Searching for Variables and Models to Investigate Mediators of Learning from Multiple Representations

Martina A. Rau
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
marau@cs.cmu.edu

Richard Scheines
Department of Philosophy
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
scheines@cmu.edu

ABSTRACT

Although learning from multiple representations has been shown to be effective in a variety of domains, little is known about the mechanisms by which it occurs. We analyzed log data on error-rate, hint-use, and time-spent obtained from two experiments with a Cognitive Tutor for fractions. The goal of the experiments was to compare learning from multiple graphical representations of fractions to learning from a single graphical representation. Finding that a simple statistical model did not fit data from either experiment, we searched over all possible mediation models consistent with background knowledge, finding several that fit the data well. We also searched over alternative measures of student error-rate, hint-use, and time-spent to see if our data were better modeled with simple monotonic or u-shaped non-monotonic relationships. We found no evidence for non-monotonicity. No matter what measures we used, time-spent was irrelevant, and hint-use was only occasionally relevant. Although the total effect of multiple representations on learning was positive, they also had a negative effect on learning, mediated by a higher error-rate. Our evidence suggests that multiple representations increase error-rate, which in turn inhibits learning. The mechanisms by which multiple representations improve learning are as yet unmodeled.

Keywords

Model search, variable search, mediation, log data, multiple representations

1. INTRODUCTION

Learning processes are complex: many factors influence learning outcomes and the mechanisms by which experimental interventions influence learning are often mysterious. Intelligent Tutoring Systems (ITSs) can easily capture large amounts of data during learning, and combined with sophisticated data mining tools, they have the potential to help understand the mechanisms underlying the effects of successful interventions. Most ITSs are instrumented to collect data on several problem-solving behaviors that might mediate learning, such as error-rate, hint-use, and time-spent [13]. Variables that assess students' problem-solving behaviors have been used to model students' learning [3,8] and to improve ITSs [17]. To make use of the potential of ITS data to gain insights into *why* we see certain learning outcomes, however, we have to overcome difficulties in modeling the mechanisms of learning outcomes. First, we may not adequately understand which variables to use to model these complex relationships. We often assume a linear relationship between measures of learning behaviors and learning outcomes, even though linear relationships may not adequately describe such complex relations [1]. Second, there are a very large number of possible models that describe *how* learning behaviors and learning outcomes relate – how can we know which is the right one? The goal of the present paper is to address both of these important issues using variable search, path analytic modeling, and model search.

Many ITSs use multiple representations to support mathematics learning. Although a vast body of research shows that multiple representations can benefit student learning [2], we know little about the mechanisms that underlie the advantage of learning with multiple representations compared to learning with only a single representation. We investigated the benefits of multiple graphical representations compared to the benefits of a single graphical representation in the context of an ITS – thus enabling us to make use of the rich log data provided in order to investigate the mediating role of student learning behaviors. Specifically, students worked with a Cognitive Tutor for fractions. Cognitive Tutors provide problem-solving tasks and individualized support for students during the learning process [10], and have been shown to lead to significant learning gains in a variety of studies [10,11]. The Fractions Tutor provides error messages tailored to specific misconceptions a student may have. Students can also request a sequence of hints for each step. We chose fractions as the domain for our experiments since fractions instruction typically uses multiple graphical representations such as circles, rectangles, and number lines [12]. Each of these representations emphasizes a different conceptual view on fractions [6] and students need to understand each of these conceptual views [12]. Furthermore, fractions pose a major obstacle for students in the elementary and middle grades [12], such that

understanding mechanisms underlying successful learning is an important educational goal.

We conducted two *in vivo* experiments to investigate the benefits from learning with a version of the Fractions Tutor that uses multiple graphical representations compared to learning with a version of the Fractions Tutor that uses only a single graphical representation. In experiment 1, students worked only with a number line (in the single representation condition), or (in the multiple representations condition) with a variety of graphical representations, including circles, rectangles, and number lines. The representations were relatively static: students could interact with the representations only by entering a number into a text field. The picture updated when the student entered the correct number. In each tutor problem, students solved a fractions problem. For instance, students were asked to add two given fractions and by typing the number of shaded sections into a text field, specifying the numerator of the sum fraction. We crossed these two conditions with a second experimental factor: whether or not students received self-explanation prompts to relate the graphical representations to the symbolic notation of fractions (e.g., ½). For example, students were asked to select "adding the number of shaded sections" to the question of what action with a circle diagram corresponds to adding the numerators using fractions symbols. Results based on an analysis of pretests, immediate posttests, and delayed posttests showed that learners significantly benefited from multiple representations, provided that they were also prompted to self-explain [15].

