Interpreting Model Discovery and Testing Generalization to a New Dataset

Ran Liu
Psychology Department
Carnegie Mellon University
ranliu@cmu.edu

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

Elizabeth A. McLaughlin
Human-Computer Interaction Institute
Carnegie Mellon University
mimim@cs.cmu.edu

ABSTRACT

Automated techniques have proven useful for improving models of student learning even beyond the best human-generated models. There has been concern among the EDM community about whether small prediction improvements matter. We argue that they can be quite significant when they are interpretable and actionable, but the importance of generating meaningful, validated, and generalizable interpretations from machine-model discoveries has been under-emphasized in educational data mining. Here, we interpret a Learning Factors Analysis model discovery from a geometry dataset to suggest that students experienced difficulty applying the square root operation in circle-area backward problem steps. We then sought to validate and generalize this interpretation in the context of a completely novel dataset. Results indicated that our interpretation of the small, automated prediction improvement not only held up in the context of a novel dataset but also generalized to new types of problems that didn't exist in the original dataset. We argue that identifying cognitive interpretations of automated model discoveries and assessing the generalizability of such interpretations are critical to translating those model discoveries to concrete improvements in instructional design.

Keywords

Cognitive model discovery, model interpretation, generalization across datasets, learning factors analysis.

1. INTRODUCTION

Much Educational Data Mining (EDM) has focused on new data mining methods for improving within-dataset predictions. There has been interest in the community concerning whether small prediction improvements matter. Although we cannot provide a firm answer, we argue that they do when the improvements are interpretable and actionable. We have shown, in past experimental results, that genuine learning improvements can result from automated discoveries of small prediction differences [16]. Further, we argue that there should be more emphasis in EDM on whether predictions are clearly interpretable from a theoretical or cognitive perspective and whether the interpretation has external validity (e.g. generalizes beyond the dataset in which it was discovered).

Here, we present one of the first attempts at taking an interpretation of an automated cognitive model (or Q matrix [1, 8, 19]) discovery and generalizing that interpretation to a novel dataset, different from the one used to make the discovery. We focused on a discovery by the Learning Factors Analysis (LFA) algorithm [4] from a geometry dataset that improved predictions beyond the best available human-generated cognitive model. Even though the prediction improvement was small within this original dataset, with the addition of some exploratory data analysis, we interpreted the discovery within the context of a cognitive skill model [15].

Our intention was not to apply the improved model directly to new data (e.g., as in [11]) nor to run an exact replication of the study but, rather, to test whether the interpretation itself held up within the context of a new dataset with direct relevance to the interpretation but whose structure and properties may differ from those of the original dataset.

2. BACKGROUND

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, when combined with an appropriate theoretical interpretation, can yield better instruction and improved learning. More accurate skill diagnosis leads to better predictions of what a student knows, thus resulting in improved assessment and more efficient learning overall. Cognitive Task Analysis [5, 6, 17] is currently the best strategy for creating cognitive models of learning, but the method has its limitations. For example, it involves many subjective decisions and requires large amounts of human time and effort, as well as a high level of psychological expertise.

Educational data mining and machine learning techniques can be used to improve cognitive models in an automated fashion. These methods involve using data and statistical inference to create or modify a cognitive model involving continuous parameters over latent variables that can be linked to observed student performance variables. In addition to saving time and effort, machine models have the potential to discover cognitive model improvements that may not otherwise be considered via human-generated methods.

In order to use techniques of automated cognitive model improvement effectively towards the primary goal of bettering instructional design and assessment, it is important to properly interpret machine discoveries in the context of a cognitive skill model. Furthermore, it is critical to demonstrate the external validity of the interpretation beyond the dataset from which the discoveries were made. There exist good techniques (e.g., various methods of cross-validation) for ensuring internal validity of automated discoveries, but there have been few demonstrations of generalization beyond the samples in which discoveries are made.

