This presents a strong need, therefore, for tools to distill the data while maintaining detail to understand the nuances and impacts of instruction on student learning.

We use a visualization, referred to as network-like graphs, for a more complete picture of the complexity of student responses. Rather than simplifying data to correct and incorrect responses, the network-like graphs (similar to Sankey diagrams) provide an easily digestible view of all students' responses to and from any answer choice. Motivated by ideas from factor analysis, we also present a vector analysis method to assess and categorize student response patterns, which may be equally valuable to both instructors and researchers. The network-like graphs and this analysis method are versatile and can both be applied to any multiple choice test that has pre- and post-test conditions.

II. VISUALIZATION METHODS

In this work, we study the student response patterns to the CSEM administered at Cornell University in an introductory electricity and magnetism physics course. The CSEM was administered online to STEM undergraduates as a pre- and post-test in the first and final weeks of the semester, respectively. Students received participation credit to encourage responses. Matched responses from 2514 students over 12 semesters from Fall 2012 to Spring 2018 were used in this study. To be included in the study, students must have been enrolled in the course, taken both the pre- and post-tests, and have responded to over 25% of the items on the CSEM. Responses to the CSEM were aggregated across all semesters, with an average (standard deviation) of $41\%(16\%)$ on the pre-test and $61\%(20\%)$ on the post-test.

For each item on the CSEM, a network-like graph is created to visualize the data. Each item on the CSEM has 6 possible responses recorded: a, b, c, d, e, and no response. The pre- and post-test item responses are represented by nodes on the left and right, respectively. The node size represents the relative proportion of students who selected a certain item response. The nodes are ordered from top to bottom, where the topmost node is always the correct answer and is colored in blue. The remaining nodes are ordered from the most to least common responses based on post-test results. Edges are also drawn between nodes, representing the transition from a pre-test response to a post-test response. The width of an edge is determined by the proportion of students who gave that specific pre- and post-test response combination. To enhance the visualization of each network-like graph, we have removed edges with widths less than the average width size.

We use one item from the CSEM as an example of the utility of the network-like graphs (Fig. 1). The results are shown by a histogram in Fig. 1a. The responses from pre- and post-test are both displayed in a single histogram to show how student responses change before and after instruction. For this question, 32.4% of students answered correctly on the pre-test, with an improvement to 47.3% on the post-test. The correct answer choice for this question was choice "d", however it is clear that a common distractor answer of choice

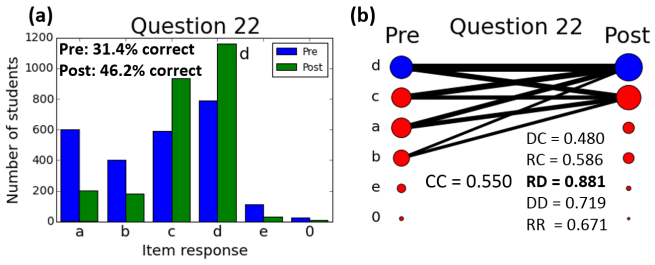

FIG. 1. Data from question 22 on the CSEM. (a) A histogram showing the pre-test and post-test responses along with the percentage of students who chose the correct answer. (b) A network-like graph showing the same data with edges representing how responses changed between pre-test and post-test. Dot product values from the representative vectors are also shown.

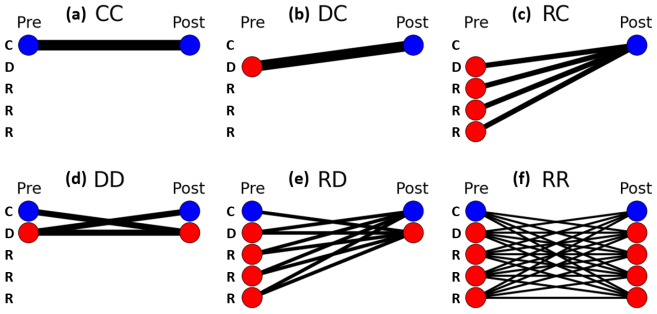

FIG. 2. The six representative response patterns used for vector analysis. (a) "Correct to Correct" (CC) (b) "Distractor to Correct" (DC) (c) "Random to Correct" (RC) (d) "Distractor to Distractor" (DD) (e) "Random to Distractor" (RD) (f) "Random to Random" (RR)

"c" is still prevalent, even after instruction. Choices "a" and "b" had a decrease in responses, with choices "c" and "d" becoming the two most common answers.

