Assessing class-wide consistency and randomness in responses to true or false questions administered online

Andrew Pawl,^{1,*} Raluca E. Teodorescu,² and Joseph D. Peterson¹

¹Engineering Physics Department, University of Wisconsin–Platteville, Platteville, Wisconsin 53818, USA

²Department of Physics, The George Washington University, Washington, D.C. 20052, USA

(Received 29 September 2012; published 2 July 2013)

We have developed simple data-mining algorithms to assess the consistency and the randomness of student responses to problems consisting of multiple true or false statements. In this paper we describe the algorithms and use them to analyze data from introductory physics courses. We investigate statements that emerge as outliers because the class has a preference for the incorrect answer and also those that emerge as outliers because the students are randomly changing their responses. These outliers are found to include several statements that are known in the literature to expose student misconceptions. Combining this fact with comments made by students and results of complementary assessments provides evidence that the tendency of a group of students to change their answer to a true or false statement or to remain consistent can serve as indicators of whether the class has understood the relevant concept. Our algorithms enable teachers to employ problems of the type described as a tool to identify specific aspects of a course that require improvement. They also enable researchers to employ such problems in experiments designed to probe aspects of students' thought processes and behavior. Additionally, our results demonstrate that at least one category of research-inspired problems (ranking tasks) can be adapted to the linked true or false format and productively used as an assessment tool in an online setting.

DOI: 10.1103/PhysRevSTPER.9.020102

PACS numbers: 01.40.Di, 01.40.Fk

I. INTRODUCTION

A. Background and motivation

A popular type of problem in the LON-CAPA [1] online homework network consists of a situation or set of situations followed by several related true or false statements [2]. An example is shown in Fig. 1. To discourage random guessing, the student is required to submit answers to all the true or false statements at once, and generally receives only correct or incorrect feedback. Thus, a student who enters a wrong answer will not know which of the statements has been answered incorrectly or even how many of the statements are incorrect. Because of these measures, a perfectly organized but extremely confused student might have to enter as many as 2^N responses before arriving at the correct answer to a problem containing N true or false statements.

An advantage of online administration is the opportunity to employ computer algorithms to examine the full pattern of student responses to the various statements that make up the problems. In this work, we develop algorithms to differentiate statements that exhibit basically consistent responses (predominantly true or predominantly false) from those that exhibit essentially randomly varying

responses. We show that this differentiation provides important benefits to the instructor. We find that statements which exhibit a consistent pattern of incorrect responses are an indicator either of an error in the problem or of a strongly held misconception, while statements exhibiting random responses could again indicate an error or could represent an incomplete understanding of the concept tested.

In the course of the analysis, we discover that ranking tasks [3,4] are readily adaptable to the problem format studied here and stand out as useful probes of student misconceptions. This suggests that ranking tasks can be administered in a manner that streamlines their evaluation and which is adaptable to fully online courses. In

You are pushing a filing cabinet across a rough floor ($\mu_k = 0.37$) in a straight line at a constant speed. Which of the following statements about the magnitudes of the forces acting on the filing cabinet are correct?

☐ the force that you exert on the filing cabinet will be more than the frictional force on the cabinet

☐ if you exerted twice the force, the cabinet would accelerate across the floor

☐ the force that you exert on the filing cabinet will be more than its weight

☐ the force that you exert on the filing cabinet will be less than its weight

☐ the force that you exert on the filing cabinet will be equal to the frictional force on the cabinet

Tries 0/99

FIG. 1. An example of the type of problem discussed in this paper. Each problem consists of five true-false statements.

*pawla@uwplatt.edu

traditional administration of such tasks, students assign a ranking to several situations and then give a written explanation for their choice and rate their confidence. In online administration, breaking the rankings up into multiple pairwise comparisons and then assessing randomness and consistency using the metrics we have developed can serve as a proxy for the combination of the written statement of reasoning and the confidence rating. We emphasize, however, that the mode of analysis described in this paper is intended to assess the conceptual understanding of an entire class rather than an individual student. Our approach is not well suited to grading a single student because it relies on statistical assumptions that are rendered invalid when examining individual response patterns.

After providing evidence for the validity of the metrics, we consider potential future applications. The metrics developed in this paper can help teachers to determine what topics need more attention in class. They offer researchers the opportunity to use linked true or false questions in experiments designed to probe student thought processes and behavior. To give specific illustrations of how the metrics might be used by teachers and researchers, we conclude the paper with a discussion of preliminary results from ongoing experiments.

B. Terminology

The discussion of problems of the type studied in this paper is complicated by the fact that each problem consists of several linked questions. In order to avoid ambiguity, we refer to the individual true or false questions that make up a problem as *statements*. We reserve the term *problem* to refer to the collection of five statements that must be submitted simultaneously for grading (see Fig. 1).

