Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains

Ran Liu
Carnegie Mellon University
ranliu@cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
koedinger@cmu.edu

As the use of educational technology becomes more ubiquitous, an enormous amount of learning process data is being produced. Educational data mining seeks to analyze and model these data, with the ultimate goal of improving learning outcomes. The most firmly grounded and rigorous evaluation of an educational data mining discovery is whether it yields better student learning when applied. Such an evaluation has been referred to as "closing the loop," as it completes the cycle of system design, deployment, data analysis, and discovery leading back to design. Here, we present an instance of closing the loop on an automated cognitive modeling improvement discovered by Learning Factors Analysis (Cen, Koedinger, & Junker, 2006). We discuss our findings from a process in which we interpret the automated improvements yielded by the best-fitting cognitive model, validate the interpretation on novel data, use it to make changes to classroom-deployed educational technology, and show that the changes lead to significant learning gains relative to a control condition.

1. INTRODUCTION

The field of educational data mining (EDM) seeks to analyze and model the rich process data resulting from educational technology use, with the ultimate goal of improving learning outcomes. However, emphasis on the "educational" aspect of educational data mining has been scarce. While EDM research has produced a rich set of automated techniques and statistical models for educational data, progress remains largely theoretical with respect to their impact on learning efficiency and outcomes. One reason for this is the inclination of researchers to evaluate EDM research primarily for model fits and predictive accuracy rather than for plausibility, interpretability, and generalizable insights. Statistical models and knowledge component models (also known as skill models or Q-matrices) are often assessed based on the ability to correctly predict successes and failures in a set of student response outcomes. Less commonly, models may also be validated on their ability to predict post-test outcomes (e.g., Corbett & Anderson, 1995) or pre- to post-test gains (e.g., Liu & Koedinger, 2015).

Assessing modeling outcomes based on predictive accuracy has much to recommend it, and it is possible in some cases to apply "black box" predictive models to education without much interpretation (e.g., Baker et al., 2006). But, more often than not, interpreting modeling improvements and testing generalization to novel datasets are important steps to advancing practical outcomes, educational theory, or both. The most firmly grounded and rigorous evaluation of a data-driven discovery in EDM is whether it yields better student learning when

applied. Such an evaluation has been referred to as "closing the loop" (Koedinger et al., 2013), as it completes cycle of system design, deployment, data analysis, and discovery leading back to design. The loop is closed through an experimental comparison of a tutoring system redesign, based on the data-driven discovery, to the original tutoring system.

Here, we present an instance of closing the loop on an *automated* discovery by a method of improving knowledge component models called Learning Factors Analysis (Cen, Koedinger, & Junker, 2006). We cycle through a process in which we interpret the model's finding (Koedinger et al., 2012), test the interpretation on novel datasets (Liu et al., 2014), use the interpretation to make changes to classroom-deployed educational technology, and demonstrate that the changes result in significant learning gains.

1.1. KNOWLEDGE COMPONENT (KC) MODELS

Cognitive models map the actual tasks that students engage with during learning to the underlying knowledge that is required to complete those tasks. Cognitive models are an important basis for the instructional design of automated tutors and are important for accurate assessment of learning. Improvements to cognitive models result in better prediction of what a student knows, allowing adaptive learning to work more efficiently.

The work described here uses a simplification of a cognitive model composed of hypothesized knowledge components. A knowledge component (KC) is a fact, skill, or principle required to succeed at a particular task or problem step. We refer to this specialized form of a cognitive model as a knowledge component model (KC model), and it is sometimes referred to alternatively in the literature as a Q-matrix (e.g., Barnes, 2005).

KC models are typically evaluated in conjunction with a statistical model. The statistical model uses the KC model mapping to make inferences about student *knowledge* based on performance across a variety of observable *tasks/items*. In the modeling work discussed here, alternative KC models were evaluated in conjunction with a statistical model called the Additive Factors Model (AFM; Spada & Magaw, 1985; Draney, Wilson, & Pirolli, 1996; Cen, 2009), a generalization of the Rasch IRT model that accommodates learning over time.

1.2. DATA-DRIVEN KC MODEL IMPROVEMENT: PRECURSORS TO LEARNING FACTORS ANALYSIS

Traditional ways of constructing cognitive models involve structured interviews, think-aloud protocols, rational analysis, and labeling by domain experts (Clark et al., 2008). These methods, however, require particularly time-consuming human input. They are also subjective, and prior research has shown that expert-engineered cognitive models often ignore content distinctions that are important for novice learners (Nathan, Koedinger, & Alibali, 2001; Koedinger & McLaughlin, 2010). Difficulty Factors Assessment (e.g., Koedinger & Nathan, 2004) moves beyond experts' intuitions by using a data-driven knowledge decomposition process to identify the problematic elements of a defined task. In other words, when one task is much harder than a closely related task, the difference implies a knowledge demand (at least one KC) of the harder task that is not present in the easier one.

Stamper & Koedinger (2011) illustrated a data-driven method based on the idea of Difficulty Factors Assessment that yields cognitive model improvements. It utilizes freely available educational data and built-in visualization tools on DataShop (Koedinger et al., 2010; http://learnlab.org/datashop), in conjunction with Difficulty Factors Assessment. The method

for human-mediated cognitive model refinement iterates through the following steps: 1) fit the statistical model with the given KC model, 2) inspect learning curve visualizations and best fitting parameter estimates, 3) identify problematic KCs and hypothesize changes to the KC model, and 4) re-fit the statistical model with the revised KC model and evaluate for improvements.

Through manual inspection of the visualizations of a Geometry data set (pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=76), potential improvements to the best existing KC model at the time were identified (Stamper & Koedinger, 2011), using domain expertise to hypothesize about additional (or different) knowledge that might be required on certain items. A new KC model was generated, resulting in significantly better prediction of student performance than the original KC model.

