Mining Student Misconceptions from Pre- and Post-Test Data

Ángel Pérez-Lemonche
Universidad Autonoma de Madrid
Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain
angel.perezl@estudiante.uam.es

Byron Coffin Drury
MIT
77 Massachusetts Avenue
Cambridge, MA 02139, USA
bdrury@mit.edu

David Pritchard
MIT
77 Massachusetts Avenue
Cambridge, MA 02139, USA
dpritch@mit.edu

ABSTRACT

We analyze results from paired pre- and post-instruction administration of the Mechanics Baseline Test to 2238 students in introductory mechanics classes. We investigate pairs of specific wrong answers given with unusual frequency by students on the pretest. We also identify transitions between pre- and post-test answers on the same question which elucidate student learning due to instruction. We define criteria for excess transitions above a random response model. Some common transitions are found to be associated specifically with students within a particular range of skills. Further, transitions from pre- to post-test revealed that incorrect pretest answers that were frequently repeated on the post-test often correspond to known misconceptions from physics or math. Thus, our data mining techniques can elucidate common student misunderstandings of mechanics concepts and how instruction affects these misunderstandings. This opens the way for finding improved interventions for specific misunderstandings revealed by analyzing results from pre- and post conceptual tests.

Keywords

Pre- and post- Testing; Common Student Misconceptions; Educational Data Mining; Analyzing Wrong Answers.

1. INTRODUCTION

The Force Concept Inventory [9] by Hestenes group revolutionized physics instruction by showing that students trained mostly on end-of-chapter problems in standard textbooks did not learn to answer easy (so teachers thought) questions based on fundamental concepts in the domain. This has led to tremendous reform of physics instruction worldwide and a series of concept tests covering introductory physics and astronomy [7]. The present study uses another research-based assessment, the Mechanics Baseline Test ("MBT"). The MBT is designed for students with more physics background and is appropriate for introductory students at MIT.

Research-based assessments such as concept inventories and surveys are typically developed by first administering the questions in open response format. Analysis often reveals clusters of related responses which are then made into distractors in a multiple-choice version of the assessment. Since these assessments typically center only a particular subdomain, e.g. force and motion, a part of Newtonian mechanics, it is expected that common misconceptions (also called alternate conceptions and misunderstandings) will manifest as correlated selections of distractors to different questions. We searched for these, as well as for statistically significant deviations of specific learning transitions from a random guessing hypothesis.

This paper addresses several questions relative to the deep assessment of students' knowledge structure based on results on the Mechanics Baseline Test. Our objective is to find the 'atomic' student conceptions and abilities that underlie their answers to the questions (possibly incorrectly)? Our approach is data mining on a large sample of pre and post-tests, and concentrates on these research questions

- Are there pairs of wrong answers to different questions that reveal common misunderstandings?
- Are there exceptionally prevalent transitions from pre- to post-test that seem to indicate learning some specific knowledge?
- Can we suggest new questions or improvements to existing ones that will improve the assessment?

We are not the first to attempt to extract actionable analysis from concept tests. Indeed, the FCI has been analyzed using factor analysis [4]; however, that analysis has been questioned [5]. The MBT has been refined using Item Response Theory analysis [3]. Recently Brewe et al. [2] have applied Network analysis to the FCI to predict post scores. The Colorado Learning Attitudes about Science Survey [1] has a nice web-based multicategory analysis based on factor analysis that is used. But it's fair to say that most concept tests are not analyzed beyond the score and whether it seems appropriate for each particular class based on quality of students & instructional style [8]. This provides a good characterization of the students' (and class) overall knowledge and gives a useful indication of the amount of learning if the assessment is administered both pre- and post-instruction. Unfortunately, such one-dimensional analysis ignores the category-specific information that the method of construction of these assessments would seem to generate. Therefore, administering these assessments neither informs the student about which concept(s) they know well or poorly nor informs the teacher about the areas in which they most need to improve their instruction.

The goal of finding specific difficulties and misconceptions of students continues to appear reasonable yet remains tantalizingly out of reach. The progress made here shows the promise of analysis of learning data at scale. But while our findings are clearly revelatory, they beg for further development to make them useful. We discuss ways of closing this gap in the last section: Future.