In experiment 2, we included self-explanation prompts in the single representation condition and in the multiple representations condition. Students in the single representation condition worked either only with a number line, only with a circle, or only with a rectangle. Students in the multiple representations condition received all three graphical representations. In this experiment, the graphical representations were interactive: students could interact with the representations by dragging-and-dropping sections from one representation into another, by using buttons to change number of sections, and by clicking on sections to highlight them. Results based on students' test data confirm the findings from experiment 1: students in the multiple representations condition significantly outperformed students in the single representation condition[1].

We hypothesize that multiple graphical representations result in more successful learning behaviors in the learning phase. We investigated these relationships with the log data that the Fractions Tutor recorded during the learning phases of both experiments. We assume that students who make very few errors, ask for very few hints, and spend very little time per step already have a very good understanding of fractions and will not benefit from working with the Fractions Tutor. On the other hand, inefficient learning such as trial-and-error [4], may manifest themselves in making many errors, asking for many hints, and spending a lot of time per step. We expect that students who show these kinds of unsuccessful learning strategies are not engaging in deep processing of the learning contents and will consequently be less likely to benefit from working with the Fractions Tutor. We hypothesize that the most successful learning behaviors will manifest themselves in moderate levels error-rate, hint-use, and time-spent. This suggests that the relationships between error-rate, hint-use, and time-spent with learning is not simple and monotonic, but rather u-shaped (or inverted u-shaped). We investigated this hypothesis by searching for non-monotonic transformations of our "raw" variables that better predict students' learning than do the raw variables. We then used the best variables in path analysis to investigate the mediating role of error-rate, hint-use, and time-spent on students' benefit from multiple graphical representations.

2. DATA SETS

The analyses presented in this paper are based on the data obtained from the two experimental studies just described. Students in both experiments received a pretest on the day before they started to work with the Fractions Tutor. The day after students finished working with the Fractions Tutor, students received an immediate posttest. About one week after the immediate posttest, students were given an equivalent delayed posttest. In experiment 1, the pretest was a shorter version of the posttests, the posttests included more advanced items which required students to transfer the knowledge covered by the tutoring system to novel situations. In experiment 2, all three tests were equivalent (i.e., they contained the same type of items, but with different numbers).

In experiment 1, 110 6th-grade students worked with either of four versions of the Fractions Tutor (i.e., with a version that included a single graphical representation without prompts, a single graphical representation with prompts, multiple graphical representations without prompts, or multiple graphical representations with prompts). Students worked with the Fractions Tutor for 2.5 hours of their regular mathematics instruction. The average number of errors made per step, the average number of hints requested per step, and the average time spent per step were extracted from the log data obtained from the tutor sessions. Table I shows the means and standard deviations per condition per and per test. Students had a broad range of prior knowledge: the minimum pretest score was 0.00, and the maximum was 1.00. As shown in Table I, students in the MGR condition with prompts outperformed the other conditions both at the immediate and at the delayed posttest. Since in experiment 1, the pretest was not equivalent to the posttests, the pretest scores are not directly comparable to the posttest scores shown in Table I.

[1] This effect was significant for number line items and conceptual transfer on the delayed posttest.

Table II gives an overview of the tutor log data for each condition. While conditions did not differ with regards to error-rate, students who received self-explanation prompts requested fewer hints than students without prompts. Students in the MGR condition with prompts spent relatively more time per step than students in the other conditions, but the differences were small.

	SGR w/o prompts	SGR with prompts	MGR w/o prompts	MGR with prompts
Pretest	0.79 (0.14)	0.70 (0.24)	0.64 (0.25)	0.75 (0.21)
Immediate posttest	0.77 (0.16)	0.70 (0.18)	0.61 (0.23)	0.83 (0.15)
Delayed posttest	0.77 (0.19)	0.74 (0.22)	0.63 (0.21)	0.85 (0.12)

Table I. Means and standard deviations (in brackets) of standardized performance on pretest and posttests from experiment 1 per condition: single graphical representations (SGR) with or without prompts, and multiple graphical representations (MGR) with or without prompts.

	SGR w/o prompts	SGR with prompts	MGR w/o prompts	MGR with prompts
Error-rate	0.27 (0.15)	0.37 (0.17)	0.31 (0.12)	0.34 (0.13)
Hint-use	0.13 (0.31)	0.04 (0.05)	0.19 (0.32)	0.04 (0.09)
Time-spent	10.37 (4.98)	8.47 (6.77)	11.93 (10.18)	13.99 (18.46)

Table II. Means and standard deviations (in brackets) of error-rate (# per step), hint-use (# per step), and time-spent (in sec) per condition: single graphical representations (SGR) with or without prompts, and multiple graphical representations (MGR) with or without prompts.

