Closing the Loop with Quantitative Cognitive Task Analysis

Kenneth R. Koedinger
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
koedinger@cmu.edu

Elizabeth A. McLaughlin
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15201
mimim@cs.cmu.edu

ABSTRACT

Many educational data mining studies have explored methods for discovering cognitive models and have emphasized improving prediction accuracy. Too few studies have "closed the loop" by applying discovered models toward improving instruction and testing whether proposed improvements achieve higher student outcomes. We claim that such application is more effective when interpretable, explanatory models are produced. One class of such models involves a matrix mapping hypothesized (and typically human labeled) latent knowledge components (KCs) to the instructional practice tasks that require them. An under-investigated assumption in these models is that both task difficulty and learning transfer are modeled and predicted by the same latent KCs. We provide evidence for this assumption. More specifically, we investigate the data-driven hypothesis that competence with Algebra story problems may be better enhanced not through story problem practice but through, apparently task irrelevant, practice with symbolic expressions. We present new data and analytics that extend a prior close-the-loop study to 711 middle school math students. The results provide evidence that *quantitative cognitive task analysis* can use data from task difficulty differences to aid discovery of cognitive models that include non-obvious or hidden skills. In turn, student learning and transfer can be improved by closing the loop through instructional design of novel tasks to practice those hidden skills.

Keywords

Cognitive task analysis, cognitive model, transfer, knowledge components, close-the-loop experiment

1. INTRODUCTION

As the field of Educational Data Mining (EDM) strives for technical innovation, there is risk of losing the "E" in "EDM", that is, of not making a clear link to the "Educational" in "Educational Data Mining". Connected with this concern is the temptation to evaluate EDM research only in terms of predictive accuracy and not place value on interpreting the resulting models for plausibility and generalizable insights. While it is possible to use uninterpretable or "black box" predictive models in educational applications (e.g., [1]), interpreting model results is an important step toward improving educational theory and practice for three reasons: 1) for advancing scientific understanding of learning or educational domain content, 2) for generalization of models to new data sets (cf., [19]), and 3) for gaining insights that lead to improved educational technology design.

Whether an educational application of EDM is through a black box model or mediated by data interpretation, the most important, rigorous, and firmly grounded evaluation of an EDM result is whether an educational system based on it produces better student learning. Such an evaluation has been referred to as "closing the loop" (e.g., [16]) as it completes a "4d cycle" of system **d**esign, **d**eployment, **d**ata analysis, and **d**iscovery leading back to design. The loop is closed through an experimental comparison of a system redesign with the original system design.

Use of the "close the loop" phrase, in our writing, goes back at least to [12]. Early examples of data-driven tutor designs, that is, of a close-the-loop experiment, can be found in [13] which tested a tutor redesign based on discoveries from data originally published in [17] and in [4], which was based on data analysis [5]. It is notable that a systematic process for going from data to system redesign was not articulated in this early work, but has been increasingly elaborated in more recent writings [especially 16].

This paper further specifies a particular class of analytic methods, namely *quantitative cognitive task analysis* methods, and how to use them to close the loop. The output of a cognitive task analysis (CTA) is a model of the underlying cognitive processing components (so-called knowledge components or KCs) that need to be learned to perform well in a task domain. Quantitative CTA uses data on task difficulty and task-to-task learning transfer to make inferences about underlying components.

1.1 Cognitive Task Analysis

In general, Cognitive Task Analysis (CTA) uses various empirical methods (e.g., interviews, think alouds, observations) to uncover and make explicit cognitive processes experts use and novices need to acquire to complete complex tasks [3]. Various representations of the resulting cognitive model (e.g., goal trees, task hierarchies, if-then procedure descriptions) are used to design or redesign instruction. Close-the-loop experiments in different domains demonstrate that students learn more from instruction based on CTA than from previously existing instruction (e.g., medicine [23]; biology [8]; aviation [20]). These results come from CTAs using qualitative research methods that are costly and substantially subjective.

Quantitative CTA methods provide greater reliability and are less costly (though ideally used as a complement to qualitative CTA). An early close-the-loop study [13] based from a Difficulty Factors Assessment (DFA) showed that algebra students are better at solving beginning algebra story problems than matched equations. In a controlled random assignment experiment, the newly designed instructional strategy was shown to enhance student learning beyond the original tutor. Besides DFA, automated techniques can further reduce human effort and can be used on large data sets. An early example used learning curve analysis to identify hidden planning skills in geometry area [16] that resulted in tutor redesign. In a close-the-loop experiment comparing the original tutor to the redesigned tutor, students reached mastery in 25% less time and performed better on complex planning problems on the post-test. Further research [15] has shown how a search algorithm (e.g., Learning Factors Analysis) can generate better alternative cognitive models.