The example problem shown in Fig. 1 can also serve to illustrate two other ideas that will be explored in this paper. One of our goals is to develop a means to detect statements to which the class consistently responds incorrectly. We will show that such response patterns correlate with what we will term *strong misconceptions*. Our definition of strong misconception is an intuitive belief that is in conflict with the principles of physics. An example is the first statement from the problem shown in Fig. 1. Research has shown that students in introductory physics have a strong tendency to believe that there must be a net force in the direction of motion, even when this belief conflicts with Newton's first law [5–8]. Thus, we might expect the class to consistently answer “true” to the statement, even when forced to answer multiple times. Another complementary goal is to investigate whether significant class-wide randomness in the answers to a given statement can serve as an indicator of *incomplete understanding*. Again, the problem of Fig. 1 provides an illustration. Suppose that a significant portion of the class is indeed convinced that a net force is necessary to produce constant velocity. This could produce a conflict in the minds of the students about

the consequences of applying more force. Will the extra force cause an acceleration in accordance with Newton's second law? Will it result in a new (faster) constant velocity? Will there be a transient acceleration dropping to zero when the appropriate velocity is reached? Because of these conflicting ideas, we might expect students to exhibit a tendency to change their answer to the second statement shown in Fig. 1. Thus, our definition of incomplete understanding is a conflict in the minds of the students about the consequences of a specific physical principle in the situation presented. We will revisit these two definitions and return to the two statements from the problem of Fig. 1 as test cases in Sec. VB after we have developed quantitative metrics that assess consistency and randomness.

C. Sample

We administered student-driven (“choose your own path”), multilevel homework [9,10] containing problems of the type described above in two introductory mechanics courses during the 2010–2011 academic year. The overall sample consists of approximately 80 students enrolled in the off-semester calculus-based mechanics course at the Massachusetts Institute of Technology (MIT 8.011) and approximately 50 students in the first semester of a two-semester algebra-based introductory physics sequence at the University of Wisconsin–Platteville (UWP 1350).

Because the students were able to choose among several problems in completing their homework and because LON-CAPA allows for randomization of the statements that appear in a given problem (to discourage cheating), the total number of students who gave responses to any individual statement is generally substantially fewer than the total number of enrolled students. We have chosen not to analyze statements to which fewer than five students responded. The median number of respondents to the statements analyzed was 13 for the MIT sample and eight for the UWP sample. Further, to simplify the analysis, we have chosen to consider only problems that display exactly five true or false statements to each student. After these cuts we have approximately 250 statements to analyze for each of the courses, corresponding to about 30 problems for each course. (Note that although each problem studied in this paper presents exactly five statements to the student, many of the problems contain a library of more than five statements because of the randomization feature mentioned above, and so the total number of statements exceeds five times the number of problems.)

II. METHODOLOGY

A. Quantifying consistency

1. Submission-weighted consistency

Our first goal is to detect statements for which the class shows a marked preference for a particular answer. Since a true or false statement has only one correct response and

one incorrect response, one way to assess how consistently the class gives a specific response is to compute the fraction of submissions to the statement that are correct. A class which yields nearly 100% correct submissions for a statement is consistently giving the correct response. A class which yields nearly 0% correct submissions is *also* consistent, but incorrect (continuing to respond with the incorrect answer). A class which is answering randomly would be expected to approach 50% correct submissions as the number of tries becomes large. Thus, one measure of consistency for an entire class of students is to take the number of correct submissions to a given statement divided by the total number of submissions and subtract 0.5, given by

$$C_{sw} = \frac{N_{s,correct}}{N_{s,total}} - 0.5, \quad (1)$$

where C_{sw} is the submission-weighted consistency, $N_{s,correct}$ is the number of correct submissions to the statement, and $N_{s,total}$ is the total number of submissions to the statement.

With this definition, $C_{sw} = +0.5$ is complete correctness and $C_{sw} = -0.5$ is complete incorrectness. Again, we emphasize that we consider this metric unreliable on the individual student level. The average is taken over the full set of submissions by the class.

Taken alone, this submission-weighted measure of consistency is insufficient to characterize the consistency of the whole class. It can be strongly biased by a single student. Imagine a situation where five students respond to a given statement and four of the five immediately get the statement and also the entire corresponding problem correct. Suppose that the fifth individual, however, makes 8 incorrect submissions, each of which includes the wrong answer to the statement of interest. The total fraction of correct submissions is only $4/(8 + 4) = 0.33$ giving $C_{sw} = 0.33 - 0.5 = -0.17$ for this statement. This is a negative consistency even though 80% of the class got the statement and the corresponding problem correct. For this reason, it is important to define a complementary respondent-weighted consistency measure.

2. Respondent-weighted consistency

To gain a complementary measure of the consistency of a class on a particular statement, we examine the *first* submission of each respondent to the statement. We subtract the number of incorrect first submissions from the number of correct first submissions and divide this difference by the total number of students who responded, given by

$$C_{rw} = \frac{N_{s,i,correct} - N_{s,i,incorrect}}{N_{r,total}}, \quad (2)$$

where C_{rw} is the respondent-weighted consistency, $N_{s,i,correct}$ is the number of correct initial submissions,

$N_{s,i,incorrect}$ is the number of incorrect initial submissions, and $N_{r,total}$ is the total number of respondents.

With this definition, $C_{rw} = +1$ means that the entire portion of the class that submitted answers to the statement responded correctly on the first try, while $C_{rw} = -1$ means the entire portion of the class that submitted answers to the statement responded incorrectly on the first try. If $C_{rw} = 0$, then half the respondents gave a correct first submission and half gave an incorrect first submission.

3. Putting it together: Overall consistency

To detect statements which yield large values for *both* submission-weighted and respondent-weighted consistency, we define an overall consistency score C_{total} :

$$C_{total} = C_{sw} C_{rw} [\text{sgn}(C_{sw}) + \text{sgn}(C_{rw})], \quad (3)$$

where C_{sw} is the submission-weighted consistency and C_{rw} is the respondent-weighted consistency. The allowed range is $-1 \leq C_{total} \leq 1$. The sum in brackets ensures that the formula remains negative if both C_{sw} and C_{rw} are negative, thus preserving the ability to distinguish strong *incorrect* consistency from strong *correct* consistency. Additionally, the formula yields zero if the signs of the two component scores do not agree, which prevents potential false signals of strong positive or negative consistency.