Table 1: The students in our dataset represent five years of an introductory mechanics course at MIT. Since some students lack either a pre- or post-test score, we have calculated grades and normalized gain using only those students who took both tests. The pretest was administered at the beginning of the semester, and the post-test was administered – often as part of the final exam - at the end of the semester.

year	#pre	#post	#both	fraction pre	fraction post	gain
2005	485	509	438	.57±.15	.66±.13	0.34
2007	356	356	355	.56±.15	.76±.12	0.46
2008	414	414	410	.58±.15	.79±.12	0.51
2009	612	565	527	.58±.14	.75±.12	0.41
2010	589	554	508	.60±.18	.78±.12	0.44
all	2456	2398	2238	.58±.15	.75±.15	0.40

2. CORRELATIONS ON PRE- AND POST-TESTS

Assuming that there are fundamental misconceptions shared by many students, the question becomes "how can we detect these in the test results". Since the MBT was designed with distractors compiled from open responses to those questions, one would expect that a specific misconception would lead students to give a specific wrong answer. If a misconception leads to wrong answers on two (or more) questions, we expect that students with this misconception would submit this particular pair of wrong answers with more than random frequency. We seek to detect such correlated pairs of wrong answers by looking for statistically excessive pairs of wrong answers, and that these will offer insight into the nature and prevalence of specific student misconceptions. We examine only correlations between wrong answers, since correct answers do not provide much information about misconceptions.

Table 2: Correlations between wrong answers on MBT pretest. For each pair of correlated wrong answers we show the overall correlation coefficient, the fraction of all students who gave the paired response, the Student's t-statistic, and the p-value. X indicates that a student did not answer the question (this is considered as a specific response).

Responses 1&2	Correlation [%]	Fraction [%]	t	p-value
Q1A Q2E	67	13	15.1	$\sim 10^{-37}$
Q4D Q5C	41	19	9.3	$\sim 10^{-18}$
Q11X Q12X	57	7	8.5	$\sim 10^{-14}$
Q9X Q11X	48	8	7.4	$\sim 10^{-11}$
Q9X Q12X	47	7	6.8	$\sim 10^{-10}$
Q13A Q14A	52	5	6.0	$\sim 10^{-8}$
Q13X Q14X	60	2	4.7	$\sim 10^{-5}$
Q20B Q21C	35	7	4.6	$\sim 10^{-5}$
Q20D Q22D	44	4	4.6	$\sim 10^{-5}$
Q20A Q22C	19	15	3.5	0.001
Q16C Q16D	18	13	3.0	0.004

We examined all possible wrong answer pairs, defining a binary variable for each possible wrong answer, specifying whether a particular student did or did not give that answer. We calculated the tetrachoric correlation between every pair of answers, as well as the amount by which the observed number of students giving the paired wrong answers exceeded the number expected assuming that each wrong answer was selected independently at random with the observed answer probability distribution for each question alone. All pretest correlations found to be significant at the p = 0.01 level are displayed in Table 2.

Because students with very low skill may have weak or inconsistent preconceptions and students with very high skill presumably have few misconceptions of any sort, we expect that certain misconceptions will be held primarily by students lying within a limited range of overall ability or perhaps in students only of low ability. To test this hypothesis, we divided the students into 7 equal partitions sorted by overall score and calculated correlation coefficients for each partition independently.

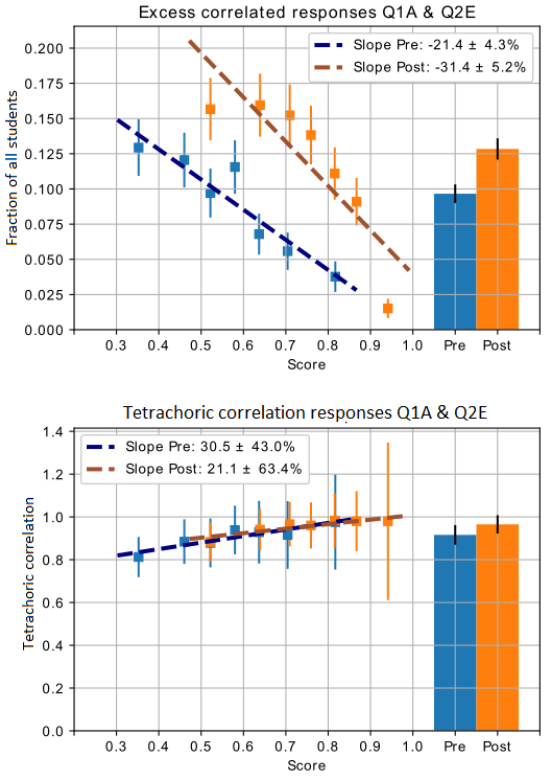
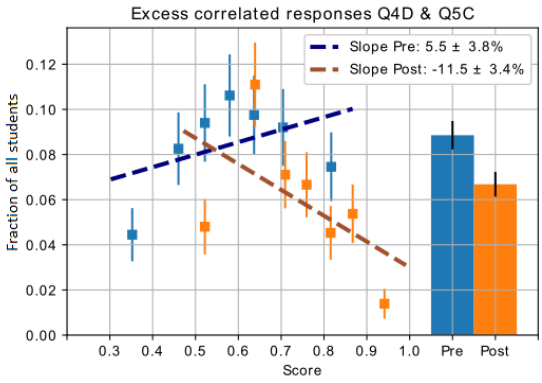