A key assumption behind DFA is that significant differences in task difficulty can be used to make non-obvious (sometimes counter-intuitive) inferences about underlying cognitive components and, in turn, these components help predict learning transfer and guide better instructional design. Similarly, statistical models of learning, including both logistic regression and Bayesian Knowledge Tracing variations, also tend to assume that both task difficulty and learning transfer can be predicted using the same KC matrix.

Recent work explored this connection [18] and found, across 8 datasets, that statistical models that use the *same* KC matrix to predict task difficulty *and* learning transfer produce better results than models that use *separate* matrices (item vs. KC). A key goal of this paper is to further investigate this difficulty-transfer linkage claim by extending evaluation of it through close-the-loop experimentation.

1.2 Illustrating Quantitative CTA

Consider the problems in Table 1 and try to answer the following question before reading on. Assuming the goal of instruction is to improve students' skill at translating story problems into algebraic symbols (e.g., translating the 2_step story in the first column of Table 1 into "62+62-f"), which will yield better transfer of learning: practice on 1_step story problems (columns 2 and 3) or practice on substitution problems (column 4)? Note that in the close-the-loop experiment we ran, similar multiple matched problem sets were created. A different problem set was used for practice than was used for transfer. For example, students who saw the 2-step problem in Table 1 as a transfer post-test item would not see the associated 1_step or substitution problems from Table 1 as practice problems. So, again, which yields better transfer to 2_step problems, practice on 1_step or substitution?

If you answered that practice on the 1_step story problems will better transfer to 2_step story problems, you are in good company as learning commonalities underlying problem formats (i.e., deep features) is a known factor in aiding analogy and transfer [9; 10]. But, the following quantitative analogy cognitive task analysis suggests a different answer.

Table 1. Examples of problem variations and their solutions.

2_step	1_step	1_step	substitution
Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches ***f fewer boys*** than girls. Write an expression for how many students Ms. Lindquist teaches.	Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches ***b boys***. Write an expression for how many students Ms. Lindquist teaches.	Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches ***f fewer boys*** than girls. Write an expression for how many boys Ms. Lindquist teaches.	Substitute 62-f for b in 62+b Write the resulting expression.
62+62-f	62+b	62-f	62+62-f

Using DFA, [11] explored the struggle beginning algebra students have with translating story problems into symbolic algebra expressions. A common belief is that story problems are hard due to comprehending the story content. However, two results indicate that comprehension is not a major roadblock. First, students are better able to solve 2_step problems when given a value (e.g., answering 116 when f is given as 8 in the 2_step story shown in Table 1) than when asked to write the symbolic expression (e.g., 62+62-f or even 62+62-8) [11]. Second, students do not do better when given explicit comprehension hints of the needed arithmetic operations than they do on 2_step symbolization problems without hints [11]. If comprehension is not the key challenge, perhaps production of the target algebraic symbols is. Their results show students perform consistently better (62% vs. 40%) symbolizing both 1_step problems (e.g., producing 62+b and 62-f for the 1_step problems in Table 1) than on 2_step problems (e.g., producing 62+62-f for the 2_step story problem in Table 1).

These results suggest inferences about unobserved or "hidden skills" that are needed to translate 2_step stories into symbolic expressions such as learning how to put one algebraic expression inside another (e.g., as the one- operator expression 40m is inside the two-operator expression 800- 40m). The results are consistent with a need for skills that extend the implicit grammar for generating expressions for 1_step symbolization to recursive structures (e.g., "expression => expression operator quantity" and "expression => quantity operator expression"). Furthermore, they suggested that practicing non-contextual substitution problems (see last column of Table 1) should help students (implicitly) learn the desired recursive grammar structures and the corresponding production skills for constructing more complex expressions.

1.3 Analysis Methods

Our first analysis explores how much substitution practice transfers to story symbolization. We pursue this question with respect to broad outcomes and learning processes. This analysis replicates the high level analysis of the prior study (2008-09) [14] with a full dataset accumulated across four school years (2008-12). Our second analysis probes, more specifically, the question of the cognitive model link between task difficulty and learning transfer that underlies quantitative cognitive task analysis and, more generally, adaptive tutoring models like Bayesian Knowledge Tracing. Practically, the theoretical claim that learning transfer can be inferred from task difficulty data suggests that we can design instruction that produces better transfer of learning using models built from difficulty data (which is easier to collect).