In experiment 2, 290 4th- and 5th-grade students worked on one of two versions of the Fractions Tutor (i.e., SGR with prompts, or MGRs with prompts) for about 5 hours of their regular mathematics instruction. As in experiment 1, we extracted the average number of errors made per step, the average number of hints requested per step, and the average time spent per step from the log data. Table III summarizes students' performance on each test for each condition in experiment 2. Again, students started with a broad range of prior knowledge: the minimum pretest score was 0.06, and the maximum pretest score was 0.96. Students in the MGR condition perform slightly better than students in the SGR condition at the immediate and at the delayed posttest. Since in experiment 2, the pretest was equivalent to the posttests, we can compare the pretest scores to the posttest scores: students' average scores improved from pretest to the posttests (see Table III). Table IV shows that students in the MGR condition make slightly more errors and ask for slightly more hints, while spending the same time per step as students in the SGR condition. As in experiment 1, the differences between conditions on the log data variables are small.

	SGR	MGR
Pretest	0.54 (0.23)	0.57 (0.21)
Immediate posttest	0.60 (0.23)	0.63 (0.21)
Delayed posttest	0.62 (0.23)	0.67 (0.20)

Table III. Means and standard deviations (in brackets) of standardized performance from experiment 2 per condition and test: single graphical representations (SGR) and multiple graphical representations (MGR).

	SGR	MGR
Error-rate	0.14 (0.07)	0.16 (0.08)
Hint-use	0.04 (0.06)	0.06 (0.09)
Time-spent	0.14 (0.04)	0.14 (0.05)

Table IV. Means and standard deviations (in brackets) of error-rate (# per step), hint-use (# per step), and time-spent (in sec) per condition: single graphical representations (SGR) and multiple graphical representations (MGR).

3. DEFINING VARIABLES WITH WHICH TO INVESTIGATE MEDIATORS

In order to investigate whether a u-shaped, non-monotonic relationship between error-rate, hint-use, and time-spent with students' learning describes the association between problem-solving behavior and learning better than the monotonic relationship, we first conducted a search for a non-monotonic transformation that best predicts students' learning using the data from experiment 2. We used a simple algorithm which computed the "optimal level" of error-rate, hint-use, and time-spent by searching for the highest correlation with learning gains from pretest to the immediate posttest, and from pretest to the delayed posttest, respectively. The algorithm used intervals that varied in size and position. For each interval, we computed a binary variable that for each student indicated whether his/her error-rate (or hint-use, or time-spent) was within the interval or outside the interval. We then computed the correlation of this variable with students' learning gains. For the interval that had the highest correlation with students' learning gains, we identified the mid-point as the "optimum" level of error-rate, hint-use, and time spent. Next, we created two new, non-monotonic predictor variables for error-rate, hint-use, and time-spent, respectively: distance from the optimum, and squared distance from the optimum.

To evaluate whether the non-monotonic variables more accurately predict students' learning, we conducted step-wise regression analyses separately for error-rate, hint-use, and time-spent on both the immediate and the delayed posttests. We entered pretest performance, error-rate, hint-use, or time-spent, and the interaction of pretest performance with error-rate, hint-use, or time-spent as predictors into the regression model. Table V provides a summary of the results from the stepwise regression analyses for error-rate. The regression models with error-rate show that the regression models using monotonic variable explain more variance than the non-monotonic variables. Similarly, the best models with hint-use using the monotonic variable explain more variance than the best models with the non-monotonic variables. The most successful regression models with time-spent take only pretest performance into account; neither the monotonic variable for time-spent nor the non-monotonic variables for time-spent were significant predictors.

		pre	pre + errors	pre + errors + errors*pre
mono-tonic	IP	β_1 = 81*, R^2 = .66	β_1 = .81*, β_2 = -.27*, R^2 = .70	***β_1 = .46*, β_2 = -.48*, β_3 = .19*, R^2 = .71***
	DP	β_1 = .80*, R^2 = .65	***β_1 = .65*, β_2 = -.24*, R^2 = .68***	β_1 = .54*, β_2 = -.38*, β_3 = .13(*), R^2 = .68
distance from optimum	IP	β_1 = .81*, R^2 = .66	β_1 = .76*, β_2 = -.17*, R^2 = .68	***β_1 = .63*, β_2 = -.37*, β_3 = .23*, R^2 = .69***
	DP	β_1 = .80*, R^2 = .65	***β_1 = .78*, β_2 = -.13*, R^2 = .66***	β_1 = .71*, β_2 = -.22*, β_3 = .11, R^2 = .66
squared-distance	IP	β_1 = .81*, R^2 = .66	***β_1 = .76*, β_2 = -.16*, R^2 = .68***	β_1 = .73*, β_2 = -.25*, β_3 = .09, R^2 = .68
	DP	β_1 = .80*, R^2 = .65	***β_1 = .77*, β_2 = -.13*, R^2 = .66***	β_1 = .77*, β_2 = -.12, β_3 = -.01, R^2 = .66