B. Quantifying randomness

1. Submission-weighted randomness

With a consistency score defined, it is tempting to assume that $C_{total} \approx 0$ can be taken to indicate random submissions. We must, however, attempt to distinguish true randomness from false signals. If, for example, a large number of students submitted a string of incorrect responses followed by an equal string of correct responses, the overall fraction of correct submissions could be approximately 50%, but we would not call this random switching. Our goal is to find statements which give indications of approximately equal preference for the two options. In keeping with this goal, our first method of assessing randomness is to check for constant switching of the response by calculating the fraction of possible switches that are realized. A switch is “realized” if the current submission is different from the one immediately prior, given by

$$R_{sw} = \frac{N_{s,switch}}{N_{s,total} - N_{r,total}}, \quad (4)$$

where R_{sw} is the submission-weighted randomness score, $N_{s,switch}$ is the number of submissions that represent a switch from the immediate predecessor, $N_{s,total}$ is the total number of submissions made by the class to the statement, and $N_{r,total}$ is the total number of students who responded to the statement. The denominator represents the number of potential switches. The number of respondents must be

subtracted from the number of submissions to determine the number of potential switches because it is impossible to switch your answer on the first submission.

We call this measure the submission-weighted randomness because, just as with the submission-weighted consistency score, it can be strongly biased by a single individual who makes substantially more submissions than his or her classmates to a certain statement. Unlike the consistency scores, however, the submission-weighted randomness score for a given statement cannot be negative. The allowed range is $0 \leq R_{sw} \leq 1$, with 0 indicating that no switches were made to the responses and 1 indicating that all possible switches were made.

2. Respondent-weighted randomness

To define a respondent-weighted measure of randomness for a given statement, we determine what fraction of the students who responded to the statement ever switched their response from correct to incorrect, given by

$$R_{rw} = \frac{N_{r, \text{correct} \rightarrow \text{incorrect}}}{N_{r, \text{total}}}, \quad (5)$$

where R_{rw} is the respondent-weighted randomness score and $N_{r, \text{correct} \rightarrow \text{incorrect}}$ is the number of respondents who ever switched their response to the statement from correct to incorrect.

The allowed range is $0 \leq R_{rw} \leq 1$ with 0 indicating that no student ever changed their response to incorrect after giving a correct response and 1 indicating that all students who responded to the question shifted from a correct to an incorrect response at some point during their interaction with the statement.

3. Putting it together: Overall randomness

Just as for the consistency scores, we can define an overall randomness score for a given statement by multiplying the submission-weighted and respondent-weighted scores. This time, both scores are inherently positive. Thus, we can define the overall randomness (R_{total}) in terms of the submission-weighted randomness (R_{sw}) and respondent-weighted randomness (R_{rw}) as

$$R_{\text{total}} = R_{sw} R_{rw}, \quad (6)$$

which gives the allowed range $0 \leq R_{\text{total}} \leq 1$.

III. CHARACTERIZING THE SAMPLE

Our sample is made up of two distinct populations located at different universities and instructed in different curricula by different instructors. Thus, before we assess the validity of our overall consistency and randomness measures, we characterize the reliability by investigating the variation of these scores between the subsamples. We

FIG. 2. Cumulative probability distribution of the overall consistency score [C_{total} defined in Eq. (3)] by subsample. The total number of statements used to generate the distributions are 250 for UWP 1350 and 257 for MIT 8.011.

do this by examining cumulative probability distributions for the two overall scores. The relevant distributions are shown as Figs. 2 and 3.

The two populations yield significantly different distributions for both scores ($p < 10^{-8}$). In the consistency distribution (Fig. 2), the MIT class is considerably more likely to yield a high positive consistency, but considerably less likely to yield a strong negative consistency. In the randomness distribution (Fig. 3), the MIT class is far less

FIG. 3. Cumulative probability distribution of the overall randomness score [R_{total} defined in Eq. (6)] by subsample. The total number of statements used to generate the distributions are 250 for UWP 1350 and 257 for MIT 8.011.

likely to produce high scores. The performance of the MIT students on the questions examined in this study is significantly better than that of the UW–Platteville students according to our metrics. We note, however, that the questions were administered in a slightly different manner to the two populations. The MIT students were given only seven tries while the UWP students had 99 tries (the maximum allowed in LON-CAPA). Further, the MIT students were assigned to attach comments to a certain number of problems each week. These comments were visible to other students and sometimes contained hints about the correct answers. It is possible that these differences contributed to the disparity observed between the MIT and UWP distributions. We will explicitly consider the impact of the comments in Sec. VI.