Our third analysis examines whether statistical models of the learning process data support conclusions drawn from the outcome data. Does learning curve analysis indicate whether and how tasks (e.g., substitution problems) designed to isolate practice of CTA-discovered hidden skills (e.g., recursive grammar) transfer to complex tasks that theoretically require these skills (e.g., 2_step story problems)?

2. METHOD

The original 2008-09 study [14] and current close-the-loop study were run with middle school students as part of a math course. In the original study, students were randomly assigned to either a substitution practice condition (N=149) or 1_step story practice condition (N=154). Since then, additional data with random student assignment was collected over three school years from 2009-12 (N=234 for substitution practice, N=174 for 1_step story practice) using the same problem set in ASSISTments. As previously described [14], the study involved a pre-test, instruction, and post-test. For the substitution condition, substitution problems were embedded as instruction interleaved with 2_step story problems (posttest). For the 1_step condition, 1_step problems were used as instruction interleaved with the same 2_step story problems. The pretest for a given version and order was the same for both conditions. Order was determined by difficulty of 2_step problems from a pilot study and included a sequence of 2_step problems from easy to hard or hard to easy.

Small changes were made to the automated scoring to give better feedback on unusual but arguably correct answers (e.g., d60 instead of 60d). For consistency in scoring, manual corrections made to the 0809 dataset [14] were combined with the corrections to the 0912 dataset and automatically applied to every answer in the combined dataset (0812).

3. RESULTS AND DISCUSSION

3.1 High Level Transfer

In our first study [14], we reported significant main effects for condition and order while controlling for pretest, and no significant two-way or three way interaction effects when version was added as an independent variable. In the new study, we add a fifth factor for when the data was collected (i.e., from 0809 or from later years 0912). Most importantly, in a full five-factorial ANCOVA (in R with pretest as the covariate), we found a main effect for condition (F(1,679) = 4.5, $p < 0.05$, $d = .21$). Main effects were also found for pre-test (F(1,679) = 235.3, $p < 0.001$), order (F(1,679) = 117.8, $p < 0.001$), and version (F(1,679) = 19.8, $p < 0.001$), but study year was insignificant. Significant two-way interactions were found for pre-test and condition (F(1,679) = 4.05, $p < 0.05$), pre-test and order (F(1,679) = 18.69, $p < 0.001$), and order and year (F(1,679) = 10.77, $p < 0.01$). No other higher- level interactions were significant (all $p > 0.05$).

The pre-test by condition interaction is a consequence of the substitution treatment having a greater effect for students with higher pre-tests. Based on a median split of pre-test scores, students with a higher pre-test, showed greater benefits of substitution practice (52% posttest) over 1_step practice (44%). In contrast, students with a lower pre-test show less benefit of substitution practice (24% posttest) over 1_step practice (20%). This interaction is theoretically consistent with the cognitive task analysis in that students who cannot generate symbolizations for 1_step problems (e.g., 800-y and 40x) will not have the raw material they need to compose 2_step expressions (e.g., 800-40x). Figure 1 illustrates the interaction. Substitution practice produces transfer to story problem symbolization for the 82% of students (580 of 711) with pre-tests of at least 40%. For the 18% of students without 1_step story skills (below 40% on the pre-test), substitution practice does not provide a benefit.

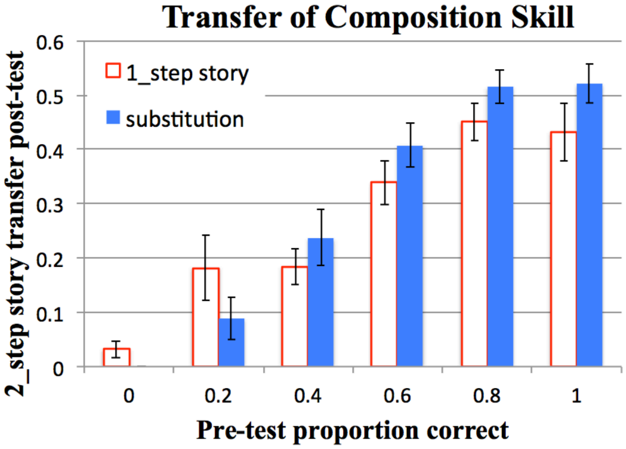

Figure 1. The benefit of substitution practice for symbolizing 2_step story problems is present for the 82% of students with some incoming competence in 1_step story symbolization (at least 40% correct).