This human-mediated cognitive task analysis method circumvents some of the issues of expert bias by utilizing data to drive KC model refinement and has produced genuine close the loop results (Koedinger et al., 2013). However, it still demands the involvement of significant human effort. In the next section, we describe an advancement that is inspired by Difficulty Factors Assessment but automates the process of generating new KC models based on potential difficulty factors and evaluates them based on predictive accuracy.

1.3. Learning Factors Analysis (LFA)

Learning Factors Analysis (LFA; Cen, Koedinger, & Junker, 2006) was developed to automate this type of data-driven method of cognitive model refinement. LFA performs a combinatorial search process across hypothesized knowledge components (KCs) drawn from existing cognitive models. It alleviates human effort and error by providing an automated way of improving and evaluating cognitive models. It also outputs the cognitive model with the best predictive accuracy in the form of a symbolic model. As such, LFA greatly eases the burden of interpretation even if it does not automatically accomplish it.

Koedinger and colleagues applied the LFA search process across 11 datasets spanning different domains and different educational technologies, all publicly available from DataShop. This automated discovery process improved KC models' fit to data, beyond the best existing human-generated KC models, across all of the datasets (Koedinger et al., 2012). Due to familiarity with the dataset as well as the domain of geometry, we focused on a geometry dataset (Dataset #76 in DataShop) to interpret why the best-fitting KC model discovered by LFA might be better than the previously best-fitting human-generated KC model.

In the present research, we tested the generalizability of this interpretation to a completely novel dataset (new school and non-identical problem content), used the results of these findings to redesign the relevant unit in an intelligent geometry tutor, and assessed the learning improvements driven by the redesign when deployed in a classroom setting.

2. Interpreting Model Discovery and Testing Generalization to Novel Data

A manual KC model comparison between the best-fitting LFA model and the best-fitting human-generated model revealed one critical difference. The LFA model split circle area items into two separate KCs, one involving forward (i.e., find area given radius) calculations

and the other involving backward (i.e., find radius given area) calculations. In contrast, the original human-generated model labeled both types of items as a single KC (i.e., circle area). The LFA-discovered model did not do this type of KC split for other shapes in the dataset, which included rectangles, triangles, and parallelograms. Data on student performance corroborated this finding. Applying domain expertise to interpret this KC split, we hypothesized that the automated model improvement might have captured a hidden difficulty factor: knowing when and how to apply a square root operation for backward circle area items. This is not a difficulty that is present in forward circle area items, or for backward items for other shapes' areas.

To assess the external validity of the interpretation beyond the dataset from which the discoveries were made, we tested it on a novel dataset. The goal was not to simply apply the improved model directly to new data (e.g., as in Feng et al., 2009) or to run an exact replication of the study. Rather, we aimed to test whether the interpretation itself held up within the context of novel data relevant to the interpretation but whose structure (i.e., problem types) and properties (i.e., school district, students) differed from those of the original dataset on which the LFA discovery was made.

For example, the tutor unit for the Geometry Area 1996-1997 dataset had only three unique items associated with the circle area backward (i.e., find circle radius given area) computation. It had no items associated with a square area backward (i.e., find square side length given area) computation. We sought to test the generalizability of the interpretation of the LFA finding in a dataset with substantially more backward circle area and square area items, and with more students from a different school district.

To this end, we investigated the shape-area unit of a much more recent dataset, Motivation for Learning HS Geometry 2012 (geo-pa) (Dataset #748 on DataShop). This dataset is an excerpt from regular classroom use of a Geometry Cognitive Tutor (Ritter et al., 2007) by 82 HS students (10th graders) with a total of 72,404 student problem steps. It contains similar shape area modules and questions as the original dataset but has many more (49) unique backward circle area items. It also contains many (57) unique backward square area items. This makes it possible to validate (i.e. by investigating performance on forward and backward circle area and other shape-area items) and generalize (i.e. by investigating performance on forward and backward square area items) our interpretation of the original LFA discovery.

An exploratory analysis of the 2012 dataset reveals a substantially higher proportion of correct first attempts at forward circle (0.89) and square (0.86) area items than backward circle (0.72) and square area (0.59) items, respectively. To validate the specificity of the square root interpretation, we also investigated performance on backward vs. forward items that do not require a square root application. These included other shapes' area calculations, circle circumference items, and square perimeter items. Results of this exploratory analysis are summarized in Figure 1. Except rectangle area items, the differences in forward and backward item performances were specific to cases where the backward calculations required the application of a square root (circle area and square area).

