Identifying characteristics of pairs of questions that students answer similarly

Trevor A. Balint[1], Raluca Teodorescu[2], Kimberly Colvin[3], Youn-Jeng Choi[4], and David E. Pritchard[4]
[1]*Department of Physics, George Washington University, Washington, DC 20052*
[2]*Department of Physical Sciences, Montgomery College, Takoma Park, MD 20912*
[3]*Department of Educational and Counseling Psychology, University at Albany SUNY, Albany, NY 12222*
[4]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139*

Abstract: We discover pairs of questions (items) that students answer in a dependent manner by applying Fisher's Exact Test to a sample of 1080 students answering 257 items in the MOOC 8MReVx. To eliminate false positives arising from the large range of student abilities, we divided students into groups of similar ability by: 1) similar percentage correct on first try (CFT) of all available problems, 2) similar percentage CFT of problems attempted, and 3) similar skills determined by Item Response Theory. All three methods dramatically reduce the number of false dependencies in a similar fashion. Importantly, using 5 groups 70% (vs. 5% without groups) of dependent pairs of items have identical ratings in all four categories of the Taxonomy of Introductory Physics Problems, [1] 70% (vs. 23%) are of the same pedagogical type, 45% (vs. <1%) are items in the same problem, and 100% (vs. 12%) reside in the same course Chapter.

PACS:01.40.Fk,01.40.Ha

I. INTRODUCTION

Dependence testing is commonly used to determine whether an association between two items is significant. In the context of education, dependence (or correlation, which implies dependence [2]) has been used to discover relationships between solution methods and effectiveness [3]; whether a student's interest in introductory physics is related to their becoming a physics major [4]; and for testing the relationships between elements of a course or assessment [5] - among other uses. In this paper, we test the dependence of student responses to two questions or items in a course to determine which items in the course are most closely related. This will allow the course materials to be rearranged in a way that is most beneficial to students.

We use Fisher's Exact Test of independence to analyze data from the 2013 offering of the Massive Open Online Course 8.MReVx designed by the RELATE (REsearch in Learning Assessing and Tutoring Effectively) Group [6] at the Massachusetts Institute of Technology and offered on the edX platform. [7] This course contains a wide variety of physics problems targeting various levels of thinking. [8] Our analysis focuses on 1,080 participants (out of 16,787 enrolled into the course) who achieved greater than 50% of available points, as this is a strong indicator of students who participated actively in the entire course.

In this paper we first illustrate how we use dependence testing to analyze physics items (individual questions whether within a multi-part problem or stand-alone). Then, we use three different methods to sort participants into groups of similar ability to eliminate "false positive" dependencies. Finally, we qualitatively analyze the dependent question pairs using the Taxonomy of Introductory Physics Problems (TIPP) [1] and search for commonalities between the cognitive processes and the type of knowledge involved in the dependent item pairs.

II. DEPENDENCE TESTING

In dependence testing two items are defined to be independent if the probability of a student answering either one correctly is unrelated to the student's answer on the other one. Equivalently, the joint probability for all students is equal to the product of the separate probabilities:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad (1)$$

Note that students with higher ability have higher probabilities on both problems; the point is that the probability of a student answering X correctly is $p_X(x)$ independent of their answer on Y. This implies that if we find a pair of items that are not independent then those items are dependent and have a non-trivial relationship of some kind (e.g. the answers are correlated or uncorrelated).

We are searching for violations of the null hypothesis - that every pair of questions in the course is independent. We examine each pair of questions by constructing a contingency table, an example of which is shown in Table 1. For this hypothetical pair of items we see that 12 out of 20 (60%) of students got item 1 correct and 11 of the 20 (55%) got item 2 correct. This would predict that 33% (i.e. $a = 7$) would get both correct and 18% ($d = 4$) would get both wrong. [9]

2015 PERC Proceedings, edited by Churukian, Jones, and Ding; Peer-reviewed, doi:10.1119/perc.2015.pr.009
Published by the American Association of Physics Teachers under a Creative Commons Attribution 3.0 license.
Further distribution must maintain attribution to the article's authors, title, proceedings citation, and DOI.

TABLE 1. Example of contingency table.