Figure 1: Questions 1 & 2: Velocity and Acceleration Graphs, correlation of 1A and 2E. The tetrachoric correlation and excess paired responses are plotted in each of seven cohorts divided by overall score.

In Questions 1 and 2, shown in Figure 1, the paired errors both correspond the same misinterpretation of a stroboscopic image of an accelerating object. The very high correlation coefficient implies that roughly 90% of the students who answered 1A also answered 2E. This suggests that the students determined the

acceleration (Q2) from the answer to the velocity (Q1), thereby making the same time-base error. This hypothesis is supported by the fact that the better cohorts made relatively fewer mistakes carrying out this prescription, hence had (even) higher correlations, as did all students on the post-test.



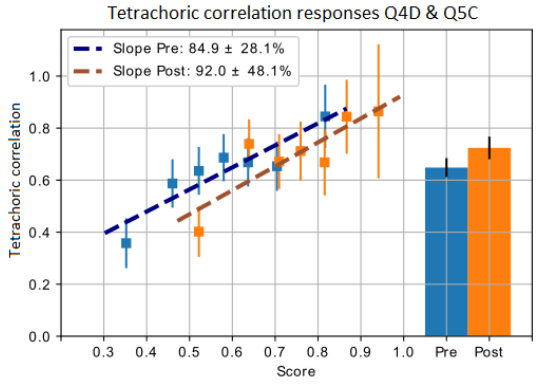
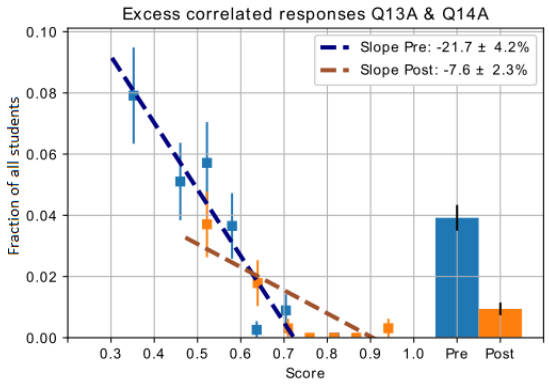


Figure 2: Question 4 and 5: Direction of Acceleration on Ramp - correlation of 4D with 5C.

Correlated wrong answers on Questions 4 and 5, shown in Figure 2, both correspond to ignoring real forces when applying F=ma. It is apparent that the prevalence of this error maximizes at score levels ~ 0.6 suggesting a specific misconception that shows some, but not too much, knowledge.

Correlated responses 13A and 14A both correspond to confusing the mass of a system with the force required to support it. This correlation is very strong (R ~ 0.9), but the probability of making