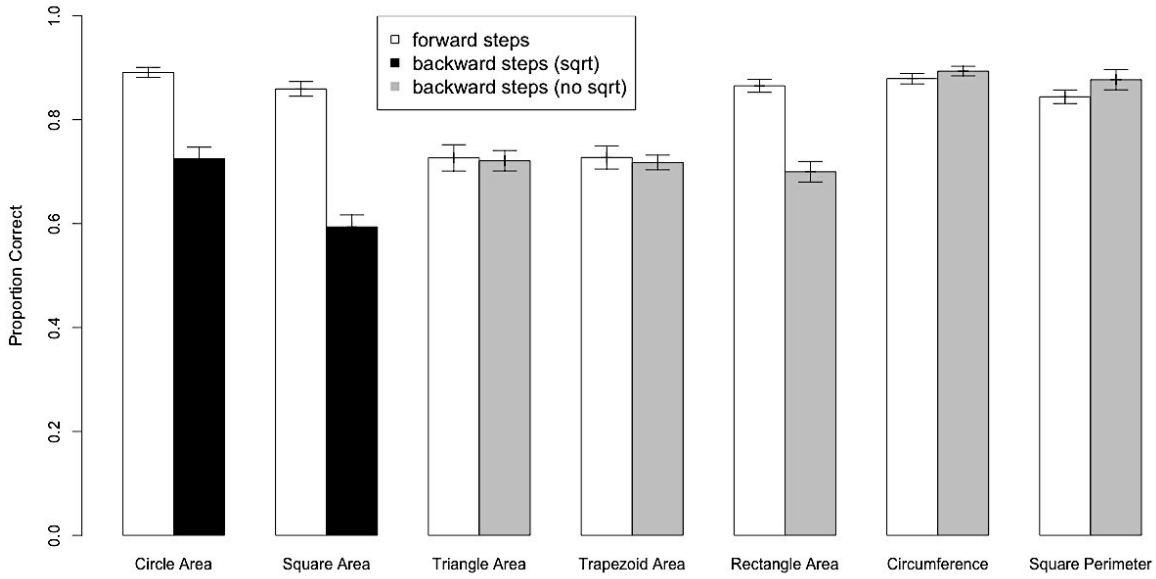


Figure 1: Exploratory analyses investigating the generalization of our hidden square root difficulty hypothesis. Average proportion correct on first attempts at geometry area items, grouped by shape/computation and color-coded based on whether the problem step requires a forward strategy (white), a backward strategy that requires a square root calculation (black), or a backward strategy that does not require a square root calculation (grey)..

To further examine the robustness of our square root hypothesis, we made a KC model comparison. The hypothesis-driven KC model (SQRT SKILL CIR-SQ DISTINCT, 58 KCs) distinguishes forward and backward (F/B) circle and square area items but does not make this F/B distinction for other shapes. We compared this to a KC model that makes no F/B distinctions for *any* shapes (a single 'area' KC for each shape; ALL SHAPES F-B MERGED, 56 KCs). This KC model is analogous to the original human-generated model for the Geometry Area 1996-1997 dataset in which LFA discovered the circle area KC split. To test the specificity of the square root hypothesis, we also made a comparison with a KC model that makes F/B distinctions for *all* shape-area items (ALL SHAPES F-B DISTINCT, 66 KCs).

The models were evaluated using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and 10-fold cross validation. Due to the random nature of the folding process, we repeated each type of 10-fold cross validation (item-stratified and student-stratified) 20 times and calculated the RMSE on each run, as has been done in previous work to handle this run-to-run variability (Koedinger et al., 2012). In Table 1, we report the average root mean square error (RMSE) values across 20 runs each of 10-fold item-stratified and 10-fold student stratified CV. The SQRT SKILL: CIR-SQ DISTINCT model performs best by BIC, item-stratified, and student-stratified CV measures. Consistent with our previous work comparing machine-discovered models to baseline models (Koedinger et al., 2012), we focus on itemstratified cross validation as the primary metric, because we are concerned with improving cognitive tutors. Item stratified cross validation corresponds most closely with a key tutor decision of selecting the next problem type. Furthermore, the BIC measure concurs with the item-stratified cross validation results in suggesting that the SQRT SKILL CIR-SQ DISTINCT model is the best-performing model.

Table 1: Comparison between prediction accuracies of the four hypothesis-driven KC models, evaluated using AIC, BIC, and both item-stratified and student-stratified 10-fold cross validation. Results are reported as the average root mean-squared error (RMSE) values across twenty runs of 10-fold cross validation. The best-performing model, by each of the measures, is bolded.

Model Name	# KCs	AIC	BIC	Item-Stratified Cross Validation RMSE	Student-Stratified Cross Validation RMSE
ALL SHAPES: F-B MERGED	56	20992	22652	0.28208	0.28702
ALL SHAPES: F-B DISTINCT	66	**20839**	22670	0.28104	0.28588
SQRT SKILL: CIR-SQ DISTINCT	58	20857	**22551**	**0.28087***	**0.28584**

The superior performance of SQRT SKILL CIR-SQ DISTINCT over ALL SHAPES F-B MERGED (on all measures) supports and extends the original LFA finding that splitting F/B on circle and square area items is better than leaving the KCs merged. Notably, SQRT SKILL CIR-SQ DISTINCT even performs better, by cross validation and BIC measures, than the ALL SHAPES F-B DISTINCT, the KC model that contains the same F-B distinctions for circle and square but with more fine-grained distinctions for other shapes. This validates the specificity of the square root operation hypothesis.

Thus, there is good evidence that treating forward and backward calculations as distinct KCs specifically for circle and square area items but not other shape-area items predicts student learning best. This demonstrates that our interpretation of the initial automated cognitive modeling discovery generalizes to both novel students and novel problem types.

3. Closing the Loop: Tutor Redesign and Deployment in the Classroom

By isolating improvement in an interpretable component of student learning, elements of instructional design can be modified to more efficiently address student learning. An improved cognitive model can be used in multiple possible ways to redesign a tutor (Koedinger et al., 2013). These include adding or deleting KCs that are tracked in knowledge tracing, creating new tasks to scaffold difficulties, adding/changing feedback or hint messages, and resequencing (positioning problems requiring fewer KCs before ones needing more).

We created an intelligent tutoring system covering the shape-area unit and resembling the geometry tutoring system that had produced the original "discovery" dataset. This was the control tutoring system. Based on the interpreted KC model improvements discovered by LFA, we created a redesigned tutoring system that included separate knowledge-traced KCs differentiating forward and backward circle and square area items. For other shapes, the forward and backward area items remained merged within single shape-area KCs for each shape. The KC differentiation for circle and square area items leads to changes in knowledge tracing which, in turn, lead to different amounts of practice on forward vs. backward items. Students using the revised tutor should receive increased practice on backward circle and

square area items relative to items tagged with other KCs. We also made small modifications to the interface and hints given on backward circle and square area items to better scaffold application of the square root.