Here, we discuss an example of an automated model discovery that improved a Knowledge Component (KC) Model, a specific type of cognitive skill model, beyond the best existing human-generated model. Knowledge Components represent units of knowledge, concepts, or skills that students need to solve problems. A KC Model is composed of a set of KCs mapped to a set of instructional tasks (e.g. problem steps). The LFA algorithm [4] automates the search process across hypothesized knowledge components (KCs) across a number of possible models. A tool such as the LFA algorithm not only reduces human effort and error by providing an automated method for discovering and evaluating cognitive models, but it outputs a most predictive Q matrix [19], thus producing a statistical version of a symbolic model. As such, LFA eases the burden of interpretation, but it does not in itself accomplish interpretation.

We applied the LFA search process across 11 datasets using different domains and technologies (available from DataShop at http://pslcdatashop.org; [13]). This automated process improved models, by cross-validation measures, across all of the datasets beyond the best manual models available [15]. However, the improvements in root mean square error (RMSE) were quite small. We questioned whether such miniscule changes in measurement are interpretable, generalizable, and—most importantly—actionable.

To investigate these questions, we focused on a particular dataset called Geometry Area 1996-1997, which is available to the public, has been analyzed for several other studies and shown to be reliable, and has produced findings we can test for generalization [12]. These data included 5,104 student steps completed by 59 students. Within this dataset, we compared the best LFA-discovered model (according to item-stratified cross validation) against two human generated models—the original model and the best hand-generated model (according to item-stratified cross validation). The LFA algorithm split circle-area problem steps into those that use a forward strategy (find area, given radius) and those that use a backward strategy (find radius, given area). It did not split any other area formulas for the backward-forward distinction. Thus, LFA essentially discovered an unforeseen "new" knowledge component (i.e., circle-area backward) for this dataset. As mentioned, the cross-validation results provided evidence of the internal validity of the discovered cognitive model improvement.

In the current paper, we aim to assess the external validity of this result in a novel dataset whose structure and properties are different from the original dataset in which the discovery was made. Since it was not possible to test the original LFA-discovered model directly on a new dataset due to its differing structure and problem types, it was critical that we generated a cognitive *interpretation* of the finding. The interpretation makes it possible to generate predictions and models that are appropriate to the novel dataset in which we aim to test the validity and generalization of our findings.

3. INTERPRETING MACHINE-DRIVEN MODEL IMPROVEMENTS

The LFA discovery within the Geometry Area 1996-1997 dataset yielded a result that we interpreted by combining information from the algorithm split and other relevant exploratory measures from the dataset itself. Analysis of the automated model revealed a forward-backward split only predictive for circle area (i.e., not for the other geometric shapes in the dataset nor for other circle formulas such as find diameter or radius given circumference). Data on student performance corroborated this finding. Circle-area backward problems were substantially more difficult for students than circle-area forward problems (54% vs. 80%), but performance on the other shapes exhibited small or negligible differences in forward vs. backward steps (Figure 1a). The circle-area split illustrates an important factor discovered by the LFA algorithm that had not been anticipated by human analysts.

Delving into the problem steps associated with circle-area backward computations revealed the necessity of a square root operation ($r = \sqrt{(A/\pi)}$) that was not a requirement in any of the other backward formulas. Given the unique feature of square root operation in the context of this dataset and the absence of a forward-backward model split or performance discrepancy on all other shapes' area calculations and all other circle formula calculations, we hypothesized that the automated model improvement was more about the difficulty knowing when and how to apply a square root operation than about the difficulty applying a backward strategy more generally.

