Is the Doer Effect Robust Across Multiple Data Sets?

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
koedinger@cmu.edu

Richard Scheines
Department of Philosophy
Carnegie Mellon University
scheines@cmu.edu

Peter Schaldenbrand
Human-Computer Interaction Institute
Carnegie Mellon University
pschalde@cs.cmu.edu

ABSTRACT

The "doer effect" is the assertion that the amount of interactive practice activity a student engages in is much more predictive of learning than the amount of passive reading or watching video the same student engages in. Although the evidence for a doer effect is now substantial [6, 7, 12], the evidence for a causal doer effect is not as well developed. To address this, we mined data for evidence of a causal doer effect across multiple domains. We examined data from two online courses in Psychology, one in Biology, one in Statistics, and two in Information Science, applying causal discovery algorithms [14] in Tetrad to each. Assuming that factors driving a student's choices regarding how to spend their time in an online course are temporally prior to their performance on quizzes and exams, we found evidence of a causal relationship in every domain we studied. We did not find evidence that a unique causal model held in every domain we studied, but when we estimated the size of the causal relationships in the models we found in each domain, we did find evidence in every case that doing has a much stronger quantitative effect on learning than either reading or watching video. This work may be the first EDM effort to explore the generalizability of a causal claim about learning across multiple datasets from a variety of courses and contexts of use. It makes vivid the role of causal data mining algorithms in educational research. The evidence presented furthers the case for doer effect causality, but also recommends a need for richer data with more student background and learning process variables to better isolate causal directionality without assumptions about temporal order and unmeasured confounds.

Keywords

Doer effect; learning by doing; causal discovery

1. INTRODUCTION

When students take an online course, or use a cognitive tutor, a log of data is created that records their interactions with the course or tutor. Mining this data for causal information concerning what sorts of student behaviors cause better learning outcomes is crucial if we are to intervene, either on the design of the online material, or on the student's behavior more directly.

In this paper, we explore the causes of learning in several online courses using Tetrad and Tigris/LearnSphere. Tetrad (http://www.phil.cmu.edu/tetrad/) is a causal discovery tool that has already proved helpful in educational data mining [6, 10], and LearnSphere is a collaboration dedicated to providing data and tools for analyzing information pertaining to student learning (http://learnsphere.org/). LearnSphere combines data and analysis tools with Tigris, a workflow tool that connects data from the educational data repository DataShop [5] to analytical programs such as Tetrad. Tigris runs in a web browser and has functionality to use the abilities of Tetrad and share results of analyses with other Tigris users. Tigris allows users to test theories across diverse datasets, and this was precisely our goal in the work we describe here. Tigris connects analytical tools to data and users via their research. LearnSphere users can upload datasets to DataShop [5] and make them available in workflows. They can also share their own analytics as well as workflows they construct in Tigris. The causal models and analysis in this paper were executed using the Tetrad implementation in Tigris.

The causal discovery algorithms in Tetrad operate on graphical causal models [14], which allow us to rigorously represent the qualitative causal structure of a domain with a directed graph, and to connect the structure of the graph to statistical constraints that we can test on measured data. The algorithms compute the equivalence class of causal structures that are consistent with background knowledge about the domain. In some cases the equivalence class is not very informative - for example the equivalence class of a system of two variables X,Y that are correlated is: $X \rightarrow Y$, $X \leftarrow Y$, $X \leftarrow \text{Confounder} \rightarrow Y$. In systems involving more than 2 variables, the causal information from an equivalence class can be much more informative.

The question of how to judge whether or not to believe an equivalence class output by the algorithms is very complicated and very interesting. All models within an equivalence class have the same "fit" with data, but whether the statistical fit is "good enough" to warrant belief depends on a large number of factors. This is by no means a problem that is special to causal discovery algorithms, however, and it is not the subject of our work. It is one that should concern all data-mining procedures, including ones that involve a single human building a hypothesis and then testing it on a single dataset.

Our concern in this paper is whether or not evidence for a causal doer effect generalizes across courses and contexts. We studied courses with diverse subject matter and diverse student populations.

The "doer effect" is the assertion that the amount of interactive practice activity a student engages in is much more predictive of learning than the amount of passive reading or watching video the same student engages in. We want evidence of a causal doer effect, that is, intervening to increase the amount of interactive practice would result in better learning outcomes.