The cumulative probability distributions of Figs. 2 and 3 allow us to determine cuts in the scores that characterize a strong negative consistency (potentially indicative of a strongly held misconception) and strong randomness (potentially indicative of incomplete understanding) on a given statement. The consistency score distribution (Fig. 2) for the UW–Platteville course exhibits a bend at a consistency score of about -0.08 which corresponds to a cumulative probability of about 5%. This is also the standard probability threshold for significance, corresponding approximately to a 2 standard deviation cut in a normal distribution. Thus, we will take a consistency score of -0.08 to mark the boundary of the outliers in consistency. Since we are interested in high randomness scores rather than low, the equivalent 5% outlier cut for the randomness distribution (Fig. 3) is to determine the 95% cumulative probability randomness score. For the UW–Platteville sample, the 95% cumulative probability occurs at a randomness score of 0.23. Thus, our first estimate is that a consistency score less than -0.08 represents strong negative consistency and a randomness score greater than 0.23 indicates strong randomness. If these cuts are applied to the MIT sample, then only two statements (0.8% of the total) qualify as outliers in negative consistency score and another two as outliers in randomness.

IV. CONSISTENCY VS RANDOMNESS

Another description of the data that can add insight is a scatter plot of the overall consistency score versus the overall randomness score for each of the statements in the sample. Figures 4 and 5 give these plots for the different populations of students. Several statements have been outlined with symbols. These groupings will be discussed in the following section as examples to illustrate the type of information that can be gained by examining statements that emerge as outliers in consistency or randomness according to our metrics.

FIG. 4. Overall randomness [R_{total} defined in Eq. (6)] vs overall consistency [C_{total} defined in Eq. (3)] for the 250 statements in the UWP 1350 subsample. The dashed lines indicate the strong randomness cut of 0.23 and the strong (incorrect) consistency cut of -0.08 discussed in Sec. III. Points outlined with symbols are discussed in Sec. V.

FIG. 5. Overall randomness [R_{total} defined in Eq. (6)] vs overall consistency [C_{total} defined in Eq. (3)] for the 257 statements in the MIT 8.011 subsample. The dashed lines indicate the strong randomness cut of 0.23 and the strong (incorrect) consistency cut of -0.08 discussed in Sec. III. Points outlined with symbols are discussed in Sec. V.

V. EVIDENCE FOR VALIDITY

A. Squares: Sufficient but not necessary

Points outlined with squares appear only in the UW–Platteville sample (Fig. 4). These squares represent 14 statements associated with two problems for which the figures *did not appear* to the students due to a file permission error. Because the format of the homework allowed the students to earn full points without completing every problem, this error was not reported to the instructor. Although the figures are absolutely necessary to correctly answer the statements without guessing, some students chose to attempt the problem anyway. These statements are interesting because they clearly represent random guessing. We have highlighted all the relevant statements because they illustrate two important aspects of our analysis.

First, the randomness measure can serve as an indicator of errors in problem construction. Six of the statements for one of the two problems exceed the strong randomness threshold, which is a sign that something is wrong.

Second, the question format studied in this paper likely makes it impossible to develop a measure that is an infallible indicator of random guessing (i.e., a *necessary* condition). When “randomly” changing the answer to five linked statements, students will tend to follow some internally preferred pattern, which can give the appearance of consistency on some statements. (For example, some students might prefer to explore all the possibilities that involve answering “true” for the first statement before trying any of the possibilities that answer “false” to the first statement.) Hence, several of the statements for these two problems appear safely in the densely populated region of the consistency versus randomness plot.

B. Arrows: Revisiting the hypothesis

In the Introduction we discussed the example problem shown in Fig. 1. We hypothesized that the first statement shown in the problem would qualify as a strong misconception because research has demonstrated that students tend to expect a net force in the direction of motion [5–8] and that the second statement would exhibit significant randomness for the same reason. The two points marked by arrows in the UWP 1350 subsample correspond to these statements. In agreement with the hypothesis, the first statement has a strongly negative consistency score (-0.33) while the second statement has a strong randomness score (0.25). In this case, we can also attempt to validate the conclusions in a more specific way.

The Force Concept Inventory (FCI) [11] is used as a pre- and postinstruction assessment in the UWP 1350 course. The situation described in Fig. 1 is essentially identical to the situation of problems 25–27 on the FCI, except that the FCI replaces the filing cabinet with a “large box.”

Examining the answers given by the class to these questions can lead to a deeper understanding of their responses to the statements shown in Fig. 1.

Question 25 of the FCI asks about the force applied by the person pushing the box, which is analogous to the first statement of the example of Fig. 1. Two of the five answer choices for FCI question 25 indicate that the pushing force must be greater than the “total force which resists” the motion of the box. Of the 41 students in the sample who took the FCI as both a pre- and postassessment, 28 (68%) indicated the pushing force must exceed the resistive force on the pretest and 31 (76%) did so on the posttest. Of the 28 students who chose these options on the pretest, 21 (75%) chose one of these options again on the posttest. This is certainly indicative of a strong misconception within the class that is not being resolved by instruction.

Question 26 of the FCI asks about the result of doubling the pushing force acting on the box, which is analogous to the second statement of the example of Fig. 1. Two of the five statements indicate that the box will then move “with a constant speed” that is greater than the original speed. The other three statements all make some mention of “increasing” speed. Interestingly, the students do not seem at all random in their approach to this question on the FCI. Of the 41 students who took both the pre- and postinstruction FCI, 32 (78%) selected one of the constant speed responses on the pretest and 31 (76%) did so on the posttest. Of the 32 students who picked a constant speed option on the pretest, 26 (81%) again chose one of them on the posttest.