Although data mining techniques helped discover the split, it took

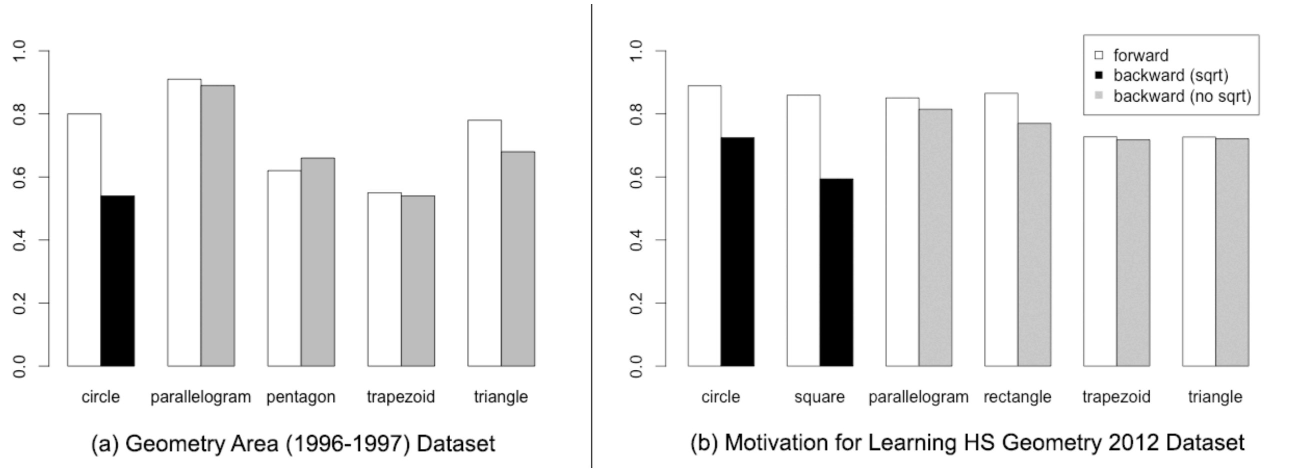

Figure 1. Average proportion correct on first attempts at geometry area problem steps, grouped by shape and color-coded based on whether the problem step requires a forward strategy, a backward strategy that requires a square root calculation, or a backward strategy that does not require a square root calculation. Panel (a) reflects the Geometry Area 1996-1997 dataset, where LFA discovered that merging forward and backward for all shapes but circle yielded the best predictions. Our interpretation was that this split reflected a difficulty applying (or knowing to apply) the square root, which only affects the circle-area backward computations. Based on this interpretation, we predicted a split between forward and backward problem steps for circles *and* squares but not other shapes. Panel (b) shows that performance in the Motivation for Learning HS Geometry 2012 dataset confirms this predicted split.

a rational cognitive analysis to identify an underlying cognitive process (e.g., square root operation) from the information obtained via the LFA output. To move from data analysis to data interpretation requires domain knowledge and cognitive psychology expertise beyond just methodological skills in EDM techniques.

4. VALIDATING AND GENERALIZING THE INTERPRETATION

Before using interpretations from machine-model discoveries to redesign instructional principles, it is often important to assess the external validity of the interpretations themselves. For example, the tutor unit for the Geometry Area 1996-1997 dataset had only three unique problem steps associated with the circle-area backward (i.e., find circle radius given area) calculation. Furthermore, it had no problem steps associated with a square-area backward (i.e., find square side length given area) calculation. Due to the limited task variety available in the Geometry Area 1996-1997 dataset, it remains unclear from that dataset alone whether our interpretation of difficulty applying the square root operation will generalize to data containing a broader set of tasks.

Thus, we sought to validate this interpretation of a machine-driven cognitive model discovery in an independent dataset containing substantially more circle-area backward problem steps as well as the existence of square-area backward problem steps, which were entirely absent from the original dataset. To this end, we investigated the geometry portion of a much more recent dataset, Motivation for Learning HS Geometry 2012 (geo-pa) [3]. This dataset is an excerpt from regular classroom use of a Geometry Cognitive Tutor [18] by 82 HS students (10th graders) with a total of 72,404 student steps. It contains similar types of shape-area modules and questions as the original dataset but has many more (49) unique circle-area backward problem steps. It also contains many (57) unique square-area backward problem steps. This makes it possible to validate (i.e. by investigating circle-area and other shape-area forward and backward performance) and generalize (i.e. by investigating square-area forward and backward performance) our interpretation of the original LFA-based discovery.

A first-pass exploratory analysis of the 2012 dataset reveals a substantially higher proportion of correct first attempts at forward, compared to backward, circle- and square-area problem steps (Figure 1b). In order to validate the specificity of the square root operation interpretation, we also investigated performance on backward vs. forward steps for all other shapes' area formulas. These data confirm that the performance differences between forward and backward area KCs are substantially smaller for the other shapes that don't require a square root operation in their backward steps (parallelogram backward=81%, forward=85%; rectangle backward=77%[1], forward=86%; trapezoid backward=72%, forward=73%; triangle backward=72%, forward=73%).