	Item 1 Correct	Item 1 Incorrect	Fractional Totals
Item 2 Correct	9 *(a)*	2 *(b)*	11 *(a+b)*
Item 2 Incorrect	3 *(c)*	6 *(d)*	9 *(c+d)*
Fractional Totals	12 *(a+c)*	8 *(b+d)*	20 *(N=a+b+c+d)*

Given the fractional totals for each row and column, Fisher's Exact Test determines the probability that the distribution of students in the table could occur randomly by using the Eq. 2, in which the brackets represent a binomial coefficient:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}. \quad (2)$$

Performing the test on the example pair we obtain a value of p=.0367. For dependence testing a p-value of .05 is usually chosen meaning that p<.05 indicates that the items are likely dependent at the 95% level as the above pair is.

In addition to finding pairs of items with statistically dependent responses we examine the determinant ($C = ad-bc$) of the contingency table to determine whether the item pair is correlated or anti-correlated. A positive determinant (associated with positive correlation) indicates that students are more likely to have the same results to both items - either both correct or both incorrect. [9] Similarly, a negative determinant (associated with anti-correlation) indicates that students are more likely to get one item correct and one item incorrect. For the above example in Table 1 we observe a = 9 and d = 6, suggesting (along with the item pair's dependence as found using Eq. 2) a positive correlation between the answers to the two items. In this case the table's determinant is positive.

A. Skill-based student sorting

In practice, variation in the abilities of the students within a class can produce false positive correlations. High (low) ability students will answer both questions correctly (incorrectly) thereby populating cell a (d) with neither group populating the cells b and c; hence C will be greatly positive even though each group answered the questions independently. To avoid using the large sample size of the MOOC's student body (which is recommended against in the literature [10]) and to reduce these false positives we sorted the students into N even groups using three ability-related methods and performed dependence testing on each. For these smaller groups, p=.05 is an appropriate dependence criterion. We consider a pair of items "strictly dependent" if they meet this dependence criteria (p<.05) for all groups. Increasing the number of groups increases the strictness of this requirement (see Fig. 1) and dramatically reduces the number of dependent pairs identified without adding any new pairs (e.g. every pair found with N groups is also found in the longer list with N-1 groups).

To illustrate the false positives due to ability variation we also used random binning, doing it repeatedly to obtain a smooth average curve shown as the solid red line in Fig. 1. By examining the difference between the random and the three performance-based grouping methods described below we can see how many false positives are due to ability variation.

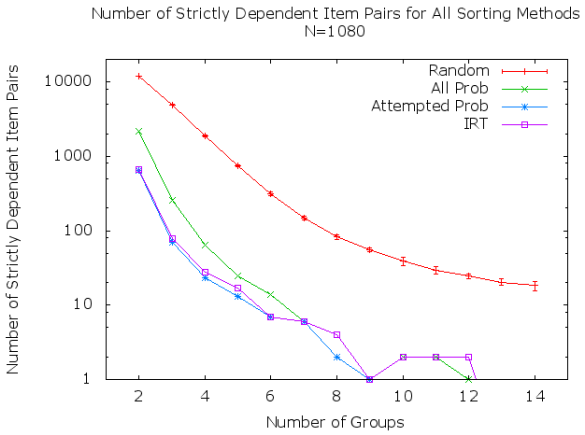

FIG 1. Number of strictly dependent item pairs for each division of the class population into even groups using all sorting methods. The x-axis represents the number of groups that the class is divided into after being sorted by the four methods. The y-axis represents the number of statistically dependent item pairs. A total of 32,640 item pairs exist in this course.

Our first non-random sorting method, named the "All Problem" sorting method, sorts the students by the percentage of all items in the course that they completed successfully on the first try:

$$\frac{\text{Number of items solved correctly on the first attempt}}{\text{Number of items in the course}} \quad (3)$$

The number of strictly dependent item pairs found with this sorting method is shown as the green x's in Fig. 1.

Our second sorting method, named "Attempted Problem", sorts students based on the percentage of questions attempted by each student that were answered correctly on the first attempt:

$$\frac{\text{Number of items solved correctly on the first attempt}}{\text{Number of items attempted}} \quad (4)$$

The number of strictly dependent item pairs found using this method is shown as the blue filled boxes in Fig. 1.

The final method used to sort students involves a two-parameter logistic Item Response Theory (IRT) model. [11] This model uses an iterative process to generate a skill for every student and two parameters for every item that

represent its difficulty. From these parameters one can generate the probability of a certain student getting an item correct. We sorted the students based on their IRT skills and performed dependence testing. The number of strictly dependent question pairs found using this sorting method is shown as the purple empty boxes in Fig. 1.