This leads us to consider whether the second statement in the example of Fig. 1 should have emerged as a strong misconception instead of strongly random. Six students in our sample submitted answers to the statement. Comparing their responses to the true or false statement with their answer to question 26 on the FCI shows that the randomness in their responses to the second statement of Fig. 1 is very likely a real effect due to incomplete understanding of the technical meaning of “acceleration.” All 6 of the students initially responded that the filing cabinet “would accelerate across the floor” if you exerted twice the force. By contrast, all 9 responses from these 6 students to question 26 on the FCI (5 pre- and 4 postinstruction) indicated a discontinuous change in velocity as a result of the doubled force. In fact, of the 41 students with matched pre- and postinstruction FCI data, only 7 pretest respondents (17%) and 6 posttest respondents (15%) answered that the box would move “for a while with an increasing speed, then with a constant speed thereafter” when the pushing force is increased. With this added information, the randomness of the responses to the second statement in Fig. 1 suggests that the students are unsure whether a discontinuous increase in velocity constitutes acceleration or not. The FCI avoids this randomness by the use of the term “increasing velocity” in place of acceleration.

Thus, the hypothesis put forward in the Introduction that randomness on the second statement of the problem shown in Fig. 1 would arise from conflict between the Newtonian concept that unbalanced force produces acceleration and the intuitive concept that an unbalanced force is needed to produce motion was incorrect. By a wide margin the students believe that the unbalanced force will produce a constant speed and that the amount of the imbalance is related to the size of the speed. Instead, the observed randomness appears to stem from an incomplete understanding of the fact that in physics the term acceleration implies a *continuous* change in velocity.

C. Open circles: Known misconceptions

Points outlined with an open circle occur in either the consistently wrong or the significantly random areas of both subsamples (Figs. 4 and 5). The circled statements all belong to a single problem that shows several drawings of masses supported by ropes (see Fig. 6), which is a variant of a ranking task [3] adapted from Ref. [4]. The statements involve assertions about the tension in the ropes labeled R in each picture. The statement, “The tension in the rope in picture [A] is 80 N,” is true, but students tended to respond that it was false. This statement qualified as an extremely strong misconception (consistency score of -0.67) in the UW–Platteville course.

It might be expected that the students assumed the tension in picture A was 160 N, but the statements, “The tension in the rope in picture [B] is 80 N” and “The tension in the rope is greater in picture [A] than in picture [C],” exhibited strong randomness (0.25 and 0.24, respectively) rather than emerging as a strong misconception (which would have yielded significant negative consistency scores). A clue to why these statements elicit random responses is provided by the fact that the statement, “The tension in the rope is equal in pictures [A] and [B],” also yielded a strong randomness score (0.23). This suggests that the students are entertaining the possibility that the tensions are zero for pictures A and B. Further support for this supposition is provided by the MIT subsample, whose responses exhibit strong randomness for the statement, “The tension in the rope is zero in pictures [A] and [B].” The signal seen in the MIT data is corroborated by student comments. The MIT course required students to periodically post comments as part of their grade. This

FIG. 6. Pictures from the problem labeled with circles (○) in Figs. 4 and 5. The problem is adapted from Ref. [4].

problem was confusing enough to prompt several posts, including: “I ran out of tries for this question, but I’m still confused about the diagrams for [A] and [B]. Is the tension for those two zero?”

Taken together, the pattern of responses to these questions indicates that the UWP 1350 class holds a strong misconception that the tension in the rope for a stationary Atwood’s machine is *not* equal to the weight on either side and that there is an associated confusion present in both the UWP and MIT subsamples regarding whether the tension would be the sum of the weights or would instead be zero. The fact that students are disposed to believe that the tension is the sum of the weights or zero in such a case has been studied previously and published in the physics education research literature [12]. The fact that we recover known misconceptions is one confirmation of the validity of our approach. In this case, additional confirmation of validity is found in the fact that the comments of the students corroborate the conclusions drawn from the data.

D. Up-pointing triangles: More ranking tasks

Points outlined with upward-pointing triangles encompass the strongest randomness (overall randomness score of 0.37 in the UWP subsample) for the entire pool of statements we studied. These points represent statements from two related problems involving projectile motion, which are again adaptations of ranking tasks. Each of the two problems presents the students with six pictures of cannons firing (see Figs. 7 and 8). In one of the problems, the initial velocities of the six cannonballs have the same

FIG. 7. Pictures from one of the problems labeled with upward-pointing triangles ($\triangle$) in Fig. 4. The horizontal component of the initial velocity is always 20 m/s in this problem. The problem is adapted from Ref. [4].

FIG. 8. Pictures from one of the problems labeled with upward-pointing triangles ($\triangle$) in Fig. 4. The vertical component of the initial velocity is always 20 m/s in this problem. The problem is adapted from Ref. [4].

horizontal component. In the other problem, the vertical components are identical.

Looking at the pattern of strong misconceptions and strong randomness points to the conclusion that the UWP students tend to believe the angle is the only important factor determining the range and the maximum height of the cannonball (they are failing to account for the variation in overall initial velocity). This conclusion is supported by the fact that the statements, “The highest vertical height reached [among the shots in Fig. 8] is in [picture C]” (false), “The highest vertical height reached [among the shots in Fig. 8] is the same in all cases” (true), and “The longest horizontal distance traveled [among the shots in Fig. 7] is in [picture C]” (true), all score as strong misconceptions, meaning that the students tend to give the incorrect answer.