Table V. Regression with error-rate: standardized regression weights and variance explained by each regression model for performance on immediate posttest (IP) and delayed posttest (DP). The best model is displayed in bold-italics. β_1 = pretest (pre), β_2 =error-rate (errors), and β_3 = errors*pre.

In sum, the results from the stepwise regressions show, the non-monotonic variables do not predict performance on the immediate or the delayed posttest better than the monotonic variables do. For that reason, we decided to use the original, monotonic variables of error-rate, hint-use, and time-spent for the subsequent path analytical analyses.

4. HYPOTHESES AND PATH ANALYSIS MODELING

In order to investigate the mechanisms by which the intervention (multiple graphical representations) might have affected learning, we first specified, estimated and tested two path analytical structural equation models [5,20] for each of the two experiments. Structural equation models provide a unified framework within which to test mediation hypotheses, to estimate total effects, and also to separate direct from indirect effects. The models that represented our hypotheses in both experiments were decisively rejected by the data, and in such a case it is not appropriate to use the model to test mediation hypotheses or estimate effects. Our strategy was to use the Tetrad IV program[2] to search for alternative models that are both theoretically plausible and consistent with the data. In this section, we describe the path analytic models that represent our hypotheses, describe the search algorithms we use to search for alternative models, and briefly summarize the results of our search.

4.1 Modeling Our Hypotheses

We hypothesized that multiple representations lead to learning via the three different mechanisms discussed above: error-rate, hint-use or time-spent per step. As each of these variables might also be affected by a student's prior knowledge of fractions, our hypothesis included paths from our intervention variables to each of these mediator variables as well as paths from *pretest* to each of these variables. One of the path models we specified to represent and test our hypothesis about mediation for experiment 1 is shown in Fig. 1.[3] Fig. 2 shows one of the models we specified for experiment 2. Each node in the path models refers to a variable in the data set: *mult_rep* = single vs. multiple representations, *se* = self-explanation prompts, *mr*se* is variable representing a intervention interaction, *pre* = pretest, *time*, *errors*, *hints* = average time spent, # of errors, and # of hints requested per step, *post* = performance on the immediate posttest, *delpost* = performance on the delayed posttest. For both experiments, we hypothesize that pretest performance predicts performance on the immediate and on the delayed posttests, as well as error-rate, hint-use, and time-spent.

[2] Tetrad, freely available at www.phil.cmu.edu/projects/tetrad, contains a causal model simulator, estimator, and over 20 model search algorithms, many of which are described and proved asymptotically reliable in [20].

[3] In path models of this type, also called "causal graphs" [20], each arrow, or directed edge, represents a direct causal relationship relative to the other variables in the model. For example, in Fig. 1 the conditions are direct causes of the mediator variables, but only affect the post-test indirectly through these mediators.

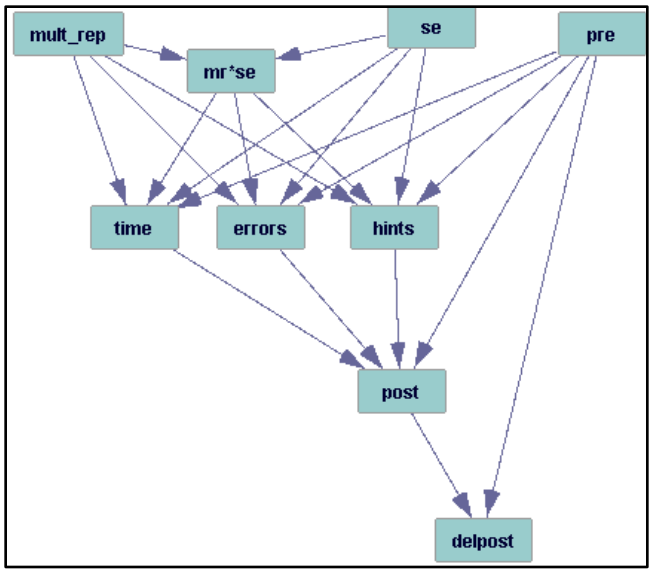

Fig. 1. Path model for experiment 1.