The three performance-based grouping methods shown in Fig. 1 decline similarly with increasing N. One can use this behavior to tune this analysis to an appropriate number of groups (or item pairs) for the study at hand; the behavior of how this analysis changes with N is understood, making quartiles or quintiles (or any other group division) equally valid for future research using this method.

We now confine our attention to N=4 groups as that provides a large but not too large number of strictly dependent item pairs. Doing so (Fig. 2) shows that the All Problem sorting method is the least strict. Importantly, over 80% of the pairs it identifies are confirmed by IRT sorting, and of those all but one is found by the strictest All Problem sorting method. This suggests that we can isolate a core of strictly dependent problem pairs common to all three methods.

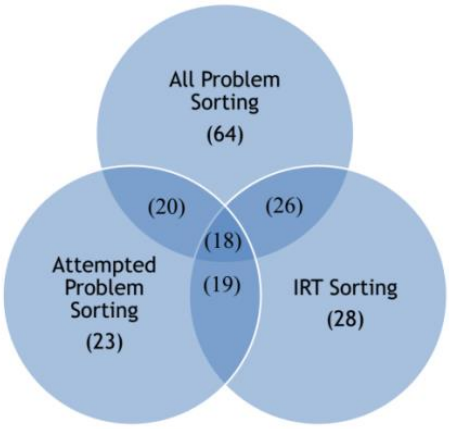

FIG 2. Comparison of the number of strictly dependent item pairs between sorting methods at 4 group divisions.

III. PEDAGOGICAL ANALYSIS OF DEPENDENT QUESTION PAIRS

We now show that the strictly dependent items (found by the above quantitative analysis using any of the three methods) have commonalities of cognitive level, topic, and pedagogical type.

We used the Taxonomy of Introductory Physics Problems (TIPP) [1] to classify the items according to the cognitive processes and the knowledge they involve. In TIPP each item receives four labels that describe: the type of declarative knowledge, the type of procedural knowledge, the highest cognitive process involved related to the declarative knowledge, and the highest cognitive process involved related to the procedural knowledge. Here we examine the similarities in the mental processes required to solve each item in an item pair.

We define the TIPP Similarity Score of an item pair as the number of matching labels. It ranges from 0 if no labels match to 4 when they all match. Figure 3 shows the distribution of TIPP Similarity Scores for all item pairs in the course with no grouping or dependence testing performed. Out of 32,640 item pairs in the course, 1708 (5.2%) have a TIPP similarity score of 4.

Also illustrated in Fig. 3 is the distribution of the TIPP Similarity Scores for all pairs of items found to be strictly dependent by any performance-based grouping method using two through five groups; that is, any item pair found to be strictly dependent by at least one (but not necessarily more than one) of the three sorting methods. Most significantly, as the number of groups is increased the percentage of strictly dependent item pairs with a Similarity Score of 4 (identical TIPP classification in all 4 categories) rises to 60% (80%) with 4 (5) groups. This demonstrates that items with the same TIPP classification are about 14 times more likely to be answered in a dependent manner than randomly selected item pairs.

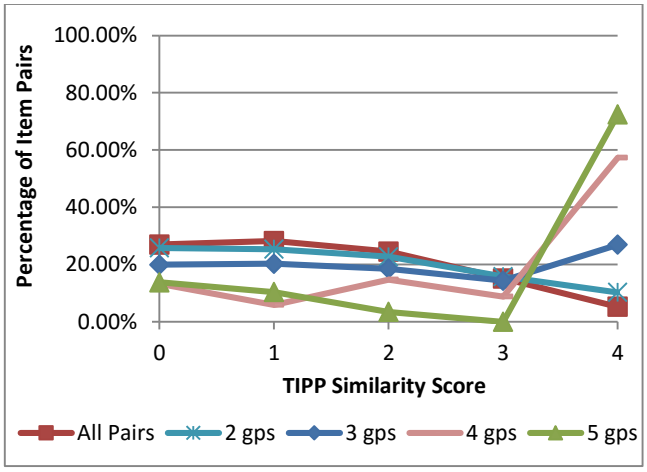

FIG 3. Distribution of TIPP Similarity Scores for all item pairs in the course (red squares, 32,640 pairs) and for strictly dependent pairs at 2-5 group divisions.

We also gave each question a label called "Pedagogical Type" which can be Symbolic, Conceptual, Numerical, Sense Making, or Multiple Concept. As the number of groups is increased this measure increases from ~20% (expected randomly given only 5 pedagogical types) to 60% as shown in Fig. 4. In addition, increasing the number of subgroups dramatically restricts the discovered item pairs to the same unit and very significantly restricts them to be different questions in the same problem (as would be expected since part b often depends on answering part a correctly).