Previous work has provided some evidence for a causal doer effect. In [12], 52 students at the University of Pittsburgh took an online course in which five variables were measured: pretest, percent of

modules printed, percent of interactive exercises completed as a measure of "doing", average end of module quiz score, and score on final exam.

Printing out modules was convenient and more common among good students, but it reduced the likelihood that students would complete interactive exercises (they could not do these on the printed modules). It thus served as an "instrument" for the doing → Quiz → Final exam relationship.

This relationship between performing active assignments and a learning outcome was directly researched in [6] and coined the "doer effect" in [7]. A dataset with six variables was examined in [6]. In this data, the relationship between doing and learning was far stronger than the relationship between passive activities such as watching videos or reading course material and learning.

Furthering the evidence for the doer effect, in [7], the relationship was tested on four other datasets, using regression methods. These were a diverse set of courses, but all had shown a strong link between doing and learning. While a strong correlation between doing and performance was shown in [7], the causal relationship was not tested. In this paper, we extend the investigation of whether the doer effect is causal by explicitly employing causal discovery techniques in Tetrad to these additional datasets.

We examined relationships between approximately six variables that are persistent throughout course subject matter, student populations, and time. Our research question: Is there evidence that the doer effect is causal across multiple contexts/datasets?

2. RELATED WORK

Much of the EDM research has investigated correlational relationships in predictive models. In [11], correlations of variables predict whether a student will enroll in college. While having a successful predictor of college attendance is good, it would be more useful to educators to understand the causes of college attendance so they can make interventions and increase applications and yield. In [13], correlation mining is used to explore a relationship between the features of a math problem and student learning. They acknowledge that future work would have to go into determining if these relationships are causal. Only once the relationships are determined to be causal can they assuredly be used to influence course design. Analyzing whether these relationships are causal by performing a randomized assignment experiment is the gold standard for making causal inferences, but this is often impractical, and there are thousands of non-experimental datasets available with which we can test the external validity (or generalizability) of hypotheses across multiple contexts [8]. Thus, it is worthwhile to pursue the use of causal discovery methods designed for non-experimental data on such datasets [6, 9].

Research into students' attitudes toward a math tutor [4] conclude that correlations exist between empathetic messages in the tutor and a student's mood toward it. They suggest that the positive correlation they found is indicative of a relationship in which increasing the empathy of these messages would cause a better mood amongst the users of the tutor. This implies a causal relationship, but they do not consider confounding variables or causal discovery algorithms [14].

Previous work in EDM that has researched causal relationships include [3] and [9]. Both of these use causal discovery algorithms and [9] uses Tetrad. Rather than resource use variables found in this paper, [3] uses variables that measure a student's interest and actions in a tutor, and it provides evidence for causal relationships between these variables and a final exam grade.

These past efforts [3, 6, 9, 12] have performed analyses on single datasets and, as such, there remains an opportunity to use the vast number of datasets available to probe external validity. This paper is distinctive in this regard -- to our knowledge, this is the first EDM effort to explore the generalizability of a causal claim about learning across multiple datasets from a variety of courses and contexts of use.

3. METHODS: CONFIRMATORY & EXPLORATORY WITH CRITERIA

We pursued both confirmatory and exploratory approaches to addressing our research question by analogy, for example, to confirmatory and exploratory factor analysis [15].

3.1 Method 1: Confirmatory Analysis of Causal Model Generality

Our confirmatory analysis involved testing a causal model that displayed the doer effect that was derived from data aggregated from a class offered at Georgia Tech in 2013, (https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=863) on five other datasets. We tested if the model statistically fits each dataset, according to the goodness-of-fit measures common in linear causal models [14]. We know of no successful attempts to test a specific causal model discovered on one dataset on other datasets collected in widely varying contexts, as in our datasets which have different kinds of course activities collected in different educational settings and with different available measures of student performance and different sizes of data. In attempting this confirmatory analysis, we discovered that it was neither going to confirm nor deny the doer effect hypothesis. We present it nevertheless as a cautionary message for others who may be tempted to do the same and to explain how dataset variations, particularly dataset size, make inferences from a confirmatory analysis problematic.