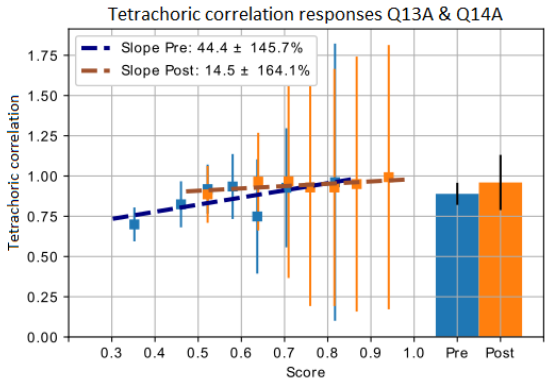
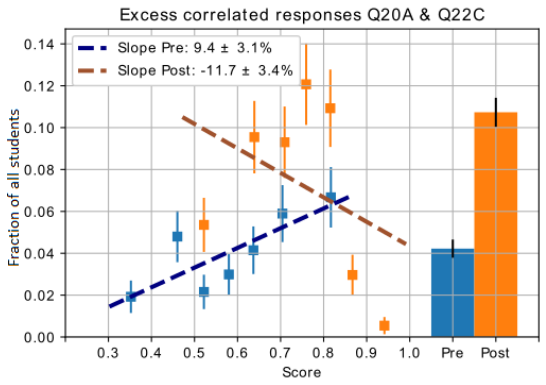


Figure 3: Questions 13 & 14: Elevator with Two Hanging Blocks – correlations between 13A and 14A.

this error drops dramatically with score, reflective of the fact that the associated error is virtually at a random rate with prevalence < ½% for all students scoring above 75% (where the correlation has huge errors). This seems to be an error predominantly made by low-ability students, and we suggest that it results from omitting g=10 m/s^2 when calculating weight from mass.



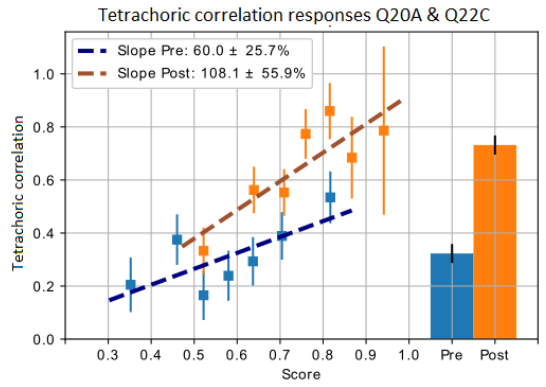


Figure 4: Questions 20 & 22: Pushing Different Masses the Same Distance with the Same Force, correlation of 20A & 22C.

The triplet of questions 20 through 22, Pushing Different Masses the Same Distance with the Same Force, yields several highly correlated pairs of wrong answers. These problems, particularly 20 and 22, are among the most difficult on the test, with respectively 36% and 47% of students answering them correctly on the pretest.

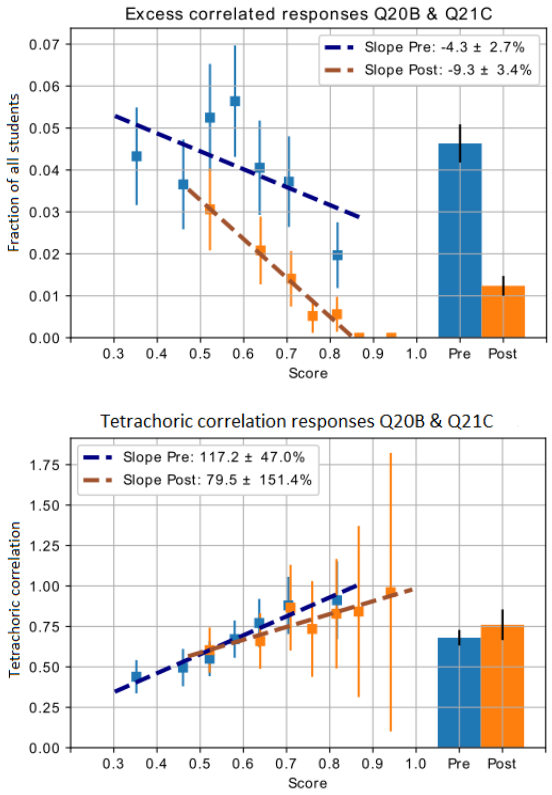


Figure 5: Questions 20 & 21: Pushing Different Masses the Same Distance with the Same Force, correlations between 20B and 21C.

The correlated pair consisting of 20A and 22C correspond to confusing the change in energy of a system with the change in momentum. About 4% of students showed this excess pairing on the pretest, rising to 10% on the post-test, the most dramatic of the only two increases in excess correlated responses found in this study. There is clear evidence that this excess correlation has a peak, probably around score 75%. Together with the dramatic increase in excess correlated responses on the post, we argue that this paired response requires confusion of work with impulse augmented by some understanding of momentum.