3.1. METHODS

3.1.1. Design of Geometry Tutoring Systems

In developing this study, a key initial decision was what the control condition should be. While a natural choice would have been the original tutor that generated the dataset on which LFA's discovery was made, it was not possible because we did not have access to the tutoring technology used to generate the original dataset. Thus, we re-created a tutor that was as similar as possible to the original tutor using freeware called Cognitive Tutor Authoring Tools (Aleven et al., 2006). This tutor was deployed for the "control" condition of our study.

The control tutoring system contained problem content extremely similar to the system that originally generated the original dataset. It included problems in which students had to apply shape-area formulas to triangles, rectangles, squares, circles, and parallelograms. The problem base for this tutoring system included fewer backward circle area items (9) than forward circle area items (24), just as in the original discovery dataset (3 backward circle area items; 19 forward circle area items). There were an equal number of forward and backward square area items, none of which were present in the original discovery dataset. These were included to fairly test the redesign around the square root difficulty interpretation beyond only circle area problems. For all shapes, forward and backward area items were merged into a single KC. That is, there was one KC for each shape (e.g., triangle area, circle area).

We then modified this tutor based on our interpreted LFA discovery to create the "redesigned" condition of the study. The redesigned tutoring system had forward and backward KCs split in knowledge tracing for only circles and squares.

One of the necessary differences between the control and redesigned tutoring systems was the number of skills tracked (the redesigned system had two additional skills, backward circle area and backward square area). That is, students would likely have to complete more problems overall in the redesigned tutor to reach mastery on all skills. Ideally, we would have handled this issue by controlling for time on tutor. We proposed this to the teachers with the caveat that this meant some students would not have a chance to complete all skills to mastery. The teachers, however, wanted every student to have an opportunity to complete the tutor to mastery and were only willing to participate in the study under this condition.

Thus, we adjusted initial parameters to try to equate for overall time-on-tutor with the control tutoring system. The learning rate parameters for the split forward/backward circle and square area KCs in the redesigned condition[1] were slightly higher than those for the merged circle and square area KCs for the control condition. The parameters for the remainder of KCs were

[1] BKT parameters. Control condition (circle area, square area KCs): $P(L_0)=0$, $P(L_T)=0.03$, $P(G)=0.2$, $P(S)=0.2$. Redesigned condition (circle & square area forward & backward KCs): $P(L_0)=0$, $P(L_T)=0.05$, $P(G)=0.2$, $P(S)=0.2$.

the same between the conditions. Since knowledge tracing depends so strongly on the actual performance of students, though, there is no way to ensure that a static set of parameters can yield the same amount of total practice between conditions for every sequence of corrects/incorrects. Thus, we also conducted post-hoc analyses of the process data to quantify actual time-on-tutor and problems completed and ensure that there were no significant differences between conditions.

Table 2. Summary of differences between the control and redesigned tutors.

	GUI	Hints	Knowledge-Traced Skills	BKT Parameters
Control Tutoring System	No scaffolding for square root.	First-level hint for backward circle and square area items did not contain any mention of the square root. All other hints were the same between conditions.	Parallelogram Area Triangle Area Rectangle Area Square Area Circle Area	Same parameters [$P(L_0)$=0, $P(L_T)$=0.03, P(G)=0.2, P(S)=0.2] for all skills.
Redesigned Tutoring System	Scaffolding to draw attention to the need to apply a square root (see Appendix 1).	First-level hint for backward circle and square area items did contained *additional text to prompt application of the square root.* All other hints were the same between conditions.	Parallelogram Area Triangle Area Rectangle Area *Forward* Square Area *Backward* Square Area *Forward* Circle Area *Backward* Square Area	Same parameters as control condition except for a higher $P(L_T)$=0.05 for circle and square area skills (to try to control for overall time on tutor).

The redesigned tutoring system also included a first-level hint that targeted the application of the square root on backward circle and square area items. For example, the redesigned tutor's first-level hint for backward circle area items states: "The formula for the area of a circle is A=R^2*pi (radius squared times pi). You can find R by taking the SQUARE ROOT of R^2." The control tutor's corresponding first-level hint states only the first part: "The formula for the area of a circle is A=R^2*pi (radius squared times pi)." The remaining hint levels were the same for the two conditions. Finally, the redesigned tutoring system included additional interface scaffolding to draw attention to the square root function. Appendix 1 shows a comparison of the interface, between conditions, for the same backward circle area problem screen.

Both versions of the tutoring system used four-parameter Bayesian Knowledge Tracing (Corbett & Anderson, 1995) to determine when students had mastered each KC. Students stopped receiving practice problems that only required a KC that they were estimated to have mastered already (95% probability) based on Bayesian Knowledge Tracing.

3.1.2. Classroom Experiment

115 students across six geometry classes at a public high school in the greater Pittsburgh region were enrolled in the study. They were randomly assigned, at the individual student level, to either the control condition (57 students) or the redesigned condition (58 students). Of the total enrolled students, 91 completed both the pre-test and post-test and were present for at least one class period of tutor use (49 students in the control condition, 42 students in the experimental condition). We performed a chi-square test to ensure that there was no differential attrition based on condition. Based on this test there was no relationship between condition and the resulting number of students who completed the full study, X^2 (1, N = 91) = 0.54, p = 0.46. Because the results of interest were primarily pre-test-post-test gains as the result of condition differences in tutor experience, we restricted our analyses to these 91 students.

The study took place across five consecutive school days. On the first day, students completed a pre-test consisting of 12 questions. Of these questions, three were backward area problems that required the application of a square root (two were "given area, find radius" circle problems, and one was a "given area, find side length" square problem). There were two pre-test forms alternately given to adjacent students to deter cheating. The problems were identical across the two forms but problem order was reversed. On days 2, 3, and 4, students moved at their own pace through the shape-area geometry tutor to which they were randomly assigned. They completed problems for each knowledge-traced KC until the KC was estimated to be mastered based on Bayesian Knowledge Tracing. On the final day, students completed a post-test consisting of 12 questions. The problems were identical to those on the pre-test forms but with slightly different numbers.