The causal model in Figure 1a was the model discovered on data from a 2013 Georgia Tech psychology course [6]. The model was previously [6] discovered using the Tetrad Java application, but in this paper, the analysis was performed using Tetrad's implementation in LearnSphere's Tigris workflow tool resulting in the same model structure, with negligible edge coefficient differences. The dataset features six variables measured on 939 students. One variable is a prior knowledge assessment (Pretest), one is a measure of doing in terms of the number interactive activities students performed (activities_started), two are measures of student use of passive learning resources including text page reading (non_activities_pageview) and video watching (play), and two are measures of learning outcome including the total across 11 unit quizzes (T.Quiz) and a final exam score (Fina_Exam). A directed edge in a causal model depicts evidence of a direct causal relationship between the variables. The coefficient on the edge is an indication of the strength of the causal relationship.

The primary feature to note in the causal model in Figure 1a is that while the outcome measures (T.Quiz and, indirectly, Fina_Exam) are effects both of passive resource use (non_activities_pageview and play) and active resource use (activities_started), it is the active resource use that exhibits the much stronger relationship. This large difference (0.44 vs. .06) is the doer effect. It is also important to note that the edges in the model do not represent correlations between the variables; they express and quantify direct causal relationships. For example, while activities_started and Fina_Exam have a correlation coefficient of 0.28, the causal inference algorithm determines they do not have a direct causal relationship.

Table 1. How the various naming schemes of datasets relate to each other.

	Psychology Georgia Tech	UMUC: Bio, Psych, Stat, InfoSci	C@CM
Pre-assessment	Pretest		Pretest
Doing activities	activities_started	activities_started	activities_started
Reading text pages	non_activities_pageview	non_activities_reading	non_activities_pageview
Watching lecture videos	play		
Unit level assessments	T.Quiz	total_quiz_proportion	T.Quiz
Cumulative assessment	Fina_Exam	final_grade_in_number	C@CM_Final_Exam

It does so by finding that when conditioned on T.Quiz, Fina_Exam and activities_started are independent.

A final note is to emphasize that the causal claims are about the *constructs* being measured not about the *measures* themselves. For example, the causal link between T.Quiz and Fina_Exam indicates that better competence attained during the course (the construct that T.Quiz measures) causes better competence at the end of the course (the construct that Fina_Exam measures). It *does not* imply that merely raising a T.Quiz (e.g., by making the quiz easier) would cause final exam scores to increase

A difficulty with testing a model on different datasets is the fluctuating naming schemes of variables and the inconsistency with which variables are contained within datasets. For instance, GTech's psychology dataset contains seven variables while a dataset from The University of Maryland University College, which is also used in this paper, has four variables. The four variables in the UMUC data are a subset of GTech's psychology data. For each dataset, we used the closest set of variables we could construct. Table 1 shows our decisions.

To facilitate comparison across datasets in the confirmatory analysis, we used the maximum number of variables that were common to the original dataset and the dataset being tested. We used five variables when we tested the original model on C@CM and four variables when we tested it on the UMUC datasets.

While we received UMUC data from the previous study [7], we added a sixth dataset from an online course on basic computing offered at Carnegie Mellon which we call Computing@Carnegie Mellon. A pre-assessment variable was created for each student by averaging the highest scored attempt at each pre-assessment quiz. The same process was performed on unit level assessments for each student. The number of active activities was the number of activities that each student started, and the number of passive activities was calculated in the same way as [6]. For a student to get to an activities page, they needed to visit a readable page. To accurately represent the number of pages read by a student, the total number of readable pages each student visited was subtracted by the number of activities they performed divided by a ratio. This ratio was the number of activities started to the total pageviews of the student with largest number of activities started. Therefore, the page viewing variable would not quantify the pages that students viewed merely as a stepping stone to get to activities. Once we made these datasets compatible with GTech's data, we could test our original model on five datasets.

3.2 Method 2: Exploratory Analysis with Criteria

Our second pass at answering our research question involved exploratory analysis whereby we applied a causal discovery algorithm to each dataset instead of confirming the original model on the other datasets. In this approach, we don't expect to find the same model on each dataset, but we do hope to see evidence of a causal doer effect in each context. We asked the question: What are the properties of the search output that would constitute evidence of the causal doer effect? These properties will be the criteria that we use to determine if each different context provides evidence of a causal doer effect. We identified them as:

Properties of a causal model exhibiting evidence of the causal doer effect.