Beyond these performance data, we compared the performance of a KC model that aligns with our square root interpretation against KC models representing alternative hypotheses. Our hypothesis-driven KC model distinguishes backward-area steps from forward-area steps for circles and squares (since the backward steps require a square root operation) but does not make this forward-backward distinction for any other shapes. We compared this to a KC model that makes no forward-backward distinctions for any shapes (merges F-B across all shapes) as well as a KC model that makes all forward-backward distinctions for all shapes.

Since this dataset contained both circle-area and square-area problems, we also asked whether there might have been transfer between circle- and square-area backward problem steps on the basis that both require application of the square root operation. If there were full transfer, we would expect that a KC model merging square- and circle-area backward steps into a single skill should outperform a KC model that distinguishes square- from circle-area backward steps. To test this question of transfer, we created the former KC model and included it in our model comparison.

We compared performance across these four hypothesis-driven single-skilled[2] KC models:

1. SQRT SKILL CIR-SQ DISTINCT (58 KCs): Forward-backward steps coded as distinct for circle and square area problems; forward-backward steps merged (into a single "area" KC) for each of the other shapes. This KC model is structured based on our interpretation that backward steps requiring a square root operation should be coded as separate skills.
2. ALL SHAPES F-B MERGED (56 KCs): No forward-backward distinction for any shapes' areas (a single "area" KC is coded for each shape). This KC model is analogous to the original hand model for the Geometry Area 1996-1997 dataset from which LFA discovered the circle forward-backward area split on.
3. ALL SHAPES F-B DISTINCT (66 KCs): Forward-backward steps are coded as distinct[3] for all shapes' area problems. The comparison of our interpretation-based model (SQRT SKILL CIR-SQ DISTINCT) against this one is important for establishing the specificity of a square root operation hypothesis and rules out the possibility that the best split should, more generally, be forward vs. backward area steps across all shapes.
4. SQRT SKILL CIR-SQ BACKWARD (57 KCs): Forward steps coded as distinct for circle and square area problems; backward circle- and square-area steps merged into a single skill; forward-backward steps merged (into a single "area" KC) for each of the other shapes. The comparison of our interpretation-based model (SQRT SKILL CIR-SQ DISTINCT) against this one will inform us as to whether there was full transfer between backward circle- and square-area skills.

[1] Adjusted value reflecting the omission of 7 problem steps for which there was an error in the problem text. The pre-adjustment value is 0.70.

[2] The original KC model from which we constructed these four single-skilled models was a multi-skilled model. To convert the multi-skilled into a single-skilled model, we selected single skills corresponding with the LFA results on the Geometry Area 1996-1997 dataset.

[3] This model codes forward vs. backward steps with the finest-grain distinction possible: some shapes have multiple backward steps that are coded as distinct from each other (e.g., for parallelograms, "find height given area" and "find base given area" are coded as separate KCs).

Model Name	**KCs**	**AIC**	**BIC**	**RMSE:** ***Item-Stratified*** **Cross Validation (Average of 20 runs)**	**RMSE:** ***Student-Stratified*** **Cross Validation (Average of 20 runs)**
ALL SHAPES: F-B MERGED	56	20,992	22,652	0.28208	0.28702
ALL SHAPES: F-B DISTINCT	66	**20,839**	22,670	0.28104	*0.28588*
SQRT SKILL: CIR-SQ DISTINCT	58	20,857	**22,551**	**0.28087***	**0.28584**
SQRT SKILL: CIR-SQ BACKWARD	57	20,883	22,560	0.28113	0.28621

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 (CV). Cross validation results are reported as the average root mean-square error (RMSE) values across twenty runs of 10-fold CV. The best performing model, by each of the measures, is bolded. *Significant at the p<0.001 level in t-tests comparing model performance against all other models, except the one italics entry, over the twenty runs of cross validation.

The models were evaluated using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and 10-fold cross validation (CV). Due to the random nature of the folding process in cross validation, we repeated each type of 10-fold CV (item-stratified and student-stratified) 20 times and calculated the RMSE on each run, as has been done in previous work to handle this variabiltiy in CV [16].