In addition, we predict that in experiment 1, multiple representations (*mult_rep*), self-explanation prompts (*se*), and the interaction between multiple representations and self-explanation prompts (*mr*se*) predict error-rate, hint-use, and time-spent. In other words, we predict that the effects of the intervention variables are entirely mediated through students' learning behaviors. Similarly, for experiment 2, we predict that the effect of multiple representations (*mult_rep*) predicts error-rate, hint-use, and time-spent, which corresponds to a full mediation of the intervention through learning behaviors. Hence, the path model for experiment 2 corresponds to the one shown in Fig. 1, except that self-explanation prompts (*se*), and the interaction between multiple representations and self-explanation prompts (*mr*se*) were not present in experiment 2.

Using normal theory maximum likelihood to estimate the parameters of these models, we find that in each case the deviation between the estimated and the observed covariance matrix is too large to be explained by chance (for the model for experiment 1 in Fig. 1: $\chi^2 = 53.8$, df = 16, $p < .0001$,[4] and for the model for experiment in Fig. 2: $\chi^2 = 59.41$, df = 6, $p < .0001$), thus the models do not fit the data and the parameter estimates cannot be trusted.[5]

4.2 Model Search

To search for alternatives, we used the GES algorithm in Tetrad IV along with background knowledge constraining the space of models searched [7] to those that are theoretically tenable and compatible with our experimental design. In particular, we assumed that our intervention variables are exogenous, that in experiment 1 our intervention variables are causally independent but direct causes of the interaction variable, that the pretest is exogenous and causally independent of intervention, that the mediators are prior to the immediate posttest and to the delayed posttest, and that the immediate posttest is prior to the delayed posttest. Even under these constraints, there are at least 2^{32} (over 4 billion) distinct path models of experiment 1 that are consistent with our background knowledge, and 2^{25} (over 33 million) distinct path models of experiment 2.

The qualitative causal structure of each of these linear structural equation models can be represented by a Directed Acyclic Graph (DAG). If two DAGs entail the same set of constraints on the observed covariance matrix,[6] then we say that they are empirically indistinguishable. If the constraints considered are independence and conditional independence, which exhaust the constraints entailed by DAGs among multivariate normal varieties, then the equivalence class is called a *pattern* [14,20]. Instead of searching in DAG space, the GES algorithm achieves significant efficiency by searching in pattern space. The algorithm is asymptotically reliable,[7] and outputs the *pattern* with the best Bayesian Information Criterion (BIC) score.[8] The pattern identifies features of the causal structure that are distinguishable from the data and background knowledge, as well as those that are not. The algorithm's limits are primarily in its background assumptions involving the non-existence of unmeasured common causes and the parametric assumption that the causal dependencies can be modeled with linear functions.

[4] The usual logic of hypothesis testing is inverted in path analysis. The p-value reflects the probability of seeing as much or more deviation between the covariance matrix implied by the estimated model and the observed covariance matrix, conditional on the null hypothesis that the model that we estimated was the true model. Thus, a low p-value means the *model* can be rejected, and a high p-value means it cannot. The conventional threshold is .05, but like other alpha values, this is somewhat arbitrary. The p-value should be higher at low sample sizes and lowered as the sample size increases, but the rate is a function of several factors, and generally unknown.

[5] We also tested variations of these models in which we added direct paths from the condition variables to the post-test and delayed post-test. These variants are also clearly rejected by our data.

[6] An example of a testable constraint is a vanishing partial correlation, e.g., $\rho_{XY.Z} = 0$.

[7] Provided the generating model satisfies the parametric assumptions of the algorithm, the probability that the output equivalence class contains the generating model converges to 1 in the limit as the data grows without bound. In simulation studies, the algorithm is quite accurate on small to moderate samples.

[8] All the DAGs represented by a pattern will have the same BIC score, so a pattern's BIC score is computed by taking an arbitrary DAG in its class and computing its BIC score.