The two other reliable interactions in the ANCOVA are not of theoretical significance, but we report them for completeness. The pre-test by order interaction is manifest in that the difference between high and low pre-test students is bigger on the easier post-test problems (63% - 31% = 32%), which appear in the hard-to-easy order, than on the harder post-problems (38% - 10% = 28%), which appear in the easy-to-hard order. The order by year interaction is a consequence of students in the 0912 school years showing more sensitivity to the order manipulation than students in the 0809 school year, such that they do relatively better on the easy problems (46% vs. 41%), but worse on the hard problems (24% vs. 30%).

3.2 Difficulty Reliably Predicts Transfer

In this analysis, we more precisely test the following general logic: If difficulty data indicates a hidden skill that makes an important task hard, then inventing new practice tasks to isolate practice of that hidden skill will transfer to better learning of that hard task. The specific version of the logic in this domain is: If the hard part of symbolizing a two operator story problem is in composing symbolic expressions, then practice on substitution problems should transfer to better performance on story problem symbolization. Our data set affords an interesting opportunity to more precisely test this logic because the difficulty data we have indicates hidden skills for some problem types, but not others. A precise application of the "hidden-skill-transfer" logic stated above is that we should see the predicted transfer for those problem types in which the hidden skill is indicated by the difficulty data. For the other problem types, there should be no reliable transfer.

We used the current data to reevaluate the "composition effect" [11]. This analysis is shown in Table 2 where task difficulty and transfer results are shown for each of the eight problems. Consider the row for the *class* problem (referred to as "*students*" in the data file), which is illustrated in Table 1. The answer for the 2_step story and substitution problems, namely 62+62-f, is shown in the second column. The third and fourth columns show the proportion correct on the 1_step story problems, (.75 for the "a" step with the answer 62+b and .70 for the "b" step with the answer 62-f). The fifth column (labeled a*b) shows the probability of getting both of these steps correct, computed here as the product of the proportion correct on each step, .53 = 75*.70. This value is the baseline for the composition effect.

The sixth column is the proportion correct on the 2_step story problem, 0.13. This value was computed from student performance on the pre-test for both conditions and the post-test for the 1_step practice condition. We did not use the post-test for the substitution practice condition to estimate the composition effect as the theory predicts that substitution practice should reduce that effect.

A composition effect is indicated when students are less likely to correctly symbolize a two operator story than to correctly symbolize both of the matched one operator stories. The seventh column displays this difference (.40 = .53 - .13 for the *class* problem). The eighth column shows the estimated conditional probability that students can compose a single two-operator expression (e.g., 62+62-f) given they have correctly formulated the two source one-operator expressions (e.g., 62+b and 62-f). Since p(2_step) = p(a*b) * p(2_step | a*b), we get p(2_step | a*b) = p(2_step)/p(a*b), thus for the *class* problem p(2_step | a*b) = .13/.53 = .25. The lower this value, the bigger the composition effect.

The important feature to note about values in the composition effect columns is that they indicate there is no composition effect for the *cds* and *mcdonalds* problems (see the last two rows). Both are relatively well-practiced forms, the 5h-7 for *mcdonalds* is a high frequency linear form (i.e., mx+b) and the *cds* form 5*12*c involves a repetition of the same operator which can be treated as a 1-operator solution, namely, 60c (as 17% of students did). Students may have specialized knowledge for producing these forms that do not require general recursive grammar knowledge.

Table 2. Composition effects are found for all but the bottom two problems

Problem name	2_step solution	1_step (a)	1_step (b)	a*b	2_step	Composition Effect		Subst transfer
						a*b - 2_step	2_step/(a*b)	
trip	550/(h-2)	0.65	0.78	0.51	0.11	0.40	0.22	0.08
class	62+62-f	0.75	0.70	0.53	0.13	0.40	0.25	0.12
jackets	d-1/3*d	0.58	0.54	0.29	0.16	0.13	0.56	-0.02
sisters	(72-m)/4	0.71	0.63	0.45	0.32	0.13	0.72	0.15
rowboat	800-40m	0.75	0.55	0.38	0.28	0.10	0.73	0.07
children	(972+b)/5	0.66	0.75	0.5	0.38	0.12	0.76	0.09
cds	5*12*c	0.71	0.74	0.52	0.52	0.00	1.00	0.14
mcdonalds	5*h-7	0.66	0.85	0.56	0.72	-0.16	1.29	-0.06

The final column (Subst transfer) shows how much substitution practice transferred to 2_step symbolization as computed by the difference in post-test scores on each problem for the two experimental groups.