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, on average, by both item-stratified and student-stratified CV measures.

To ensure that it performed better than the next best model (ALL SHAPES: F-B DISTINCT) consistently, as opposed to by chance (due to random selection of folds), we compared the RMSEs from the 20 runs of item-stratified CV and student-stratified CV between the two models using a paired t-test. For *item-stratified* CV, the SQRT SKILL: CIR-SQ DISTINCT model had consistently lower RMSEs than the ALL SHAPES: F-B DISTINCT model across every one of the 20 runs, and this pattern was significant based on a paired t-test ($t = -10.249$, $df = 19$, $p < 0.0001$). For *student-stratified* CV, the SQRT SKILL: CIR-SQ DISTINCT model had lower RMSEs than the ALL SHAPES: F-B DISTINCT model on 14 of 20 runs, which was not statistically significant by a paired t-test.

Consistent with our previous work comparing machine-discovered models to baseline models [15], we focus on item-stratified 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.

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 circles and squares is better than leaving F-B merged. Notably, SQRT SKILL CIR-SQ DISTINCT even performs better, by item-stratified CV 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 even more fine-grained distinctions in the form of F-B separation for other shapes. This validates the specificity of the square root operation hypothesis and rules out the possibility that the major split should be for general forward vs. backward strategies among all shapes' area problems.

Thus, there is good evidence from KC model comparisons that distinguishing forward from backward steps specifically for circle- and square-area problems but not other shape-area problems predicts student learning best. This validates and generalizes our original interpretation that knowing when and how to apply the square root operation is the basis for the cognitive model improvements.

We did not observe full skill transfer between backward circle- and square-area steps, since the SQRT SKILL CIR-SQ BACKWARD model performed consistently worse than the SQRT SKILL CIR-SQ DISTINCT model by all measures. To investigate whether this may have been due to a lack of variability in the order that students complete circle-area backward vs. square-area problem steps, we examined the relative ordering of the two shapes' backward area steps. We discovered that each individual student completed all square-area backward steps before any circle-area backward steps. These data show a lack of variability in the relative ordering of the opportunities for the two shapes' backward-area practice, which suggest the combined model may only reflect partial transfer. This interpretation is supported by the observation that the end of the square-area backward learning curve (Figure 2, middle panel) does not align well with the beginning of the circle-area backward learning curve; rather, there is an increase in error rate (computed by taking the inverse logit of model values) from the end of square-area backward (22.2%) to the start of circle-area backward (47.3%).

We investigated learning curve prediction improvements yielded by our hypothesis-driven models (SQRT SKILL CIR-SQ