Perhaps more interestingly, there is also an indication that the UWP students do not understand the direct correspondence between maximum height reached and time of flight. The statement exhibiting strongest randomness is, “The time in the air [for the shots of Fig. 8] is greater in [B] than it is in [C].” The misconceptions discussed above, however, indicate that the students are strongly disposed to believe shot C reaches the highest vertical height. If the students understood the correspondence between height and time of flight, they would confidently give the correct answer (false) to this statement, though for the wrong reasons. The fact that this statement exhibits randomness instead of consistency could be an indication of a conflict in the students’ minds between assigning a longer time of flight to the shot that goes higher (which their misconceptions indicate they believe to be shot C) versus the shot that has a longer horizontal range (which their misconceptions indicate they believe to be shot B). This conflict has been previously remarked by physics education researchers [13].

E. MIT subsample: different criteria

Comparing Figs. 4 and 5 reinforces the conclusion drawn from the cumulative probability distributions of Figs. 2 and 3 that the MIT students perform better than the UWP students on the homework problems in our sample. Because of this, we can identify as outliers in the MIT subsample statements that would be lost in the noise for the UWP subsample. For example, the five statements with randomness scores greater than 0.19 but less than 0.23 in Fig. 5 seem to constitute at least marginal outliers. The fact that these marginal outliers contain statements from only three problems and that each of these problems also contains a statement that exhibits either unambiguous strong randomness or strong incorrect consistency is one indicator that we should take their outlier status seriously.

The statements marked with right-pointing triangles belong to a projectile motion problem. Namely, a battleship is shooting two shells with identical initial speed to hit

two targets which have different horizontal range from the battleship (the shells must therefore have different launch angles). The shells hit the target at the same height that they were shot from. The statement that qualifies as a strong outlier in negative consistency is, “[The farther ship] gets hit first.” which is true because the trajectory is drawn with a lower maximum height. The statement that qualifies as a marginal outlier in randomness is, “[The closer ship] gets hit with a greater impact speed,” which is false because the identical launch speeds imply identical impact speeds because the launch speed equals the impact speed if the parabolic trajectory is symmetric. The student discussion posts for this problem include “I honestly couldn’t get this question, my friend and I used all of our turns together, and we didn’t figure it out. Could someone clarify what the impact speed is?” This comment implies that we should take the outlier status of the statement seriously.

The diamonds, which dominate the marginal outliers, represent statements from a problem involving a super ball and a lump of putty thrown against a door (the super ball rebounds and the putty sticks). The problem compares the evolution of the momentum and kinetic energy of the two objects as they hit the door. The statement that exhibits the strongest randomness is, “The impulse imparted by the door to the putty is greater than the impulse imparted to the rubber ball” and the others are similar. The physics education research literature has considered a situation analogous to this one and remarked student difficulty with this concept [14], and so again we find reason to take the outlier status seriously.

The marginal outlier marked with a circle belongs to the Atwood machine ranking task discussed in Sec. V C. This statement indicates that the tension in picture B of Fig. 6 is 80 N. We have already established that this is a source of confusion.

This analysis of the MIT subsample leads us to conclude that the cuts employed to identify strong randomness and strong misconceptions must be sample dependent.

VI. APPLICATIONS

In this section we provide examples of how the metrics developed in this paper might be used by teachers and researchers. To add detail to this discussion, we provide preliminary results of several studies that we have begun since developing the metrics. These results are presented to give examples of possible future directions and to pose questions that may be taken up by the reader. They should not be construed as complete, publishable research outcomes in their own right.

A. For teachers: Assessing pedagogy

One of us (A.P.) was the sole instructor for the UWP 1350 class that constitutes part of the sample. Prior to undertaking this study, he was familiar with a large body

of education research on the difficulties students have in learning forces (e.g., Refs. [5–8]), but was unaware of any systematic studies of student difficulties in understanding projectile motion. Thus, he was surprised that the statements about projectiles discussed in Sec. VD comprised such a significant fraction of the outliers in the UWP 1350 sample. One natural question to ask is whether similar results would be observed in his other courses. He therefore administered problems of the format described to two other courses: UWP 1050, an algebra-based “principles of physics” course primarily aimed at students in the industrial technology program, and UWP 2240, a calculus-based mechanics course primarily aimed at engineering students.

In the UWP 1050 course three of the 10 strongest negative consistency scores were on statements from the cannons problems of Figs. 7 and 8. The strongest negative consistency for this sample was on the statement “The longest horizontal distance traveled [among the shots in Fig. 7] is in [picture C]” (true), which yielded $C_{\text{total}} = -0.84$.

In the UWP 2240 course, once again three of the 10 strongest negative consistency scores were on projectile problems. The second strongest was on the statement “The highest vertical height reached [among the shots in Fig. 8] is in [picture C]” (false), with a score of $C_{\text{total}} = -0.18$.

These outcomes strongly suggest that the instructor should revise his approach to teaching projectile motion. The pedagogy employed currently is not adequately teaching the fundamental concepts.

B. For researchers: Investigating student behavior

In Sec. III it was mentioned that the two subsamples studied in the main body of this paper were subject to different constraints in working the problems. Two differences stand out in particular. First, the MIT students had a limited number of tries on each problem (seven) while the UWP students had an essentially unlimited number (99). Second, the MIT students were frequently able to read student comments that related to the problems while solving them because they were assigned to periodically post comments as a graded activity, while the UWP students had essentially no comments to help them. Each of these differences could be adapted into a physics education research project.