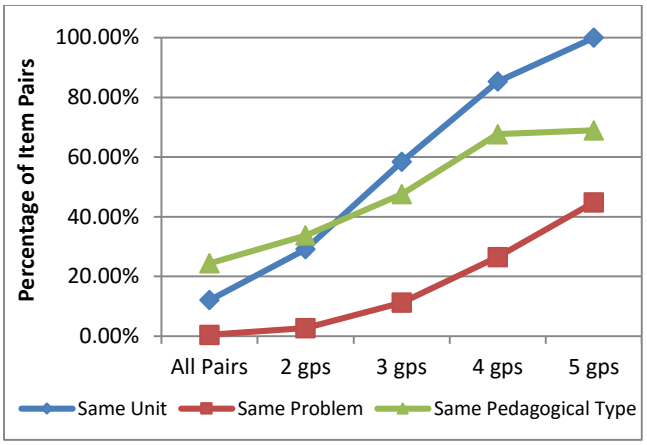

FIG 4. Percentage of all item pairs and strictly dependent item pairs at 2-5 group divisions that satisfy the requirement of being in the same unit, being a sub-part of the same problem, or having the same pedagogical type.

IV. CONCLUSION AND OUTLOOK

In this paper we have applied standard dependence testing methods to students in a MOOC who have widely varying ability. We have demonstrated that three different performance-based methods of grouping the class all dramatically reduce the number of false positives generated by this variability without further correction. Furthermore, we have shown that using 4 or 5 groups with the criterion of "strict dependence" results in identifying item pairs that almost all lie in the same unit, 70% of which have identical TIPP labels, and 60% have the same pedagogical type. This not only suggests that these item pairs are answered in a highly correlated manner, but also that these pedagogical classifications are valid measures of student' abilities in that they indicate which pairs will be answered similarly.

While a different methodology could have been used to find common item pairs (such as factor analysis or MIRT), these analyses tend to find broader similarities such as topic or format between items that they classify as similar. Here our goal was to discover dependent pairs of items, then determine those items' similarities in cognitive processes and other qualitative measures.

A limitation of our method is that it does not distinguish between dependency that originates from correlated responses and that which originates from anti-correlated responses. However, we can extract information about correlation from an examination of the determinant of the contingency tables, as discussed above under "Dependence Testing". For the class unsorted and undivided into subgroups, approximately 90% of the item pairs' contingency tables had positive determinants. At two class divisions, over 95% of strictly dependent item pairs had all determinants positive in the subgroups' contingency tables for that item pair. This percentage increases linearly to 98% as the number of group divisions increases to five.

We think that our method of determining dependent item pairs is too strict, and therefore fails to identify many pairs that are quite correlated, but this work is a first step towards the development of a method for evaluating the overall probability of a positive correlation. We believe that this will allow advances in ordering and revising physics curricula.

ACKNOWLEDGEMENTS

We thank MIT and the Google Faculty Awards Program for financial support.

[1] R. Teodorescu, C. Bennhold, G. Feldman, and L. Medsker, Phys. Rev. Spec. Top. - Phys. Educ. Res. **9**, 010103 (2013).

[2] M. Hazewinkel, editor , *Encyclopaedia of Mathematics* (Springer Science & Business Media, 2013).

[3] K. R. Koedinger and M. J. Nathan, J. Learn. Sci. **13**, 129 (2004).

[4] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, Phys. Rev. Spec. Top. - Phys. Educ. Res. **2**, 010101 (2006).

[5] T. Mzoughi, Procedia - Soc. Behav. Sci. **191**, 235 (2015).

[6] RELATE | Research in Learning, Assessing and Tutoring Effectively, http://relate.mit.edu/.

[7] Mechanics Review, https://www.edx.org/course/mechanics-review-mitx-8-mrevx.

[8] R. E. Teodorescu, D. T. Seaton, C. N. Cardamone, S. Rayyan, J. E. Abbott, A. Barrantes, A. Pawl, and D. E. Pritchard, in *AIP Conf. Proc.* (American Institute of Physics (AIP), 2011), pp. 5–8.

[9] B. S. Everitt, *The Analysis of Contingency Tables, Second Edition* (CRC Press, 1992).

[10] A. Rubin, *Statistics for Evidence-Based Practice and Evaluation* (Cengage Learning, 2012).

[11] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory* (SAGE Publications, 1991).