Similarly, the responses 20B and 21C, shown in Figure 5, seem to correspond to the idea that equal force results in equal acceleration, regardless of mass. This response decreases with increasing score and also from pretest to post-test. The correlation coefficient increases dramatically with score on both pre- and post-test.

The final correlated pair that comes from questions 20-22 is 20D and 22D, shown in Figure 6. These answers are both "too little information" to calculate the energy and momentum of two pushed pucks. Not surprisingly, this paired response shows the greatest decrease from pre- to post-test (~ 5:1), presumably because most students learn about either energy or momentum during the course.

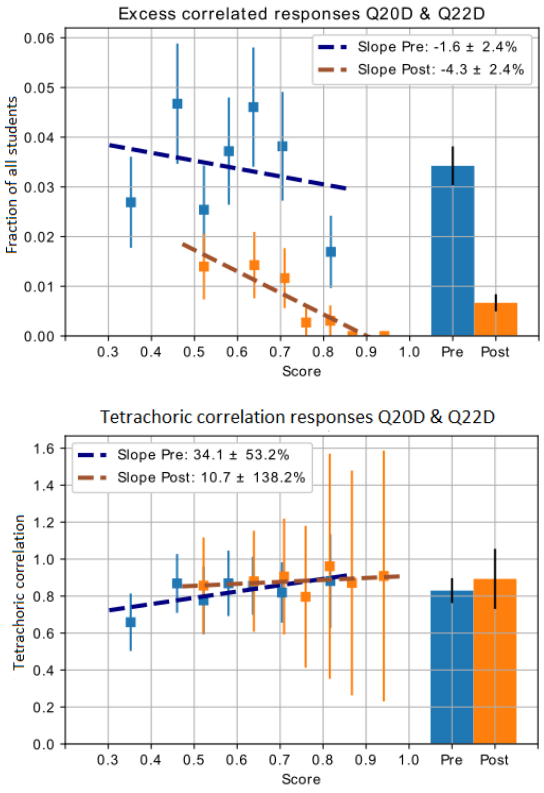


Figure 6: Questions 20 & 22: Pushing Different Masses the Same Distance with the Same Force: correlation between 20D and 22D.

3. TRANSITION ANALYSIS: PRE →POST ON THE SAME QUESTION

3.1 Robust Wrong Answers: Null Hypothesis and Findings

If a certain wrong answer on the pretest corresponds to an entrenched misconception, students should give that same answer on the post-test. We therefore use a baseline null hypothesis for comparison that assumes that students answer the post-test independently of their response on the pretest. We search for "excess" transitions above this null. When looking for wrong answers which are unusually strongly held (what we call "robust wrongs"), for example, our null hypothesis is that the student is unaffected by instruction and would answer with the same probability on the post-test as on the pretest. If 30% of all students answered correctly on the pretest, this would imply a 9% robust rate. This null hypothesis would reflect reality if all students were guessing on both pre and post.

The most robust wrong answer seen in Table 3, answer E on question 12, is "none of the above" on a numerical question, which does not suggest a specific physics misconception. The next two correspond to the same error in interpreting the motion diagram in a related pair of questions, namely reversal of the time axis. The fourth corresponds to claiming that the middle of the range of a graphed function is its average value. The fifth indicates that students have erroneously used the mass of part of a system instead of the total mass of the system in F=ma, and the sixth involves treating the speed of an object as its acceleration in an F=ma

problem. The first four of these give little insight into physics misconceptions, though they do seem to highlight mathematical deficiencies, but the robust wrong responses in Q17 and Q13 reveal difficulty with applying Newton's Second Law. Confusing speed and acceleration is a well-known student misconception.