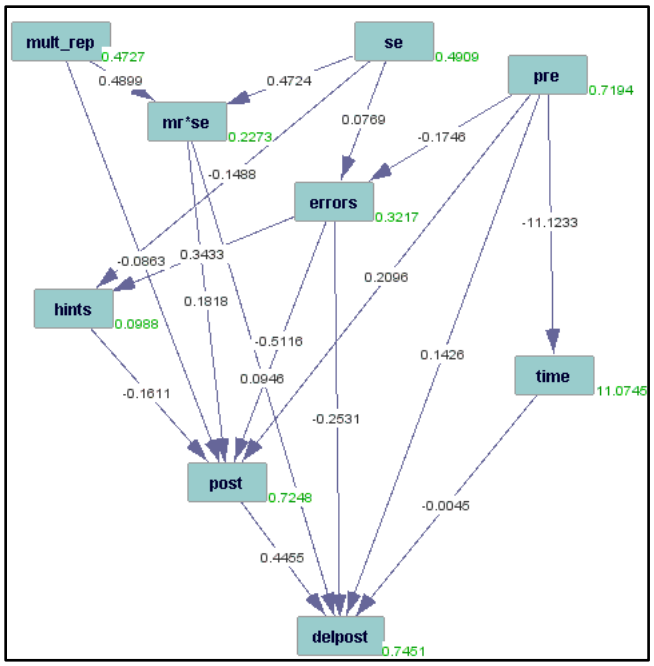

Fig. 2. The model found by GES on data from experiment 1, with parameter estimates included. This model fits the data well: $\chi^2 = 22.11$, df = 19, $p = .29$.

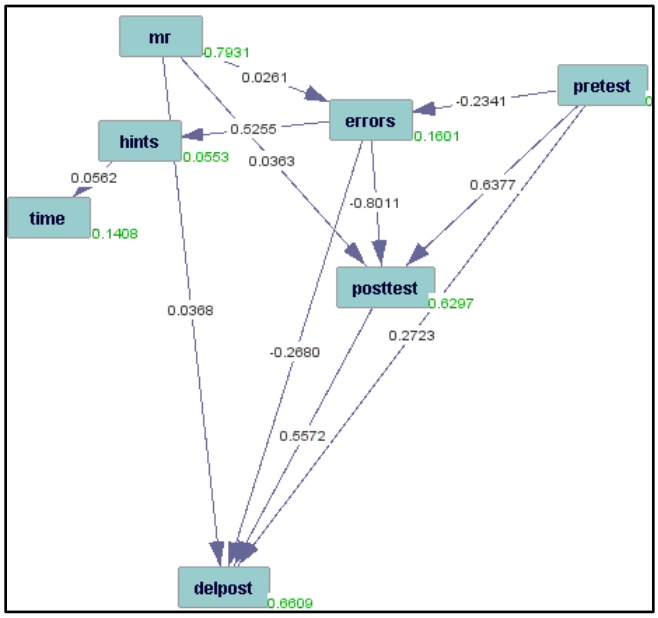

Fig. 3. The model found by GES on data from experiment 2, with parameter estimates included. This model also fits the data well: $\chi^2 = 6.89$, df = 10, $p = .74$.

Fig. 2 shows a model found by GES on the data from experiment 1, with path coefficient estimates included. The model fits the data well ($\chi^2 = 22.1$, df = 19, $p = .28$), and contains a number of interesting properties. For one thing, students with higher pretest scores spend much less time per problem, but none of our intervention variables had any influence on time, and the apparent effect of time spent per step during the learning phase is minimal. Multiple representations had a positive effect on learning, but only when self-explanation prompts were also part of the learning environment.[9] Further, there is no evidence that the positive effect of multiple representations is mediated by either error-rate, hint-use, or time-spent. When *not* combined with multiple representations, self-explanation prompts appear to slightly increase error-rate and thus inhibit learning, but slightly decrease hint-use, which, because they appear to inhibit learning, have an overall positive effect on learning.

Fig. 3 shows a model found by GES for experiment 2 that fits the data very well ($\chi^2 = 6.89$, df = 10, $p = .74$). This model indicates that although multiple representations (*mr*) have a positive direct effect on both the immediate posttest and the delayed posttest, they also have a negative indirect effect on both outcomes through error-rate. Learning with multiple representations seems to cause students to make slightly more errors during learning, possibly because the greater variability in tutor problems leads to higher cognitive processing demands. The higher error-rate during the learning phase seems to have a negative influence on test performance. Note that there are two paths from multiple representations to the posttests in the model in Fig. 3, and that the positive direct effect (a bit over 3 ½ percentage points on both) is larger than the indirect negative effect through errors in both cases (2 percentage points on the immediate posttest and about ½ a percentage point on the delayed posttest).

As in experiment 1, hint-use and time-spent do not discernibly mediate the influence of multiple representations on learning. However, students appear to ask for more hints in response to making more errors, and they spend more time on a problem when they have asked for hints.

5. DISCUSSION

We used data mining in two ways: first to search for mediator variables that are monotonically related to learning outcomes and thus amenable to analysis with standard tools like linear regression and path analysis, and second, to search for causal models of learning that allowed us to investigate mediation relationships and to estimate the total and indirect effects of multiple representations on learning.