3.2. RESULTS

To ensure that students did in fact learn from pre- to post-test across both conditions, we first ran a one-way repeated measures ANOVA with test (pre vs. post) as the factor and test score as the outcome. Results indicated that students' test scores were significantly different at the two testing points [$F(1,90)=61.06$, $p < 0.0001$]. This was driven by post-test scores (mean = 0.78) being significantly higher than pre-test scores (mean = 0.61).

A one-way ANOVA showed pre-test scores were not significantly different depending on the condition students were in [$F(1,89)=0.793$, $p=0.376$]. Post-test scores *were* significantly different depending on the condition students were in [$F(1,89)=5.04$, $p = 0.027$, Cohen's d = 0.47]. Table 3 and Figure 2 summarize the pre-test and post-test results by condition.

We ran a linear regression with pre-test total scores and condition as predictors of post-test total scores. Unsurprisingly, pre-test scores significantly predict post-test scores ($\beta = 0.59$, $p < 0.001$). Controlling for pre-test scores, condition also significantly predicts post-test scores ($\beta = 0.09$, $p = 0.027$). The redesigned condition had significantly higher post-test scores (mean = 0.85) than the control condition (mean = 0.73), even when controlling for their pre-test scores (redesigned condition mean = 0.63; control condition mean = 0.58).

Table 3. Mean and standard errors of the pre-test and post-test scores for each condition.

	Pre-test Score	Post-test Score
Control Condition	0.58 (0.04)	0.73 (0.04)
Redesigned Condition	0.64 (0.05)	0.85 (0.03)

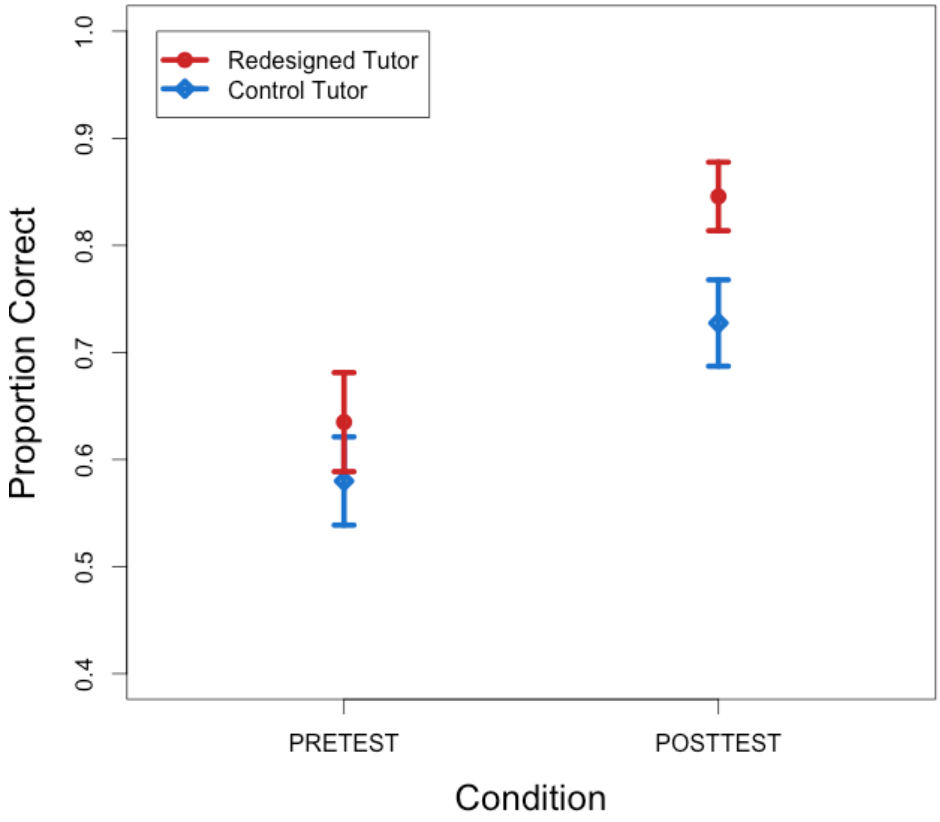


Figure 1: Summary of pre- to post-test learning gains by condition. The redesigned tutoring system yielded higher post-test outcomes than did the control tutoring system, after controlling for pre-test scores.

The changes in the redesigned condition were created to help students overcome the hidden difficulty of knowing when and how to apply the square root operation on certain backward circle area items. Thus, we would predict that any advantages exhibited by the redesigned condition in their post-test outcomes should be primarily driven by better performance on problems requiring backward circle or square area calculations. We tested this prediction by examining condition differences on a post-test subscore composed of only the three problems requiring a square root application. The effect of condition on this subscore ($\beta = 0.13$, $p = 0.034$) was significant in a regression that again controlled for pre-test scores. We then examined condition differences on the post-test subscore composed of all the problems that did not require a square root application, and condition was not a significant predictor of this subset of problems ($\beta = 0.06$, $p = 0.085$) when controlling for pre-test score. This suggests that

the relatively greater post-test improvements in the redesigned condition were predominantly driven by performance on the problems that required applying a square root.

Total time-on-tutor. One of the anticipated effects of the redesigned tutoring system was increased practice on backward circle and square area items. Without any other adjustments, this would lead to the redesigned condition receiving overall more practice than the control condition. To control for this, we had adjusted the circle and square area parameters between conditions so that the redesigned condition would receive less practice on each of the split KCs (e.g., "find area given radius"; "find radius given area") for circles and squares than the control condition would receive on a merged KC (e.g., "circle area"). Since Bayesian Knowledge Tracing depends so strongly on the actual performance of students, though, it is not possible for any set of fixed parameters to yield the same amount of total practice between conditions for every sequence of corrects/incorrects. Thus, we conducted some post-hoc analyses on actual practice time and problems completed across the two conditions.