To examine the dependence of student behavior on availability of comments from other students, we have administered problems of the type studied in this paper to two separate UWP 2240 classes, each with an enrollment of approximately 50 students. In one of those classes posts were not required of the students and so essentially no comments were posted. In the other class one post was required per week. Both classes were allowed 99 tries to complete each problem. The consistency score distributions for these separate UWP 2240 populations are shown in Fig. 9. The distributions indicate that the presence of

FIG. 9. Cumulative probability distribution of the overall consistency score [C_{total} defined in Eq. (3)] for two distinct populations of students enrolled in UWP 2240. One population was not required to post comments to the problems while the other was required to post one comment per week. An identical set of 136 statements was used to create the distribution for each population.

student posts does not dramatically affect the frequency of confidence scores. The Kolmogorov-Smirnov (KS) test indicates no significant difference between the distributions ($p = 0.50$). The corresponding randomness score distributions shown in Fig. 10, by contrast, clearly show a difference by eye. The KS test indicates that this difference is significant at the $p = 0.003$ level. The student comments dramatically reduce the frequency of strong randomness scores. This implies that students are unlikely to search the posts for help on a statement about which they feel confident, but they are likely to seek answers for statements about which they are unsure.

FIG. 10. Cumulative probability distribution of the overall randomness score [R_{total} defined in Eq. (6)] for two distinct populations of students enrolled in UWP 2240. One population was not required to post comments to the problems while the other was required to post one comment per week. An identical set of 136 statements was used to create the distribution for each population.

FIG. 11. The eight projectile pairs used for the linked true or false problem described in the text. The figures were adapted from Ref. [13].

C. For researchers: Insight into student thought processes

Once a researcher has exposed a widespread student misconception, it is possible to refine the true or false question format to explore the thought process of the students in more detail. We have discussed the fact that the dependence of projectile time of flight on the height of the trajectory tends to be incompletely understood by the students enrolled in one of the author's (A. P.) courses. We might surmise that students would believe that the higher trajectory corresponds to a longer time of flight for cases where the ranges of the two trajectories are comparable, but that they would believe the longer-range trajectory corresponds to a longer time of flight for cases where the heights of the two trajectories are comparable. An open question for research is what level of disparity is needed to produce confident responses from the students. Is there a clear division that separates the "higher is more time" response from the "longer range is more time" response, or is it a fuzzy boundary? To investigate this we can design what might be termed a modified ranking task. Consider the pictures in Fig. 11, which are adapted from Ref. [13]. Instead of presenting the students with a large number of individual situations to rank in series, it is possible to present the students with a large number of separate pairwise ranking tasks. In doing so, we can deeply probe one specific thought process that contributes to the ranking.

The images of Fig. 11 were used to construct a five-statement linked true or false problem that was administered to two UWP 2240 courses in order to determine the nature of the transition discussed above. For each of the eight projectile pairs, the students could potentially be asked to respond to one of three statements: "[the higher trajectory] will have a longer time of flight than [the lower trajectory]" (true), "[the lower trajectory] will have a longer time of flight than [the higher trajectory]" (false), or "[both trajectories] will have equal times of flight"

(false). The five statements presented to each student were randomly selected from this pool of 24 possible statements with the constraint that an individual student could not receive more than one statement about a single pair of trajectories. In Fig. 12 we plot R_{total} vs C_{total} for the statements that indicate the higher trajectory will have a longer time of flight for each of the eight pairs. In Fig. 13 we plot R_{total} vs C_{total} for the statements that indicate the time of flight is equal for each of the eight pairs. Examining these figures, we can see that there is some threshold of disparity that shifts the class to confident responses. Pairs 5 and 7 apparently exceed this threshold in the direction of correct responses and pair 2 exceeds the threshold in the direction of incorrect responses. The remainder of the pairs, however, fall into a fuzzy area where the class as a whole does not give a consistent response. We note that one potentially confounding complication in this study is the

FIG. 12. Overall randomness versus overall consistency for eight statements of the form "[the higher trajectory] will have a longer time of flight than [the lower trajectory]" that correspond to the eight projectile pairs shown in Fig. 11. This statement should be answered "true" in each case.

FIG. 13. Overall randomness versus overall consistency for eight statements of the form “[both trajectories] will have equal times of flight” that correspond to the eight projectile pairs shown in Fig. 11. This statement should be answered “false” in each case.

fact that the data were gathered from UWP 2240 courses that required the students to post comments. As discussed above, it is likely that the availability of student comments alters the response patterns to the statements.

VII. CONCLUSIONS

As online homework becomes the norm and online instruction grows in popularity, it is useful for the physics education research community to develop methods of assessing conceptual understanding that exploit the power and simultaneously respect the constraints of the online environment. Such methods will be needed to allow a direct comparison of the efficacy of online-only instruction to traditional instruction and to measure the impact of efforts to improve the online experience for students.

We have developed the data-mining algorithms summarized in Table I for assessing the consistency and the randomness of student responses to individual true or false statements. These algorithms are directly applicable to problems involving several linked true or false statements,

which have been implemented in online homework. Investigation of examples indicates that the consistency score can uncover misconceptions and the randomness score can be a useful indicator of incomplete understanding. Both scores can also serve to reveal errors in problem construction.

The power of our approach for uncovering student difficulties is suggested by the fact that the conspicuous outliers reflect several misconceptions that are known to the physics education research literature. Additional corroboration is provided by student comments and class performance on the FCI.