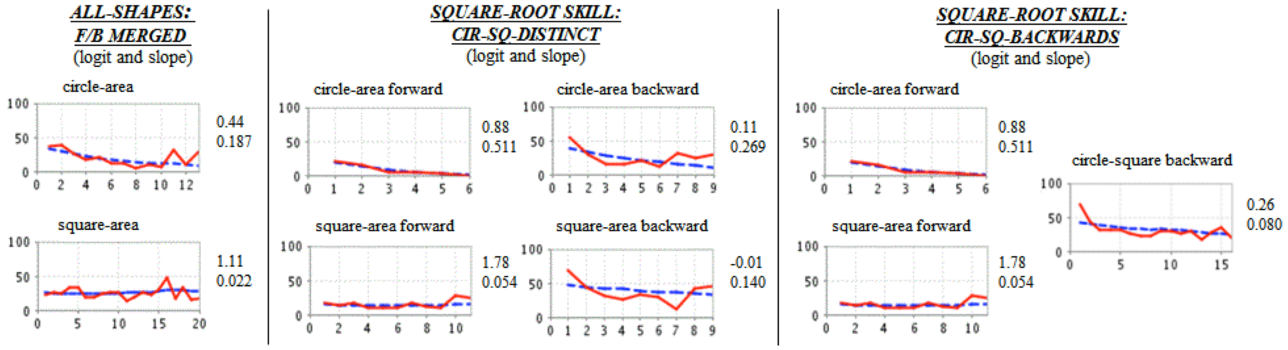

Figure 2. Learning curve prediction improvements (from the new 2012 dataset) yielded by comparing the square root KC models (middle and right panels) based on our interpretation of the LFA discovery against one that reflects what the KC model would have been (ALL-SHAPES: F-B MERGED, left panel) without the LFA discovery/interpretation. The x-axis reflects the opportunity number. Each data point was required to have at least 10 observations. The interpretation-based KC models that yielded better prediction results also exhibited higher learning slope values (bottom number to the right of each graph). This finding is consistent with what we observed using the LFA-discovered model in the original dataset.

DISTINCT and SQRT SKILL CIR-SQ BACKWARDS) compared to the baseline KC model (ALL SHAPES F-B MERGED). Figure 2 shows these learning curve predictions as well as their AFM model values (logit and slope). Our hypothesis-driven KC models, both of which consistently performed better than the baseline KC model, exhibit higher learning slope values. This finding is consistent with our general LFA results [15] that showed that models with better prediction results had higher learning slope values.

5. RELATED WORK

Beck & Xiong [2] rightfully raised concerns about the fact that many promising modeling approaches have produced only “negligible gains in accuracy, with differences in the thousandths place on RMSE.” That paper focused on differences in statistical modeling approaches, such as Bayesian Knowledge Tracing and Performance Factors Assessment, whereas our focus is on cognitive model improvements. Beck & Xiong do make a similar comment about how cognitive model (they use the phrase “transfer model”) modifications produce only “slight improvement in accuracy”. In our case, we argue that even slight improvements can yield meaningful and valid interpretations that generalize to new contexts within the same domain and can be used to produce significant differences in student learning.

We completely agree with Beck & Xiong’s suggestion that “higher predictive accuracy is not sufficient” and with their emphasis on interpretability, “is there any interpretable component relating to student knowledge?” We share a desire to connect results to student learning and address questions like “can we use this model to predict whether an intervention will lead to more learning?” However, we emphasize using interpretation of models not only to predict the impact of an intervention, but also as a *guide to design* such interventions. As we discuss below, cognitive model improvements, even ones with small impact on prediction accuracy, can be used to guide new instructional designs and high plausibility for impact in improving student learning. We need to “close the loop” and test whether designs based on cognitive model insights do improve learning, as has been done in past experiments [14, 16].

Learning Factors Analysis (LFA) requires human intervention to propose factors that may (or may not) account for task difficulty or transfer of learning from one task to another (e.g., backward application of a formula). This human intervention can be considered a downside of LFA relative to other cognitive model or q-matrix discovery algorithms [e.g., 1, 7-9, 19] that automatically produce new factors (e.g., as clusters of tasks with similar factor loadings). The results of these models, however, must be interpreted and post-hoc factor labeling is, in our experience, extremely difficult. It is quite hard to make sense of discovered factors or the task clusters they imply. We suspect that such interpretation difficulty is the reason that, to our knowledge, none of these methods have been used to produce new cognitive model explanations of task difficulty or transfer. More importantly, to our knowledge, none of them have been used to redesign instruction that can be tested in close-the-loop experiments. Thus, while LFA does require upfront human intervention to propose factors, this upfront investment appears to pay off in that LFA output affords more effective interpretation of model results on the backend.

At the other extreme, traditional methods of Cognitive Task Analysis such as structured interviews of experts [5, 6, 17] or think alouds [10] puts great emphasis on logical interpretation. They draw on qualitative data and are quite time-consuming or expensive to implement. LFA offers a quantitative alternative that may be easier to implement.

Other work besides ours has tested models produced using one dataset on another. For example, it was demonstrated that the structure and parameterization of a model using ASSISTment (www.assistments.org) system interaction data to predict state test scores in one year also works well in predicting state test scores from data in another year [11]. Here, we focused on transferring not only the specific structure of the model (e.g., the Q-matrix) but the cognitive insights from interpreting the model. The latter allowed us to make predictions on a kind of task (i.e., square-area backward) that was not even present in the original data or in the original Q-matrix. Making predictions of student performance on unseen tasks is something that a purely statistical model cannot do. We need to extend such models with logical or structural interpretations that have both explanatory power (i.e., they help us make sense of student learning) and generative power (i.e., they guide the design of better instruction).