Table 3: Six wrong answers were given by students on both the pre- and the post test at rates which were significantly greater than chance at the p<0.0001 level. Here we present the p-values for each of these responses and the frequency with which these responses were given as a percentage of all responses to the questions.

p-value	Question	%
$\sim 10^{-9}$	Q12E	14
$\sim 10^{-8}$	Q1A	6
$\sim 10^{-7}$	Q2E	5
$\sim 10^{-6}$	Q25B	4
$\sim 10^{-5}$	Q17C	10
$\sim 10^{-4}$	Q13C	3

3.2 Wrong to Correct: Null Hypothesis and Findings

Since the wrong answers on the MBT are designed to represent specific misconceptions, the question arises of whether students who give certain wrong answers on the pretest might be more or less likely than other students to subsequently provide the correct answer on the post-test. In other words, we wish to ascertain whether some misconceptions are more resistant to instruction than others. In calculating the excess (or deficit) relative to chance of students making a transition from a wrong answer to the correct answer, our null hypothesis is again that a student's likelihood of answering correctly on the post-test is independent of the answer they gave on the pretest. However, we must take into account that a non-trivial fraction of the students answer any given problem correctly on both pre- and post-test not by chance but because they understand the relevant physical concepts-- in some cases as many as 80% of students answered a problem correctly on both tests. We therefore use a slightly different null hypothesis that eliminates students who do not change their answer after instruction. This posits that the conditional probability of a student offering the correct answer to a particular problem given that they gave a particular incorrect answer to that problem on the pretest should be equal to the ratio of the number of students who transitioned to the correct answer from any incorrect answer over the total number of students who changed their answer in any direction. The most statistically significant wrong to correct transitions are displayed in Table 4 and discussed below.

3.2.1 Q1 and Q2: Find velocity and acceleration from a graph

Both transitions have moderate excess probability (~ 60%) of switching to the correct answer, and very small probability that the wrong is robust. This suggests that these wrongs are mainly due to careless errors in reading the graph.

3.2.2 Q14: Force from lower rope on top block of two hanging in stationary elevator

About 6.5% answered D (20N, twice the answer) or A (forgot multiplying by g) and at least 80% of both switch to correct. This generally shows strong growth on applying Newton's Laws. (Although most students probably saw this example in the course.) The very small number of robust wrongs shows that the initial answers may have been mostly due to lack of full understanding of tension rather than strongly held misconceptions.

3.2.3 Q23: Average acceleration from graph of velocity versus time

The two most attractive wrong answers, taking v=0 at t=0 ($p < 10-4$) and "none of above" ($p < 10^{-3}$) both exhibited excess transitions to the correct answer. Students with pre-answers switched to correct with 78% and 80% likelihood. This is a graphing question, so possibly learning about graphs is reinforced due to complimentary instruction on graphs of functions in the introductory calculus courses which a majority of students are co-registered for. NOTE: 14% of those who were correct on the pretest answered incorrectly on the post.

Table 4: Wrong to correct transitions which occur significantly more frequently than would be expected due to chance. We display the p-values and the overall frequency with which the transition occurred for all such transitions with p<0.001.

p-value	Transition	Freq. [%]
$\sim 10^{-6}$	Q2E2D	10
$\sim 10^{-5}$	Q1A2B	11
$\sim 10^{-4}$	Q23C2D	9
$\sim 10^{-3}$	Q14D2B	8

4. CONCLUSIONS AND DISCUSSION

4.1 Excess Correlated Wrong Responses

"Excess correlated responses" (ECR) are in addition to those that would occur if the correlated questions were independently answered randomly with the observed frequency of wrong answers. Correlated wrong answers between different questions were detected and described in two ways: by the excess fraction of students who selected both wrong answers (vs. assuming independently answered questions), and by the fraction of students who selected one wrong answer who also selected the other (tetrachoric correlation). Both quantities varied considerably with the overall ability of the students as measured by their overall fraction correct (score) on the assessment. For this reason, we discuss only results specific to student overall score.

The correlated wrong answers found here are surprisingly prevalent, with ~10% or more of the students in one of the score groups selecting both of the paired wrongs in all cases except the last two which have the lowest statistical significance. Our most important findings are:

1. The percentage of correlated wrongs always drops for students with score >0.7, and typically decreases to 1% or lower for the top score group on the post-test.
2. In the two cases suggesting a real misconception, force from ramp and kinetic energy of masses, the percentage of correlated wrongs also decreased for the lowest-scoring groups.

The tetrachoric correlation measures the "purity" of the observed correlations. In every case presented, it reaches or exceeds 0.8 for groups with high test scores. This shows that essentially every skilled student giving one of the paired wrong answers also gives the other. Equivalently, the mistake or misconception is the main cause of the wrong answers on both questions. In cases where low-skill students appear to lack the correct physics knowledge (energy/momentum and direction of force on curved ramp), the correlation decreases to well below 0.5. Low tetrachoric correlation probably indicates that students are using a variety of incorrect reasons in their responses, so that many are led to answer one of the paired wrong answers but not the other.