The results are also displayed in a network-like graph, as in Fig. 1b. The same information from the histogram can be determined from the nodes of the network-like graph. It is clear from both plots that four of the five responses are relatively popular on the pre-test, whereas two responses are popular on the post-test.

The network-like graph also provides additional information beyond what is given by a histogram by showing precisely how student answers changed before and after instruction. In this example, the network-like graph shows that although the correct response became more popular after instruction, a large portion of students switched answers from the correct response (d) to the most common distractor (c) after instruction. In fact, students had a nearly even split between choosing the correct response and the common distractor regardless of their pre-test answer choice. This information is not conveyed in a histogram, where pre- and post-test responses are treated independently.

III. VECTOR ANALYSIS

To classify and quantify response patterns, we represent the data using a 36-dimensional vector, where each component is the number of students that answered with a certain response combination (6 possible responses for each pre- and post-test, thus 36 possible combinations). The vector is then normalized to unit length. Thus, a vector is a representation of the edges in the network-like graphs. The power of using this representation is that vectors can be compared quantitatively to other vectors.

One idea is to classify the response patterns we observe using comparisons to predetermined patterns. Based on the responses, we determined three general possible outcomes for either the pre- or post-test responses: "Correct", where the majority of students answer with the correct answer, "Distractor", where a significant portion of students chose a specific incorrect item response, and "Random", where a significant portion of students did not tend towards any specific item response. With these three possible responses on the pre- and post-test, there are nine total possible combinations of pre- to post-test response patterns.

Based on our data, there were six common combinations which made up our representative pattern responses: "Correct to Correct" (CC), "Distractor to Correct" (DC) and "Random to Correct" (RC) all representing positive learning outcomes after taking the course, whereas "Distractor to Distractor" (DD), "Random to Distractor" (RD), and "Random to Random" (RR) all representing cases where there is still a prevalent distractor and possible lack of understanding on the post-test, even after instruction.

To quantify these categories, we constructed vectors to classify the patterns, referred to as representative vectors. We use the same procedure as outlined earlier for the vectors constructed from the data, with two differences. First, the representative vectors assume an even distribution of hypothetical students. Second, the edge between the correct response pre- and post-test was removed. This was done for two purposes: to emphasize response patterns from students who did not already know the material, and to isolate the six response patters by using relatively more orthogonal vectors, as determined by a dot product between them. The representative vectors are all normalized. A network-like graph for each of these six scenarios is shown in Fig. 2.

Using the dot product of the representative vectors with a vector constructed from the course data, the student response patterns are identified and quantified. This is a similar geometrical view to factor analysis, where the effective "angle" between two vectors yields how similar they are [8]. If two vectors are parallel, then the dot product between them is maximized at 1, and if they are orthogonal, then the dot product yields a minimum value of 0. As part of our procedure, a dot product with the CC vector is reported first, then the data vectors are renormalized without the CC edge for the remaining dot products. For the purposes of this analysis, values above 0.9 show very strong alignment between vectors, and values above 0.8 signify a strong similarity.

From qualitative comparisons of the network-like graphs in Fig. 1 and Fig. 2, it is clear that the response pattern from question 22 most closely aligns with the RD graph, i.e. a "Random to Distractor" response pattern. Thus, we expect that taking the dot product of the vector formed from the raw data in Fig. 1 with the RD vector in Fig. 2 will yield the largest value, as compared to the dot products taken with the other representative vectors. The results of the dot products with each of the representative vectors for this question are shown in Fig. 1b. The largest value obtained is indeed from the RD vector with a value of 0.881, which shows a very strong alignment between the data set and the RD vector. The next largest dot product is from the DD vector with a value of 0.719, which also demonstrates that many students actually go from the correct to distractor answer. This information would not be visible in a histogram.

We repeated the same dot product procedure with the representative vectors for every question on the CSEM. Examples of questions that are qualitatively categorized by the different response patterns are shown in Fig. 3. The dot product values are shown within the network-like graphs, and the values that are above the threshold of 0.8 are shown in bold. We can see that for each graph the representative vectors with dot product values above the threshold appear to describe the data. We will go through each question individually to highlight how one may interpret and use this information.