Contrary to our expectations, we found that raw measures of error-rate, hint-use, and time-spent were as predictive of learning as any of the non-monotonic variants we searched over. One might suspect that our variable search failed to

[9] The paths from the interaction variable mr*se track the effect of both treatments compared to either one alone or neither. The paths from the individual treatments track the effect of each treatment when the other is absent.

improve on the apparent monotonicity of the raw measures because our sample did not include high prior knowledge students. However, students' pretest scores covered a broad range from very low to very high (see Tables I and III). Although surprising, our findings can be taken as encouraging for the community of educational data mining and for the community of researchers who study ITSs. Analyzing raw measures of error-rate, hint-use, time-spent and learning is much easier than analyzing non-monotonic variants. Furthermore, most research that uses log data obtained from ITSs assumes monotonicity. Our findings do nothing to undermine this practice.

Our findings from path analysis modeling demonstrate the importance of model search. None of our initial hypotheses fit the data, but there are millions of plausible alternatives, only a small handful of which could be practically investigated by hand. Further, estimating path parameters with a model that does not fit the data is scientifically unreliable. Parameter estimates, and the statistical inferences we make about them with standard errors etc., are all conditional on the model specified being true everywhere except the particular parameter under test.

Even if our initial hypotheses had fit the data well, however, it would have been important to know whether there were alternatives that explained the same data. The GES algorithm implemented in Tetrad IV enabled us to find plausible models that fit the data well. The models we found in Fig. 2 and Fig. 3 allow us to estimate and test path parameters free from the worry that the model within which the parameters are estimated is almost surely mis-specified, as is the case for the model in Fig. 1.

Several caveats need to be emphasized, however, lest we give the false impression that we think we have "proved" the causal relationships that appear in the path diagrams shown in Fig. 1 and Fig. 2. First, the GES algorithm assumes that there are no unmeasured confounders (hidden common causes), an assumption that is almost certainly false in this and in almost any social scientific case, but one that is routinely employed in most observational studies.[10] In future work we will apply algorithms (e.g., FCI) that do not make this assumption, and see whether our conclusions are robust against this assumption. Second, although we did include intervention interaction in our model search for experiment 1, and did test for interactions between pretest and mediators in experiment 2, by no means were our tests exhaustive, and by no means can we rely on the assumption that the true relations between the variables we modeled are linear, as the search algorithms assume. Nevertheless, many of the bivariate relationships in the data we modeled appear approximately linear, so the assumption is by no means unreasonable. Third, we have a sample of 290 students, and although that is sizable compared to many ITS studies, model search reliability goes up with sample size but down with model complexity and number of variables, and is overall impossible to put confidence bounds over on finite samples [19].

Nevertheless, our searches for causal models suggest that there are indeed path models that are consistent with our background theory and with the data, and which indicate that multiple representations enhance learning, but not through any detectible mechanism involving error-rate, hint-use, or time-spent. In experiment 1, multiple representations have a positive influence on learning, but have no detectible effect on any of the mediators we measured. In experiment 2, in which interactive graphical representations were part of the intervention, it appears that there is a mediated influence on learning through error-rate, but it is a negative influence. Research from a variety of domains shows that some interventions that decrease performance during the learning phase by increasing the variability of learning tasks result in better long-term retention and transfer performance [9,16]. In other words, interventions that are beneficial in the long run often come at some cost, for instance in the form of lower performance during the learning phase. Our results show that "costs" which become apparent during the learning phase are indeed associated with lower performance also on the posttests. However, we have not yet identified the mediators of the benefits of learning with multiple representations. Given the results from the two experiments described in the present paper, it is unlikely that the advantage of multiple representations is mediated through error-rate, hint-use, or time-spent. Taken together, the results from our two experiments suggest that researchers need to look elsewhere for the cognitive mechanisms by which multiple representations improve students' learning.

The finding that error-rate partially mediates the effect of multiple representations in experiment 2 (but not in experiment 1) is an interesting one as well. One difference between experiment 1 and experiment 2 was that the graphical representations in experiment 1 were not interactive tools, but static pictures that updated when students entered the correct answer into a text field. By contrast, the graphical representations in experiment 2 were interactive: students could drag-and-drop sections from one representation into another and use buttons to partition the representation into fewer or more sections. It is conceivable that interactive representations provide a more direct learning experience for students, which will have a different effect on problem-solving behavior (as, for example, on error-rate) than relatively static representations [18]. There is currently very little research that systematically investigates the impact of interactive versus static representations on students' problem-solving behaviors and consequent learning outcomes. Our findings

[10] Although our data are from a study in which we intervened on intervention, we did not directly intervene on our mediator or outcome variables. Thus these parts of our model are subject to the same assumptions as a non-experimental study.

demonstrate, that the impact of interactive representations is an interesting question to address in future research.