1. There exists a causal edge between doing and either of the outcome measures that has a positive coefficient estimate.
2. The strength of this causal edge is larger than all the edges from passive resource use to the outcome measures.
3. The edge(s) between doing and outcome(s) is oriented *from* doing *to* an outcome.

4. RESULTS

We now provide results from the two methods, first the confirmatory analysis and then the exploratory analysis.

4.1 Confirmatory Analysis: Testing a Causal Model Across Multiple Datasets

In order to determine if the causal model discovered on GTech's psychology course data would fit other datasets, modifications to the data were made to ensure that all datasets were comparable. We show in Figure 1 the causal model that was used as a "modified original" causal model, which was in turn then tested on new data. We arrived at the "modified original" model by applying the same causal search algorithm to the original data set – but with the set of variables that were common to both it and the dataset to be tested. Happily, these models are strongly consistent with the original. For instance, when the play variable was removed, the value of the edge from non_activities_pageview to T.Quiz (i.e., non_activities_pageview→T.Quiz) should be adjusted. This adjustment should be equal to the original edge between these variables plus the product of the edges from the two edges that were removed (i.e., non-activites_pageview→play and play→T.Quiz).

Table 2. The causal model that was discovered on GTech's psychology dataset was estimated using data from datasets listed in the first row of the table.

	UMUC Biology	UMUC Info Sci	UMUC Psychology	UMUC Statistics	C@CM	UMUC Biology (sample)	UMUC Info Sci (sample)
#Students	3516	6112	89	61	383	300	300
Chi-square	78.89	18.44	1.04	28.33	14.30	11.92*	3.32*
DOF	2	2	2	2	4	2	2
P-value	0	0	0.59	0	0	0.02*	0.49*

*average of multiple trials with different samples

non_activities_pageview→T.Quiz + (non_activities_pageview→play * play→T.Quiz) = edge's new value

0.0650 + (0.1149*0.0645) = 0.0724

The model estimated the new value for the edge from non_activities_pageview to T.Quiz to be 0.0713, which is consistent with the calculation above.

The causal models in Figure 1 show yellow edges. These edges were originally either unoriented in the representation of the equivalence class or were oriented as bidirected edges. Before we can estimate and test a causal model, we must direct all edges to form a directed acyclic graph. Therefore, before estimating, the undirected and bidirected edges were *arbitrarily* converted into directed edges - and such edges are shown in yellow to caution the user against inferring any directional information from such edges. In Figure 1c, if the edge were directed the opposite way, the coefficient would still be 0.3927. Removing variables such as play and Pretest still allowed for models that show strong doer effect to be discovered, which is consistent with [6].

The structures of the models in Figure 1 were then applied to the other five datasets, and these models were estimated to determine how well the exact causal structure of the "original model" fit the new data. The results of the confirmatory analysis are summarized in Table 2. As was expected, whether the original causal structure fit other datasets was inconsistent. UMUC's psychology dataset fit very well to this causal model having a p-value of 0.59, however, the rest of the p-values from full datasets were low. It is worth noting that the only full data set to fit GTech's psychology course, was another psychology course. UMUC and GTech's psychology courses have the same content (online readings and interactive activities). The differences between these datasets were the population that created the data and the number of variables. GTech's course had all of the variables that UMUC's course had with the addition of the number of videos watched and a pretest. Therefore, once the video watching and pretest variables are removed from GTech's psychology dataset, the same causal model would be expected to be discovered on GTech's and UMUC's psychology data.

In large datasets, e.g., with N > 2000, the goodness-of-fit χ^2 statistic is of limited use, as it not only tests for causal structure, but it also becomes sensitive to small deviations from linearity, or normality, or other parametric assumptions that have little to do with the causal structure. To test whether the statistic is rejecting the model structure or fine-grained violations of the parametric assumptions, we took a random sample of 300 students from each of the UMUC datasets and then re-estimated and tested the model. The smaller biology sample showed a much better fit than its full dataset, however, at a p-value of 0.02, the model is still rejected. The sample from the information science course showed an excellent fit with a p-value of 0.49 and a chi-square value that differs from the

Figure 1. Using subsets of variables from the Georgia Tech Psychology dataset, three causal models were discovered using the PC algorithm and an alpha value of 0.05 as in [6].