Figure 2: Violin plots showing the kernel density distributions and boxplots for total time-on-tutor in the two experimental conditions.

The distributions for total time-on-tutor for the two conditions were not normal based on Wilkes-Shapiro Tests (W=0.77, p<0.001 for the Control condition; W=0.93, p=0.01 for the Redesigned condition). Considering this, we conducted a Mann-Whitney U test, which is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. Mean times-on-tutor in the Control and Redesigned groups, respectively, were 105.4 and 111.6 minutes. The distributions in the two groups did *not* significantly differ (Mann–Whitney $U = 797$, $n_1 = 42$, $n_2 = 49$, $p = 0.07$).

Furthermore, to ensure that any marginal differences in time-on-tutor did not explain the pre- to post-test gains, we also conducted a linear regression using log-transformed total time-on-tutor (to correct for normality), in addition to pre-test scores and condition, to predict post-test scores. Condition remained a significant predictor ($\beta = 0.084$, $p = 0.035$) of post-test scores, and total time-on-tutor was not a significant predictor of post-test scores ($\beta = 0.015$, $p = 0.824$).

Interaction between condition and practice. Based on the theory behind the redesign, we expected the redesigned condition to have gotten more practice on backward circle and square

area items than the control condition. We also expected the control group to have gotten relatively more practice on the corresponding forward circle and square area items. In an ANOVA using condition (control, redesigned) and item type (forward, backward) to predict total circle and square area problems completed, we did observe a significant interaction between condition and item type [$F(1,1) = 29.91$, $p < 0.0001$]. The redesigned condition completed more backward circle and square area items (mean = 13.4) than the control condition (mean = 5.75) did, whereas the control condition completed more forward items (mean = 12.3) than the redesigned condition (mean = 11.5) did. Although the differences in total circle and square area items completed were not perfectly controlled for by the initial parameter adjustments, the direction of interaction between condition and practice were aligned with the hypothesized changes expected from the tutor redesign.

4. DISCUSSION

In the present work, we interpreted and acted upon a particular automated discovery (Koedinger et al., 2012) produced by Learning Factors Analysis (LFA) in a geometry dataset covering shapes' areas and perimeters. The automated discovery split circle area items into two separate KCs, one characterizing forward calculations (e.g., find area given radius) and the other for backward calculations (e.g., find radius given area). Since it made this split specifically for circle area items but not for other shape-area items, we interpreted the split as indicating a difficulty applying and/or knowing to apply the square root. To validate this interpretation on novel data, we conducted exploratory analyses on error rates in a novel dataset that additionally contained square area items (which are relevant to the interpretation but were not present in the original discovery dataset). The results of these analyses largely confirmed that application of the square root was a hidden difficulty factor. We also constructed an interpretation-aligned KC model (KCs split for shapes requiring a square root application in the backward computation). We showed that this KC model fit the novel dataset better than both a more conservative (a single KC representing each shape-area formula) and a more liberal (KCs split forward vs. backward for every shape-area formula).

Generalizing the interpretation of the automated discovery to novel data provided a more robust validation of its plausibility and relevance to future pedagogy. With this in mind, we continued on to "close the loop" by redesigning and deploying a tutoring system based on the interpretation. For a controlled comparison, we designed a baseline "control condition" geometry tutoring system to mimic, as closely as possible, the system that generated the original dataset. We then redesigned several features of the control geometry tutoring system to improve scaffolding and practice on the items that required a square root application. This served as the "redesigned condition" tutor. Finally, we assessed the learning improvements for each condition and showed that the redesigned tutoring system led to higher post-test outcomes when controlling for pre-test scores.

To our knowledge, this is the first-ever demonstration of experimentally closing the loop on an *automated* KC model improvement. It is also one of relatively few efforts to "close the loop" on educational data mining discoveries more generally. Our process underscores the importance of considering the interpretability and actionability of educational data mining results if the ultimate goal is to move beyond predictive accuracy to improve learning outcomes, learning theory, or both. The interpretation of the LFA discovery was critical to a number of the redesign changes we made that led to improved learning outcomes. Though it would have been possible to simply "copy" the newly-discovered KC model, resulting only in

the split of forward vs. backward circle area items for knowledge tracing, we made a number of additional changes that would not have been possible without the interpretation. Specifically, we would not have known to make the same forward vs. backward split for square area problems, and we wouldn't have known how to modify the hints or introduce scaffolding to the interface to target the application of the square root skill.

There are many interesting educational data mining discoveries that have resulted in advancements that are difficult or impossible to interpret. A key aspect of LFA that we believe led to interpretable results is the up front, human-in-the-loop aspect of its feature selection. LFA derives new variables from existing, expert-labeled variables using simple split, merge or add operators. It starts with independent variables that map to clearly defined constructs, since they are based on human-labeled KCs. Incorporating some human time and effort into defining and labeling these independent variable features up front can greatly improve the interpretability and actionability of subsequent data mining efforts. The fact that methods like LFA initially requires human input has been cited as a limitation (e.g., González-Brenes & Mostow, 2012) in in arguments favoring purely automated methods of discovering KC models. We argue, however, that it is precisely this "human-in-the-loop" feature that leads the results of such modeling efforts to be interpretable. There have been a number of recent efforts to fully automate the process of discovering KC models (González-Brenes & Mostow, 2012; Lindsey, Khajah, & Mozer, 2014). These methods have much to recommend, dramatically reducing demands on human time and producing competitive results in predictive accuracy. However, the resulting KC models of these efforts have not been interpreted or acted upon with respect to improving instruction.