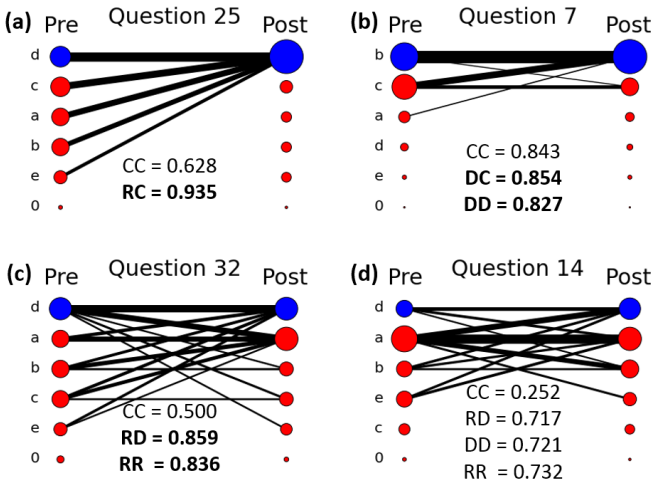

FIG. 3. Four examples of network-like graphs from data taken for questions on the CSEM. The largest values of the dot products to the six representative vectors are shown.

A. Examples

Question 25. Students typically chose randomly on the pre-test, but tended strongly towards the correct answer on the post-test (Fig. 3a). The vector from this data aligns quite closely with that of the RC vector with the dot product value being 0.935. The next highest value is from the RD vector with a dot product of 0.748 and all other dot products have values below 0.7. These results demonstrate that the dot products calculate what we believe they should. This question is an excellent example of a concept that students did not understand before instruction, but that was well covered by the course.

Question 7. Students exhibited a split between two options on the pre-test and tended towards the correct response on the post-test (Fig. 3b). The dot product with the largest value is from the DC vector with a value of 0.854, which agrees with the qualitative view of students going from a possible misconception to the correct answer after instruction. However, the dot product from the DD vector is also high, with a value of 0.827, meaning that the distractor answer was still fairly common even after instruction. From this information, an instructor may still consider this concept fairly straightforward, but may also want to emphasize a correction to that particular misconception during the course.

Question 32. The pattern of responses appears to go from a mostly random selection on the pre-test towards options "a" and "d" on the post-test (Fig. 3c). This behavior most closely aligns with the RD vector, and indeed the highest dot product value is from this vector with a value of 0.859. The RR vector also has a relatively high value of 0.836. Notably, many students go from the correct answer pre-test, to other answers post-test, which would not be seen in a histogram. In fact, the overall improvement on this question is quite small, with 33.6% of students answering correctly pre-test, and 35.3% post-test. The most common post-test response was option "a", an incorrect answer. An instructor may want to address this concept differently for future iterations of the course.

Question 14. Response patterns for this question were not as obvious as for many other questions (Fig. 3d). This question was particularly difficult for students, with a low dot product of 0.252 with the CC vector. In fact, only 28.1% of students chose the correct answer post-test, and it was not the most common response. The largest values of dot products to representative vectors were split between the RD, DD, and RR vectors with values of 0.717, 0.721, and 0.732, respectively. An instructor may view this data for comparison to other semesters to see if the concept can be covered in a different way or more in-depth. Particularly, the most common distractor answer may want to be explored further.

While questions 14 and 32 highlight the unique insights of this analysis, they also bring up potential limitations using representative vectors. It is important to note that these were the most difficult questions on the CSEM with correct answer rates of 28.1% and 35.3% on the post-test, respectively. One limitation from using representative vectors is that the ordering of item responses is important, and questions 32 and 14 both showcase the correct answer not being the most popular post-test answer. Furthermore, if two groups of students are to be compared purely using the representative vectors, then the definition of the most common distractor answer post-test is important. Another limitation is that the representative vectors do not always result in large dot product values. Question 14 in particular showcases a scenario where none of the repre-

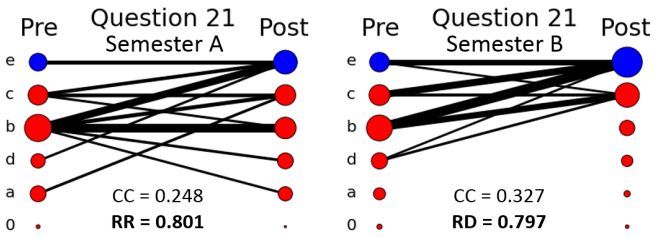

FIG. 4. Comparison of results for question 24 on the CSEM for two semesters, labeled Semester A and Semester B

sentative vectors have a very strong alignment to describe the data. One advantage of using representative vectors, however, is that we can design other representative vectors to describe the new situations and make the analysis flexible.