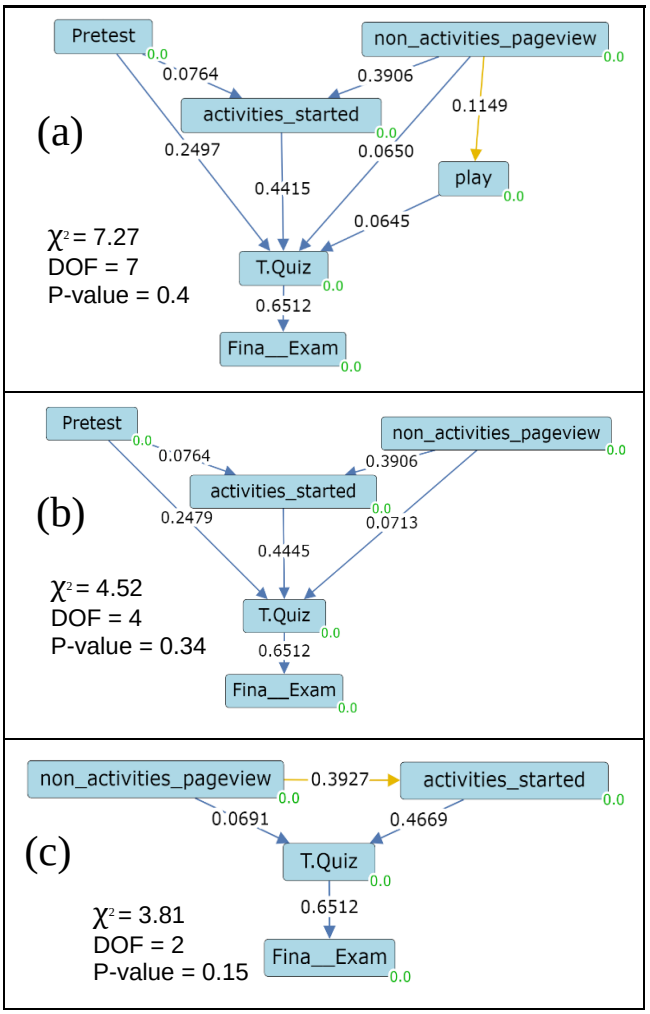
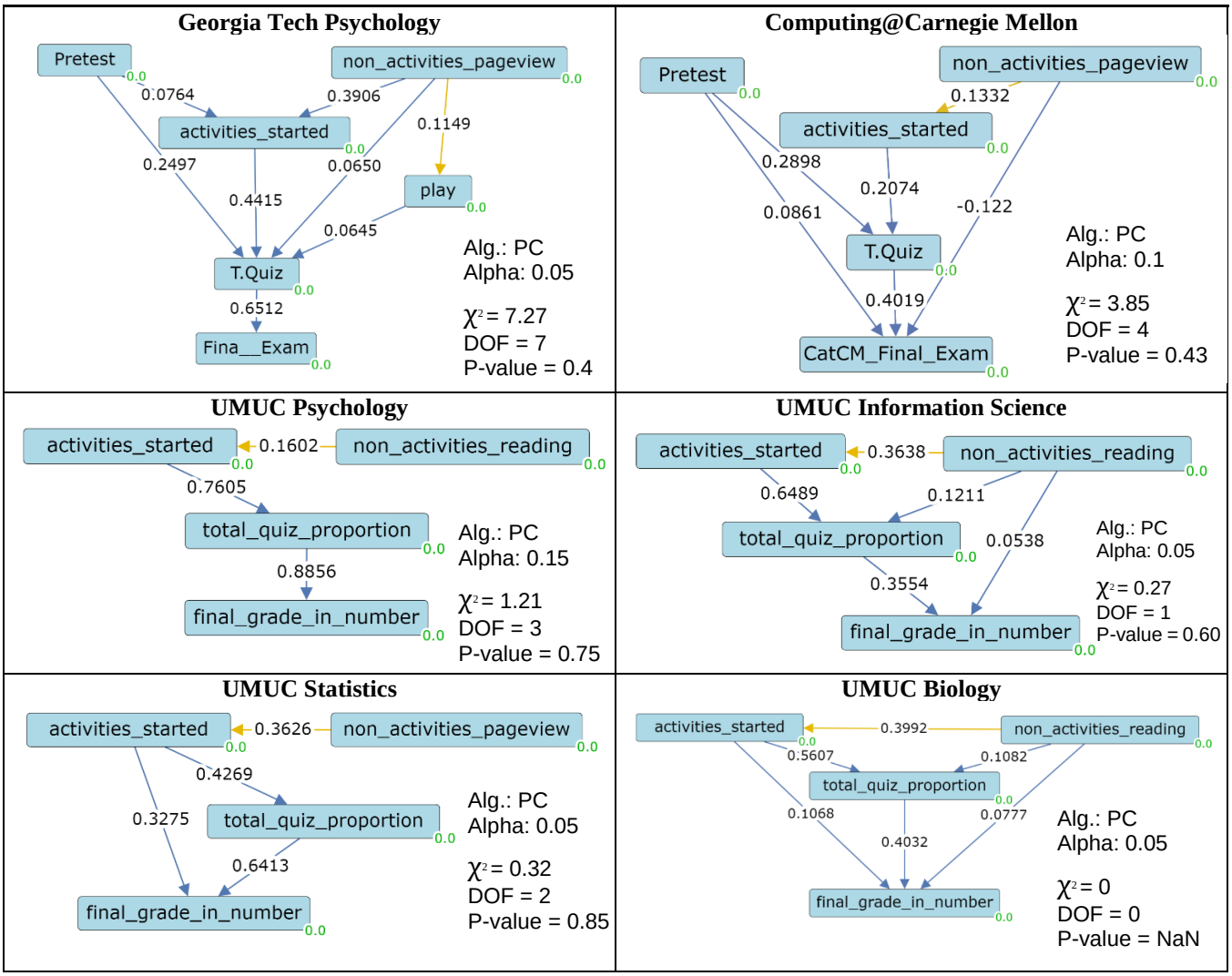