Other modeling efforts that have included a "human-in-the-loop" component, like Ordinal SPARFA-Tag (Lan et al., 2013), have yielded considerably more interpretable cognitive models than alternative methods. Although any final interpretations of modeling efforts are necessarily made by humans, methods like LFA and Ordinal SPARFA-Tag greatly improve the likelihood of generating sensible resulting models by incorporating the human effort up front. In fact, comparing the original SPARFA model (Lan et al., 2014), which only incorporates concept tags post-hoc, to Ordinal SPARFA-Tag, which incorporates domain expert concept tags in the model development process up front, shows that the latter model results in more interpretable cognitive models.

A method such as LFA, however, does have limitations in degree to which it can produce KC model discoveries. It does require an initial KC model (which does not always exist for educational data) and hypothesized difficulty factors that can be used to attempt splits and merges. Because of these features, it cannot produce large structural changes to KC models. We encourage future work that builds upon the strengths of an algorithm such as LFA but allows for automated discoveries of larger structural changes to KC models that remain interpretable and actionable.

We have highlighted many advantages of emphasizing interpretability in educational data mining. It must be acknowledged, however, that there are automated discoveries that can potentially be applied directly to improve learning without an interpretation. Developing automated detectors of student behaviors or affective states is a class of educational data mining within which models have led to successful interventions without interpreting what exactly the detectors have learned (see D'Mello et al., 2014 for a review). For example, Baker and colleagues (2006) used a machine-learned detector (Baker et al., 2004) of "gaming the system" behavior to determine when to intervene, via an animated agent, while students

engaged with an intelligent tutoring system. They showed that students who received detector-based interventions exhibited less gaming and comparatively better learning outcomes than those who did not receive such interventions.

However, even educational data mining results that could be experimentally applied without an interpretation are rarely brought back to the classroom. Recent progress in the field has remained largely theoretical with respect to impact on learning outcomes and efficiency. We encourage the field to devote more attention towards model interpretability and efforts to experimentally close-the-loop.

Towards identifying the truly robust EDM findings that are more likely to yield genuine learning gains, we note that our confidence in the interpreted LFA discovery was greatly supported by its generalization to a novel dataset and to novel problem types. Thus, moving beyond predictive accuracy to assess models based on generalization to novel data (from entirely different populations and contexts), and on predicting external outcomes such as pre- to post-test gains, will help the field produce more interpretable, explanatory, and/or actionable outcomes.

5. ACKNOWLEDGEMENTS

This research was supported in part by the Institute for Education Sciences (training grant #R305B110003 to RL) and the National Science Foundation (#SBE-0836012). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES or the NSF. We thank Mimi McLaughlin for help with data analysis and Justin Aglio for help recruiting classrooms for the study.

6. REFERENCES

ALEVEN, V., SEWALL, J., MCLAREN, B.M., AND KOEDINGER, K.R. 2006. Rapid authoring of intelligent tutors for real-world and experimental use. In *Proceedings of the 6th ICALT*. IEEE, Los Alamitos, CA, pp. 847-851.

BAKER, R. S., CORBETT, A. T., AND KOEDINGER, K. R. 2004. Detecting student misuse of intelligent tutoring systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

BAKER, R. S., CORBETT, A. T., KOEDINGER, K. R., EVENSON, S. E., ROLL, I., WAGNER, A. Z., NAIM, M., RASPAT, J., BAKER, D. J., AND BECK, J. 2006. Adapting to when students game an intelligent tutoring system. In *Proc Int Conf on Intelligent Tutoring Systems*, 392-401. Jhongli, Taiwan.

BARNES, T. 2005. The Q-matrix method: Mining student response data for knowledge. In *Proceedings of AAAI 2005: Educational Data Mining Workshop*, 978–980.

BERNACKI, M. 2012. Motivation for learning HS geometry 2012 (geo-pa). pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=748

CEN. H. Generalized learning factors analysis: improving cognitive models with machine learning. Doctoral Dissertation, Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2009.

CEN, H., KOEDINGER, K. R., AND JUNKER, B. 2006. Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems,* 164-175. Berlin: SpringerVerlag.

CLARK, R. E., FELDON, D., VAN MERRIËNBOER, J., YATES, K., & EARLY, S. 2008. Cognitive task analysis. In Spector, J. M., Merrill, M.D., van Merriënboer, J., & Driscoll, M.P. (Eds.), Handbook of research on educational communications and technology (3rd ed.). Mahwah: Lawrence Erlbaum.

CORBETT, A. T. AND ANDERSON, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling & User-Adapted Interaction*, 4, 253-278.

DRANEY, K., WILSON, M., & PIROLLI, P. (1996). Measuring learning in LISP: an application of the random coefficients multinomial logit model. In: Engelhard G, Wilson M, eds. *Objective Measurement III: Theory into Practice.* Norwood, NJ: Ablex.

D'MELLO, S., BLANCHARD, N., BAKER, R. OCUMPAUGH, J., AND BRAWNER, K. 2014. I feel your pain: a selective review of affect sensitive instructional strategies. In Sottilare R, Graesser A, Hu X, & Goldberg B (Eds.), Design Recommendations for Adaptive Intelligent Tutoring Systems: Adaptive Instructional Strategies (Volume 2). Orlando, FL: US Army Research Laboratory.

FENG, M., HEFFERNAN, N. T., AND KOEDINGER, K. R. 2009. Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI), 19*(3), 243-266.

GONZALEZ-BRENES, J. P. AND MOSTOW, J. 2012. Dynamic Cognitive Tracing: Towards Unified Discovery of Student and Cognitive Models. In *Proceedings of the 5th International Conference on Educational Data Mining*. Chania, Greece.