6. CONCLUSIONS & FUTURE DIRECTIONS

Although the reduction in overall error (RMSE) was rather small in the original LFA model discovery on dataset Geometry Area 1996-1997, we demonstrated that the theoretical interpretation of this discovery was not only validated in an independent dataset but also generalized to new problem types that were not part of the original dataset (i.e., square-area backward). Error reductions can be small as a consequence of most of the model being essentially the same as the original but can still indicate a few isolated changes that are highly practically significant for tutor redesign. In a recent close-the-loop study [16], we demonstrated how using a cognitive model discovery to redesign a tutor unit led to both much more efficient and more effective learning than the original tutor. In that case, the discovered model had a statistically significantly lower RMSE on item-stratified cross validation (0.403) than the existing human-created model (0.406). The actionable interpretation of this small difference, only 0.003 in RMSE, was demonstrated to be practically important.

Some other automated techniques discover models that are difficult or impossible to understand (e.g., matrix factorization [7, 9]), either toward deriving insights into student learning or making practical improvements in instruction. The output of LFA is more interpretable and convertible to tutor changes than these alternative methods that may produce latent variable representations without the consistent application of human-derived codes or without code labels at all.

Here, we aimed specifically to assess the generalizability of our cognitive interpretation of an LFA model discovery. We showed that our interpretation held up within the context of a new dataset with domain relevance but whose structure and properties differed from those of the original dataset. Validation and generalization were confirmed, in the 2012 geometry dataset, based on (1) performance measures and (2) superior prediction of learning by a KC model constructed based specifically on our interpretation.

These findings move beyond simply replicating the original LFA model discovery. Since the novel dataset had a different structure from the original dataset, including differences relevant to our interpretation (i.e., existence of square-area backward problem steps), it would not have been viable to directly test the discovered automated model on this new dataset. Thus, the interpretation of automated model discoveries is actually *necessary* in order to test the generalizability of such discoveries across contexts with non-identical structures. Furthermore, interpretations help anchor all subsequent data exploration and analyses to something meaningful that can then be translated into concrete improvements to instructional design.

Testing the generalization of our interpretation not only confirmed the robustness of the idea but also yielded further details about the scope of the interpretation that have relevant implications for modifying instruction. For example, the original automated discovery may have suggested that we should treat circle-area backward problems as a separate skill, but the generalization of our interpretation suggests we should treat all backward area problems involving application of the square root operation—including square area—as distinct from their forward area counterparts. Further, the demonstrated validity of our interpretation has potential implications for instructional design beyond the cognitive tutors used to generate the datasets we worked with here. For example, the Khan Academy (www.khanacademy.org) geometry area units treat all circle-area problems as one skill and all square-area problems as one skill, with no forward-backward distinction, in their practice sets. Our findings suggest, at the very least, that it may be worth investigating whether our discovered interpretation also generalizes to student performance in very different instructional contexts such as that in the Khan Academy. If so, it would suggest potential instructional improvements there as well.

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 [16]. These include resequencing (positioning problems requiring fewer KCs before ones needing more), knowledge tracing (adding or deleting skill bars), creating new tasks, and adding/changing feedback or hint messages.

From the cognitive model improvement demonstrated here, we suggest adding new skills to the tutor that differentiate backward circle- and square-area problem steps from their forward counterparts. For other shapes, in contrast, we suggest that the skills for forward and backward area problem steps be merged. These skill bar changes would lead to changes in knowledge tracing as well as the creation of new tasks. In particular, students would receive increased practice on circle-area and square-area backward problems and decreased practice on some forward and backward steps for other shapes' area formulas. Finally, we suggest that new tasks or hint messages might be added to the backward circle- and square-area practice problems. For example, we might include additional questions, or hints, that simply ask "What do you need to do to 50 in x^2 = 50 to find the value of x?" We expect that the combination of increased practice on newly discovered skill difficulties and new tasks/hints that scaffold the difficulty would significantly improve overall student learning. In future work, we aim to "close the loop" on this finding by implementing these suggested instructional design changes and testing whether a redesigned tutor yields improvements in student learning above those achieved by the current tutor.