Figure 2. Causal models of various datasets. To the bottom right of each model are the search algorithm and p-value cutoff for searching (alpha) used to discover the model. Below that are the model statistics when estimating the model on the dataset: Chi-square (χ^2), degrees of freedom in the model (DOF), and p-value.



degrees of freedom of the model by only 1.32. We take this to be evidence, although only weak evidence, that the causal structure in the "original model" is reasonably consistent with the measured data. This marginal fit exceeds expectations given the history of difficulty in fitting single models across domains. Given the diversity of the datasets and lack of control between them, any indication of generalizability adds to external validity even though the fit was marginal.

4.2 Exploratory Analysis: Causal Doer Effect Criteria Across Multiple Datasets

The inconsistencies in fitting a single, exact causal model across such diverse datasets are to be expected. A more targeted approach focuses the evaluation on just the variables of interest for assessing the causal doer effect. As described above, we defined three criteria to indicate whether a model provides causal evidence for the doer effect. We searched for causal models on each dataset and then evaluated them by these criteria. Unlike the confirmatory strategy (as shown in Figure 1), where models were discovered on one dataset and estimated on another, these models were discovered and estimated on the same data, as is the norm in causal discovery and as was done previously [6, 12].

Figure 2 shows the results of this analysis. For every dataset we discovered a model that fit the data well (with the exception of Biology, where the model discovered is untestable because it entails no constraints and thus has 0 degrees of freedom). The causal model discovered in [6] was found using the PC algorithm with a p-value cutoff (alpha) of 0.05 for detection of reliable links between variables. This is the algorithm and alpha value that produced a model with largest p-value upon estimation – indicating the model does not significantly deviate from the data and thus is a good one. For the datasets in Figure 2, we also used the PC algorithm with alpha = .05, .1, or .15.

In order to assess the goodness-of-fit of the whole model, we use the p-value of the χ^2 statistic [1]. Unlike the usual logic in hypothesis testing, the p-value in this context uses a null of the specified model. So, a low p-value indicates that we should reject the specified model, while a p-value over .05 indicates that we cannot reject the specified model from the data measured. In general, the χ^2 test is more tolerant of simple models, and simple models are also favorable since they only show the strong, important edges.