B. Semester Comparisons

A powerful use of the dot product and visualization methods is to make comparisons between semesters, institutions, demographics of students, and so on. The dot product between any two groups can be taken directly, where the value of the dot product quantifies how similarly the two groups respond to the question. This direct dot product does not have the limitations described above because no representative vectors need to be assumed and the order of the answers is always the same for both groups. Here we provide a comparison between two iterations of the same course.

Typically we found that the dot product between two semesters is well above 0.90 regardless of question, since most semesters behaved similarly on most items of the CSEM. However, there were some cases of differences. As an example, one of the largest differences is a dot product of only 0.80 between two semesters for question 21. We constructed the network-like graphs and report the relevant representative vector dot products (Fig. 4).

Results from Semester A show a spread of responses to roughly any answer post-test, whereas the instruction from Semester B seems to tend more strongly towards the correct answer and one distractor. This is also shown in the representative vectors where the largest dot product is the RR vector at 0.801 for Semester A and the RD vector at 0.797 for Semester B. All other dot products to representative vectors are at least well below 0.7 for both semesters.

This example illustrates that this method can be used to easily identify differences between semesters. It also demonstrates that instruction can play a major role in the response patterns of students at post-test, considering that Semesters A and B exhibit similar pre-test response patterns.

IV. CONCLUSIONS AND FUTURE WORK

Overall, the network-like graphs provide an efficient view of student performance, and a way to quantify response patterns to guide how the data are interpreted. Although a large amount of information is presented by a single graph, we believe it is easy to interpret and more accurate to show actual student response patterns rather than overall scores or gains.

A vector analysis approach using dot products to representative vectors is useful for quickly categorizing and quantifying the response patterns. This approach helps to guide analysis, and is effective for the majority of questions on the CSEM, as demonstrated by our examples. The utility of the vector analysis for smaller sample sizes, where random guessing more strongly influences the response patterns, may be addressed in future work.

The vector analysis was very effective for comparing the response patterns between semesters. This comparison avoids any possible limitations of the representative vectors or the ordering of the responses, since these are not needed for the comparison. The utility of this method is for addressing research questions comparing any two groups with pre- and post-test conditions. Comparisons of response patterns can be done by separately analyzing groups with different academic preparation, gender, or underrepresented minorities, for example. Our future work will expand upon these ideas, and we plan to broaden the comparison analysis to take into account differences in pre-test.

ACKNOWLEDGMENTS

We thank Professor Erich Mueller for collection and implementation of the CSEM. We thank Professors Tomás Arias, Kyle Shen, and all instructors of the electricty and magnetism course for work on reforming the courses and motivating this study. This work was supported in part by the Cornell University College of Arts and Sciences Active Learning Initiative.

[1] D. Hestenes, M. Wells, and G. Swackhamer, *Phys. Teach.* **30** (3),141-158 (1992).
[2] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, *Am. J. Phys.* **69**, S12 (2001).
[3] S. R. Singer, *J. Res. Sci. Teach.* **50**, 768 (2013).
[4] R. Hake, *Am. J. Phys.* **66**, 64 (1998).
[5] Jayson M. Nissen, et al. *Phys. Rev. Phys. Educ. Res.* **14**, 010115 (2018).
[6] James Day, et al. *Phys. Rev. Phys. Educ. Res.* **12**, 020104 (2016).
[7] Joshua Von Korff, et al. *Am. J. Phys.* **84**, 969 (2016).
[8] Lin Ding and Robert Beichner, *Phys. Rev. Phys. Educ. Res.* **5**, 020103 (2009).
[9] Karim Diff, and Nacira Tache, *AIP Conference Proceedings* **951**, 85 (2007).