More generally, this work contributes to a broader set of evidence that a deep understanding of the cognitive processes of a domain through Cognitive Task Analysis (CTA) can lead to instructional designs that produce much better learning than typical instruction created through the self-reflections of a domain expert [5, 6, 17]. Prior work on CTA involves time-consuming expert interviews and subjective qualitative analysis. We find great promise in using data mining as a form of quantitative CTA that can more automatically and efficiently produce actionable discoveries. Nevertheless, the analysis process still involves human expertise in cognitive science to interpret model output and hypothesize cognitive interpretations that can be used to generalize across datasets and make effective instructional design decisions.

7. ACKNOWLEDGMENTS

Thanks to Carnegie Learning, Inc. for providing the Cognitive Tutor data supporting this analysis, to DataShop (http://pslcdatashop.org) for dataset access and to the National Science Foundation (#SBE-0836012) for funding of the Pittsburgh Science of Learning Center.

8. REFERENCES

[1] Barnes, T. (2005). The q-matrix method: Mining student response data for knowledge. In Proceedings of the American Association for Artificial Intelligence 2005 Educational Data Mining Workshop, pp. 1-8. Pittsburgh, PA.

[2] Beck, J. E., & Xiong, X. (2013). Limits to accuracy: How well can we do at student modeling. Proceedings of the 6th International Conference on Educational Data Mining.

[3] Bernacki, M., & Ritter, S. (2012). Motivation for Learning HS Geometry 2012 (geo-pa). Dataset 748 in DataShop. https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=748

[4] Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K.D. Ashley, T.-W. Chan (Eds.) Proceedings of the 8th International Conference on Intelligent Tutoring Systems, pp. 164-175. Berlin: Springer-Verlag.

[5] Clark, R. E. (2014). Cognitive Task Analysis for Expert-Based Instruction in Healthcare. In J. Michael Spector, J. M. Merrill, M. D. Elen, J. and Bishop, M. J. (eds.). Handbook of Research on Educational Communications and Technology, 4th Edition, pp. 541-551. Springer: New York.

[6] 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.

[7] Desmarais, M. C. (2011). Mapping question items to skills with non-negative matrix factorization. ACM KDD-Explorations, 13(2), pp. 30-36.

[8] Desmarais, M. C., Behzad B., and Naceur, R. (2012). Item to skills mapping: deriving a conjunctive q-matrix from data. In Intelligent Tutoring Systems, pp. 454-463. Springer: Berlin-Heidelberg.

[9] Desmarais, M. C. and Naceur, R. (2013). A Matrix Factorization Method for Mapping Items to Skills and for Enhancing Expert-Based Q-matrices. In Proceedings of the 16th Conference on Artificial Intelligence in Education (AIED2013), pp. 441-450. Memphis, TN.

[10] Ericsson, K. A., & Simon, H. A. (1984). Verbal reports as data. Cambridge, MA: MIT Press.

[11] Feng, M., Heffernan, N. T., & 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), pp. 243-266.

[12] Koedinger, K. R. (2006). Geometry Area 1996-1997. Dataset 76 in DataShop. https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=76

[13] Koedinger, K. R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

[14] Koedinger, K. R. & McLaughlin, E. A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In: Ohlsson, S., Catrambone, R. (eds.) Proceedings of the 32nd Annual Conference of the Cognitive Science Society, pp. 471–476. Austin, TX.

[15] Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated cognitive model improvement. Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. (eds.) Proceedings of the 5th International Conference on Educational Data Mining, pp. 17-24. Chania, Greece.

[16] Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better cognitive models to improve student learning. Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED2013).

[17] Lee, R. L. (2003). Cognitive task analysis: A meta-analysis of comparative studies. Unpublished doctoral dissertation, University of Southern California, Los Angeles.

[18] Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). The Cognitive Tutor: Applied research in mathematics education. Psychonomics Bulletin & Review, 14(2), pp. 249-255.

[19] Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.