The models in Figure 2 were discovered using the same many-tiered prior knowledge as the models in Figure 1 and Table 2. This

prior knowledge assumes that the pre-assessments and weekly/unit assessments were taken before and after the doing and passive activities, respectively. This is an assumption that prohibits causal directionality that violate the temporal order, but it is *not* an assumption that a causal edge exists. That is, the assumption does not guarantee that the algorithms will find any edge between doing and learning. If it does find an edge, then it will be directed from doing to outcome as opposed to vice versa.

Setting these tiers for input in the search algorithms in Tetrad dictates that if a causal link is to be found between variables between temporal tiers, then the directionality of the edge will be from the tier earlier in time to the tier later in time. Again, putting the doing variable in an earlier tier than an outcome variable does not guarantee that Tetrad will find a causal link between the two variables.

We then asked whether the models discovered from each data set satisfy any of the three properties that indicate a causal doer effect as we had listed before. Analyzing Figure 2, all six datasets we used in this paper produced causal models that meet all three criteria of a model with a causal doer effect. For example, C@CM's causal model has a directed edge with a coefficient of 0.2074 from doing to an outcome measure, therefore displaying the first and third properties. The coefficient from the only other resource use variable (non_activities_pageview) was -0.122. The strength of the causal edge is larger than the edge from passive resource use to the outcome in C@CM, thereby showing the second property. The model for UMUC's biology class is not testable as a model, as it has 0 degrees of freedom. Nevertheless, the model along with the estimated coefficients on the edges support all three criteria of a causal doer effect.

5. DISCUSSION

We build off of the work in [6] by providing evidence to suggest that the doer effect is indeed causal. Data from a variety of different online courses (Psychology, Computing, Information Science, Statistics, and Biology) and course use scenarios (MOOCs and for-credit college courses), analyzed with causal discovery algorithms all provide evidence that the doer effect is causal and not just associational.

The correlation between doing and outcome is interesting, but establishing the correlation does not specify whether an intervention on doing would affect outcome. If the doer effect is causal, then modifying learning environments to guide or encourage students to spend more time engaging in interactive activities will result in more learning.

In addition to finding evidence for a causal relationship between doing and learning, we articulated what we hope are useful new methods for discovering and testing for cause-effect relationships across diverse datasets.

For our confirmatory strategy, we tested models discovered in one context on data from another. Finding models that fit a held-out subset of data is protection against overfitting – but it does not mean that those models will fit datasets collected in entirely new contexts, in fact, it is nearly impossible to fit across datasets as diverse as these. Although models developed for educational research seem unlikely to fit in new contexts, we found that features of the causal model of the doer effect found in Georgia Tech data did seem to generalize. The specific model discovered on Georgia Tech's Psychology course data fit extremely well on the data from UMUC's Psychology data. The courses had the same content, but they had different students and were offered in quite different settings (MOOC vs. for-credit course). A marginal fit of the causal model from GTech's Psychology course onto UMUC's Biology and Information Science courses provides some support, albeit limited, for even broader generalization of a specific causal model across different contexts. Given that task has been shown to be nearly impossible, these results are significant even though most fits were marginal.

The inconsistencies of fitting a specific model across contexts is not an indication that a causal doer effect is not present throughout the contexts, it is, however, an indication that an exact model is inconsistently present throughout the contexts. The difficulty of fitting a specific model across contexts led us to reconsider this confirmatory approach. Although a fully specified causal model failed to generalize, it appeared to be due to differences in links between variables that are not relevant to the main question of whether the doer effect is causal. Thus, we developed a method to examine just the key claims of the target theory, in our case, a theory of a causal doer effect. We did so by generating a causal model in an exploratory fashion for each dataset and then evaluating the resulting model as to whether it fit the key criteria for providing evidence of the doer effect.

In all datasets we found that: 1) there was a positive causal edge between active doing and either of the outcome measures, 2) the strength of this causal edge was larger than all edges from passive resource use (reading and watching) to the outcome measures, and 3) the edge(s) between active doing and outcome(s) was oriented from doing to an outcome.