In conclusion, our results are of interest both to the educational psychology literature and to the intelligent tutoring systems literature. First, we can gain insights into the effects of instructional interventions: although multiple representations seem to overall be beneficial, they also seem to lead students to make more errors during the learning phase, which is associated with lower performance on posttests. Second, once we gain knowledge about which learning behaviors are adaptive and which are not, we can use these insights to improve our tutoring systems. For example, perhaps multi-representational ITSs should be designed to prevent errors in the practice and learning phase. Perhaps we can help students avoid practice errors by providing more worked examples, or by designing better error feedback messages. Or perhaps the increase in errors is simply a cost associated with multiple representations that instructors have to live with. These questions and others arose from path analysis and model search and lead almost directly to new hypotheses that we, and hopefully others, will address in future research.

6. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, REESE-21851-1-1121307, and by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420. In addition, we would like to thank Vincent Aleven, Nikol Rummel, Ken Koedinger, the Datashop team, and the students and teachers who participated in our study.

7. REFERENCES

[1] Aleven, A., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. 2003. Help Seeking and Help Design in Interactive Learning Environments. Review of Educational Research, 73(3), 277-320.

[2] Ainsworth, S. 2006. A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16, 183-198.

[3] Anderson, J. R. 1993. Problem Solving and Learning. American Psychologist, 48(1), 1-35.

[4] Baker, R., et al. 2008. Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. Journal of Interactive Learning Research, 19(2), 185-224.

[5] Bollen, K. 1989. Structural Equations with Latent Variables. Wyley, New York.

[6] Charalambous, C. Y., & Pitta-Pantazi, D. 2007. Drawing on a Theoretical Model to Study Students' Understandings of Fractions. Educational Studies in Mathematics, 64, 293-316.

[7] Chickering, D. M. 2002. Optimal Structure Identification with Greedy Search. Journal of Machine Learning Research 3, 507-554.

[8] Croteau, E., Heffernan, N. T., & Koedinger, K. R. 2004. Why Are Algebra Word Problems Difficult? Using Tutorial Log Files and the Power Law of Learning to Select the Best Fitting Cognitive Model. Lecture Notes in Computer Science, 32200, 15-53.

[9] De Croock, M. B. M., Van Merrienboer, J. J. G., & Paas, F. G. W. C. 1998. High versus low contextual interference in simulation-based training of troubleshooting skills: Effects on transfer performance and invested mental effort. Computers in Human Behavior, 14(2), 249-267.

[10] Koedinger, K., & Corbett, A. 2006. Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In: Sawyer, R. K. (ed.).The Cambridge handbook of the learning sciences, pp. 61-77, Cambridge University Press, New York, NY, US.

[11] Koedinger, K., Anderson, J. R., Hadley, W., & Mark, M. 1997. Intelligent Tutoring Goes to School in the Big City. International Journal of Artificial Intelligence in Education, 8, 30-43.

[12] National Mathematics Advisory Panel. 2008. Foundations for Success: Report of the National Mathematics Advisory Board Panel, U.S. Government Printing Office.

[13] Newell, A., & Rosenbloom, P. 1981. Mechanisms of Skill Acquisition and the Law of Practice. In Anderson, J. (ed.). Cognitive Skills and Their Acquisition, Erlbaum Hillsdale NJ.

[14] Pearl, J. 2000. Causality: Models, Reasoning, and Interference. Cambridge University Press.

[15] Rau, M. A., Aleven, V., & Rummel, N. 2009. Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. International Conference on Artificial Intelligence, 441-448.

[16] Rau, M. A., Aleven, V., Tunc-Pekkan, Z., Pacilio, L., & Rummel, N. accepted. How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. To appear in the proceedings of ICLS 2012.

[17] Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. 2007. Cognitive Tutor: Applied Research in Mathematics Education. Psychometric Bulletin & Review, 14(2), 249-255.

[18] Robers, Y. 1999. What Is Different about Interactive Graphical Representations? Learning and Instruction, 9, 419-425.

[19] Robins, J., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform Consistency in Causal Inference, Biometrika, 90, 491 – 515.

[20] Spirtes, P. Glymour, C., Scheines, R. (2000). Causation, Prediction, and Search. 2nd Edition, MIT Press.