In summary, the search for excess correlated answers has revealed two cases where the excess peaks for students in a particular range of overall score. This is a clear guide for instruction: if you teach a class is in this score range, then you should carefully address situations like this to tease out and rectify the underlying misconception. Additionally, the dramatic increase in mistaking work as the source of momentum on the post-test indicates that our instruction has to be clarified on this point. We find that the correlation of all wrong answer pairs increases for better students - indicating that this misconception is the main reason for these wrong answers and is being consistently applied to both questions I.e. skillful students don't make errors on just one of the problems due to some reason unrelated to the identified misconception.

4.2 Excess and Robust Transitions

We found that the none of the transitions from wrong to right indicated that that particular wrong answer was conceptually closely related to the correct answer; rather it seemed that the wrongs were due to careless responses or fuzzy thinking. On the other hand, several of the robust wrong answers seemed to reflect physics misconceptions.

5. SUMMARY

The probability of each particular ECR varies substantially with the overall ability (measured by total score) of the students, ranging up to a maximum of 4,5,8,10, and 11%. Although it always drops to ~ 1% or less at the highest ability, we find examples where the probability of ECR peaks at low and at medium student score. In all cases, the fraction of students giving one wrong answer who also give the other exceeds 80% for the highest-scoring students. This suggests that teachers concentrate on remediating ECR's common to their students' scores. ECR's seem to be a good method to detect significant misconceptions or missing knowledge held by students of a particular ability.

The transition analysis showed that robust wrongs often reflected misconceptions in math or physics, but that excess transitions from wrong to correct generally reflected carelessness rather than a mindset primed for learning the correct response.

6. FUTURE DIRECTIONS

The present work offers a new method for finding excess correlations of wrong answers between different questions, and particularly common (or uncommon) learning transitions within one question from pre- to post-test. Two future directions seem important to explore:

1. This method should be compared with network analysis which has a similar objective [2].
2. The students at MIT are significantly stronger than most who take the MBT. It is therefore important to extend the analysis to students with lower overall ability as evidenced by lower overall scores on the pre-test.
3. We have a new way to assess misconceptions; this should enable us to find better ways to remediate them.

7. ACKNOWLEDGEMENTS

We thank the Office of Digital Learning at MIT for financial and technical support. ÁPL thanks the Distinguished Scholar program of Spain for support.

8. REFERENCES

[1] Adams, W., Perkins, K., Podolefsky, N., Dubson, M., Finkelstein, N., & Wieman, C. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, *2*(1), 10101. http://doi.org/10.1103/PhysRevSTPER.2.010101

[2] Brewe, E., Bruun, J., & Bearden, I. G. (2016). Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data. *Physical Review Physics Education Research*, *12*(2), 1–19. http://doi.org/10.1103/PhysRevPhysEducRes.12.020131

[3] Cardamone, C. N., Abbott, J. E., Rayyan, S., Seaton, D. T., Pawl, A., & Pritchard, D. E. (2011). Item response theory analysis of the mechanics baseline test. In *Physics Education Research Conference* (Vol. 1413, pp. 135–138). Omaha, Nebraska. Retrieved from http://dspace.mit.edu/handle/1721.1/78319

[4] Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The Physics Teacher*, *33*(8), 503. http://doi.org/10.1119/1.2344279

[5] Hestenes, B. D., & Halloun, I. (1995). Interpreting the Force Concept Inventory A response to Huffman and Heller. *The Physics Teacher*, 502–506.

[6] Hestenes, B. D., & Wells, M. (1992). A Mechanics Baseline Test, (March).

[7] Lindell, R. S., Peak, E., & Foster, T. M. (2007). Are they all created equal? A comparison of different concept inventory development methodologies. *Physics Education Research Conference*, *883*, 14–17. Retrieved from http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.2508680%5Cnpapers3://publication/doi/10.1063/1.2508680

[8] McKagan, S. (2018). Physport.

[9] Swackhamer, G., Hestenes, D., & Wells, M. (1992). Force concept inventory. *The Physics Teacher*. http://doi.org/10.1119/1.2343497