This work provides many possible subsequent inquiries. One area of future work is to test the assumption on the directionality of the causal link between doing and learning outcome. In this paper, we used temporal knowledge to constrain the search algorithms to direct a causal relationship, if one was found between doing and outcome, to be directed from doing to outcome. This temporal knowledge does not make it more likely to find that there is an edge between doing and outcome, it only constrains its orientation. The fact that we found a causal edge between doing and outcome in all six domains is exciting, but we need to investigate further to see if the direction of these edges can be determined from the data or from other plausible assumptions.

When we relax the assumption that doing is temporally prior to outcome, Tetrad is not as likely to orient the edges between doing and learning. Unlike the dataset from Pitt described in the introduction [12], where we were lucky to find a natural "instrument," we do not have a variable in the datasets we studied that is likely to take on that role. Identifying a broader set of variables in this dataset (e.g., by distinguishing counts of error-free doing from errorful doing) or in other datasets may lead such a natural instrument. Particularly useful datasets would involve more student background variables, such as demographics and prior aptitudes, as well as more detailed process data, such as when scrolling makes parts of a web page, whether text, video, or activity, visible or not to a student.

We also hope to perform an experiment to test and hopefully confirm the causal doer effect, much as Rau, et al., [10] did by performing an experiment to test hypotheses generated with causal discovery algorithms on non-experimental data.

6. ACKNOWLEDGMENTS

This work was supported by a National Science Foundation grant (ACI-1443068).

7. REFERENCES

[1] K. Bollen, Structural Equations with Latent Variables, John Wiley & Sons, 1989.

[2] D. Chickering, Optimal Structure Identification with Greedy Search, in *Journal of Machine Learning Research* 3 (2002) 507-554

[3] S. Fancsali, Causal Discovery with Models: Behavior, Affect, and Learning in Cognitive Tutor Algebra, *in Proc. of the 7th International Conf. on Educational Data Mining*, 2014.

[4] S. Karumbaiah, B. Woolf, R. Lizarralde, I. Arroyo, D. Allessio and N. Wixon, Addressing Student Behavior and Affect with Empathy and Growth Mindset, in *Proc. of the 10th International Conf. on Educational Data Mining*, 2017.

[5] K. Koedinger, R. Baker, K. Cunningham, A. Skogsholm, B. Leber and J. Stamper, A Data Repository for the EDM community: The PSLC DataShop., in *Handbook of Educational Data Mining*, Boca Raton, CRC Press, 2010.

[6] K. Koedinger, J. Kim, J. Jia, E. McLaughlin and N. Bier, Learning is not a spectator sport: Doing is better than watching for learning from a MOOC, in *ACM Conf. Learn at Scale*, 2015.

[7] K. Koedinger, E. McLaughlin, J. Jia and N. Bier, Is the Doer Effect a Causal Relationship? How Can We Tell and Why It's Important, in *Conf. Learning Analytics and Knowledge*, 2016.

[8] J. Pearl and E. Bareinboim, External Validity: From Do-Calculus to Transportability Across Populations, *Statistical Science*, vol. 29, no. 4, pp. 579-595, 2014.

[9] D. Rai, J. Beck and I. Arroyo, Causal Modeling to Understand the Relationship between Student Attitudes, Affect and Outcomes, in *Proc. of the 6th International Conf. on Educational Data Mining*, 2013.

[10] M. Rau, R. Scheines, V. Aleven and N. Rummel, Does Representational Understanding Enhance Fluency – or Vice Versa? Searching for Mediation Models, in *Proc. of the 6th International Conf. on Educational Data Mining*, 2013.

[11] M. San Pedro, R. Baker, A. Bowers and N. Heffernan, Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School, *in Proc. of the 6th International Conf. on Educational Data Mining*, 2013.

[12] R. Scheines, G. Leinhardt, J. Smith and K. Cho, Replacing Lecture with Web-Based Course Materials, *Journal of Educational Computing Research*, vol. 32, no. 1, pp. 1-26, 2005.

[13] S. Slater, R. Baker, J. Ocumpaugh, P. Inventado, P. Scupelli and N. Heffernan, Semantic Features of Math Problems: Relationships to Student Learning and Engagement, *in Proc. of the 9th International Conf. on Educational Data Mining*, 2016.

[14] P. Spirtes, C. N. Glymour, R. Scheines, Causation, Prediction, and Search, 2nd edition, MIT Press 2000.

[15] B. Thompson, Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC, US: American Psychological Association. 2014 http://dx.doi.org/10.1037/10694-000.