The promise of the approach is that a question format which is suitable for use in online homework or online courses can uncover the specific concepts that give a significant portion of the class problems. This offers an advantage over pen-and-paper administration of any problem that can be adapted to the format studied in this paper. In this work, we have considered ranking tasks as an example. By coding the various relationships into true or false statements, we can easily recover detailed class-wide trends without the laborious correlation of sequences of inequalities with essay-format explanations and statements of confidence necessitated by pen-and-paper administration. We reiterate, however, that the statistical approach outlined in this paper is invalid when applied to the responses of an individual student. Thus, pen-and-paper administration must be used if the primary goal is the summative assessment of an individual.

Our algorithms also allow for the use of linked true or false questions in designed physics education research experiments including the investigation of student behavior and thought processes.

We have developed a macro for Microsoft Excel that implements the algorithms described in this paper for data recovered from the LON-CAPA “student submission reports” feature. Teachers and researchers who wish to use the algorithms are encouraged to contact the corresponding author (A. P.).

TABLE I. Summary of algorithms for computing class-wide consistency (C) and randomness (R) in responses to true or false statements. In all cases the subscript s denotes submission and the subscript r denotes respondent. In the formula for C_{rw} the subscript i indicates the initial submission, in the formula for R_{sw} the subscript switch indicates a submission which is the opposite response from its immediate predecessor, and in the formula for R_{rw} the subscript correct $\rightarrow$ incorrect indicates the number of respondents who ever switched their answer to the statement from correct to incorrect.

Desired quantity	Formula	Range of allowed values
Submission-weighted consistency	$C_{sw} = N_{s,\text{correct}}/N_{s,\text{total}} - 0.5$	$-0.5 \leq C_{sw} \leq 0.5$
Respondent-weighted consistency	$C_{rw} = (N_{s,i,\text{correct}} - N_{s,i,\text{incorrect}})/N_{r,\text{total}}$	$-1 \leq C_{rw} \leq 1$
Overall consistency	$C_{\text{total}} = C_{sw} C_{rw} [\text{sgn}(C_{sw}) + \text{sgn}(C_{rw})]$	$-1 \leq C_{\text{total}} \leq 1$
Submission-weighted randomness	$R_{sw} = N_{s,\text{switch}}/(N_{s,\text{total}} - N_{r,\text{total}})$	$0 \leq R_{sw} \leq 1$
Respondent-weighted randomness	$R_{rw} = N_{r,\text{correct} \rightarrow \text{incorrect}}/N_{r,\text{total}}$	$0 \leq R_{rw} \leq 1$
Overall randomness	$R_{\text{total}} = R_{sw} R_{rw}$	$0 \leq R_{\text{total}} \leq 1$

ACKNOWLEDGMENTS

This paper relies on work done by A.P. and R.E.T. as part of D.E. Pritchard's RELATE group at MIT. We particularly acknowledge the contributions of A. Barrantes, C.N. Cardamone, S. Rayyan, and D.T. Seaton. We thank G. Kortemeyer from the LON-CAPA team for advice that was instrumental to the creation of the homework and the

recovery of the data for analysis. We acknowledge G. Feldman's inspired idea of adapting ranking tasks to the linked true or false format. We thank J.T. Lavery for suggestions which materially improved this paper. This work was funded in part by a Scholarly Activity Improvement Fund grant from the University of Wisconsin-Platteville.

-
- [1] <http://lon-capa.org>.
 - [2] E. Kashy, S.J. Gaff, N.H. Pawley, W.L. Stretch, S.L. Wolfe, D.J. Morrissey, and Y. Tsai, Conceptual questions in computer-assisted assignments, *Am. J. Phys.* **63**, 1000 (1995).
 - [3] D.P. Maloney, Ranking tasks: A new type of test item, *J. Coll. Sci. Teach.* **16**, 510 (1987).
 - [4] T.L. O'Kuma, D.P. Maloney and C.J. Hieggelke, *Ranking Task Exercises in Physics* (Prentice Hall, Upper Saddle River, NJ, 2000).
 - [5] J. Clement, Students' preconceptions in introductory mechanics, *Am. J. Phys.* **50**, 66 (1982).
 - [6] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
 - [7] R.K. Thornton and D.R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, *Am. J. Phys.* **66**, 338 (1998).
 - [8] R. Rosenblatt and A.F. Heckler, Systematic study of student understanding of the relationships between the directions of force, velocity, and acceleration in one dimension, *Phys. Rev. ST Phys. Educ. Res.* **7**, 020112 (2011).
 - [9] R.E. Teodorescu, A. Pawl, S. Rayyan, A. Barrantes, and D.E. Pritchard, Toward an integrated online learning environment, *AIP Conf. Proc.* **1289**, 321 (2010).
 - [10] R.E. Teodorescu, D.T. Seaton, C.N. Cardamone, S. Rayyan, J.E. Abbott, A. Barrantes, A. Pawl, and D.E. Pritchard, When students can choose easy, medium, or hard homework problems, *AIP Conf. Proc.* **1413**, 81 (2012).
 - [11] The question numbers refer to the reformed FCI which is available via <http://modeling.asu.edu/R&E/Research.html>.
 - [12] L.C. McDermott, P.S. Shaffer, and M.D. Somers, Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine, *Am. J. Phys.* **62**, 46 (1994).
 - [13] T.M. Scaife, A Study on Contingency Learning in Introductory Physics Concepts, Ph.D. thesis, The Ohio State University, 2010).
 - [14] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).