KOEDINGER, K. R. Geometry Area 1996-97. pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=76

KOEDINGER, K. R., BAKER, R. S. J. D., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., & STAMPER, J. C. 2010. A Data Repository for the EDM community: The PSLC DataShop. In Romero C, Ventura S, Pechenizkiy M, Baker RSJd (Eds.), Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

KOEDINGER, K. R. AND MCLAUGHLIN, E. A. 2010. Seeing language learning inside the math: Cognitive analysis yields transfer. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 471–476. Austin, TX.

KOEDINGER, K. R., MCLAUGHLIN, E. A, AND STAMPER, J. C. 2012. Automated cognitive model improvement. In *Proceedings of the 5th International Conference on Educational Data Mining,* 17-24. Chania, Greece.

KOEDINGER, K. R., STAMPER, J. C., MCLAUGHLIN, E. A., AND NIXON, T. 2013. Using datadriven discovery of better cognitive models to improve student learning. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*.

LAN, A. S., STUDER, C., WATERS, A. E., BARANIUK, R. G. 2013. Tag Aware Ordinal Sparse Factor Analysis for Learning and Content Analytics. In *Proceedings of the 6th International Conference on Educational Data Mining.*

LAN, A. S., STUDER, C., WATERS, A. E., BARANIUK, R. G. 2014. Sparse Factor Analysis for Learning and Content Analytics. *Journal of Machine Learning Research, 15,* 1959-2008.

LINDSEY, R. V., KHAJAH, M., AND MOZER, M. C. 2014. Automatic discovery of cognitive skills to improve the prediction of student learning. In *Advances in Neural Information Processing Systems, 27,* pp. 1386-1394. La Jolla, CA.

LIU, R., KOEDINGER, K. R., AND MCLAUGHLIN, E. A. 2014. Interpreting Model Discovery and Testing Generalization to a New Dataset. In *Proceedings of the 7th International Conference on Educational Data Mining*. London, UK.

NATHAN, M. J, KOEDINGER, K. R., AND ALIBALI, M. W. 2001. Expert blind spot: when content knowledge eclipses pedagogical content knowledge. In Proceedings of the 3rd International Conference on Cognitive Science, pp. 644-648. Beijing, China:USTC Press.

RITTER, S., ANDERSON, J. R., KOEDINGER, K. R. AND CORBETT, A. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review, 14*(2), 249255.

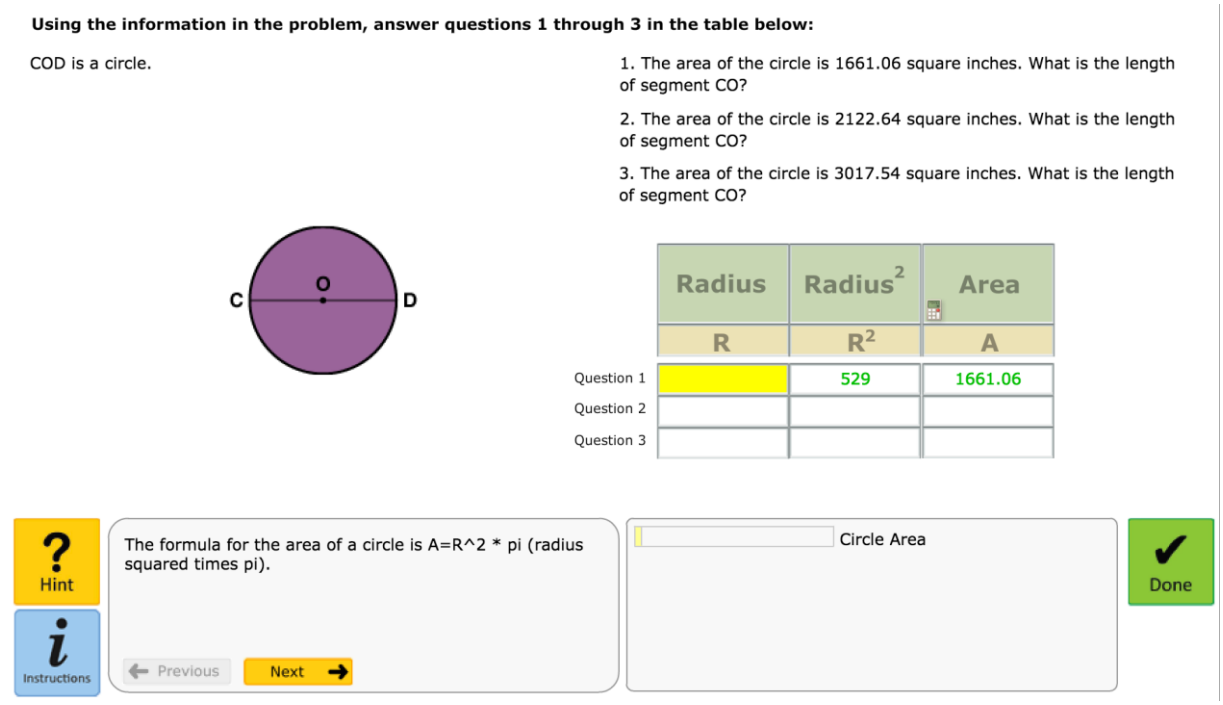
SAN PEDRO, M., BAKER, R. S., ROWERS, A. J., AND HEFFERNAN, N. T. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proceedings of the 6th International Conference on Educational Data Mining*. Memphis, TN, pp. 177–184.

SPADA, H. AND MCGAW, B. 1985. The assessment of learning effects with linear logistic test models. In: Embretson SE, ed. *Test Design: Developments in Psychology and Psychometrics*. New York: Academic Press, 169-193.

STAMPER, J. AND KOEDINGER, K. R. 2011. Human-machine student model discovery and improvement using data. (2011). In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 353–360. Auckland, New Zealand.

7. APPENDIX 1

Control condition example interface for a backward circle area problem step.



Redesigned condition example interface for a backward circle area problem step.