To test the hidden-skill-transfer hypothesis, we expect the *cds* and *mcdonalds* problems to show less transfer and the other problems to show more. While this is not strictly the case (*cds* shows transfer and *jackets* does not), there is a trend here that is illustrated in Figure 2. It shows the relationship between difficulty variation in the composition process and variation in the amount of transfer produced by substitution practice in the close-the-loop experiment. To better highlight the point, the graph shows the data from the 353 students at or above the median on the pre-test -- the ones for which improvement in composition skills should produce better post-test performance on 2_step story problems requiring such skills.

Consistent with the hidden-skill-transfer hypothesis, there is no transfer benefit (first two bars in Figure 2) for the two problem forms with no composition effect (*mcdonalds* and *cds*). There is large transfer effect for the three problems (*trip*, *sisters*, and *children*) involving parentheses (last two bars), which present greater challenges for composing expressions and the need for students to acquire more complex implicit grammar structures for generating correct parenthetic expressions. There is an immediate transfer effect for the three problems (*class, jackets,* and *rowboat)* not involving parentheses (middle bars), consistent with the fewer composition skills required. Note that success on these problems is oddly lower overall. We return to this point in the learning curve analysis where we do some search for new difficulty factors and hypothesize a new hidden skill that could be pursued in future close-the-loop instructional design. These results add to prior evidence [18] supporting the hypothesis that differences in task difficulty and transfer effects are observable manifestations of the same underlying KCs.

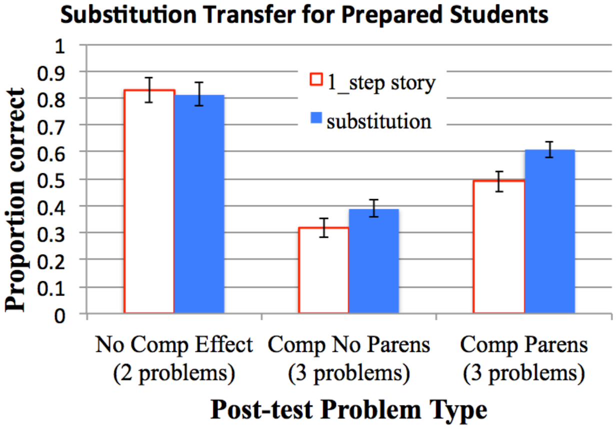

Figure 2. Transfer is limited to the problems that show a composition effect in task difficulty comparisons.

3.3 Learning Curve Analysis

As a visual representation of student performance data over time (i.e., as opportunity increases, error rates are expected to decrease), learning curves can be used to explore areas of student difficulty and transfer of learning [21]. Following this prior work, we used the statistical model for learning curve prediction built into DataShop (see PSLCDataShop.org): The Additive Factors Model is a logistic regression model that generalizes Item Response Theory by having latent variables for knowledge component difficulty in place of item difficulty and by adding a third growth term, a knowledge component learning rate, in addition to the student proficiency and knowledge component difficulty terms. We evaluate four different knowledge component models in terms of their prediction fit to all of the test and instructional items each student experienced. For our metrics, we use root mean squared error (RMSE) averaged over 20 runs of 3-fold item-stratified and student-stratified cross validation. Given the focus on understanding the difficulty and transfer characteristics of the task environment, we put particular value on predictive generalization across items (as item stratification achieves by randomly putting all data on each item in the same fold) but also report the predictive generalization across students (as student stratification achieves by randomly putting all data on each student in the same fold).

The results of a learning curve analysis are shown in Table 3. The first row displays a simple baseline no-transfer model that treats each problem type (2_step, 1_step, and substitution) as requiring a different knowledge component (KC). The second row displays a substitution transfer model that introduces transfer between substitution problems and 2_step problems by having a recursive grammar KC common to both problems. The 2_step problems have an additional KC for comprehending the story and the 1_step problems have a different unique KC. As shown in the last columns, this substitution transfer model produces a reduction in RMSE on the item stratified cross validation, down to 0.426 from 0.429. This small change is associated with a small change in the models and changes at this level (in the thousandths) have proven meaningful in producing a prior close-the-loop improvement [16]. This close-the-loop study provides further evidence that small prediction differences can be associated with significant learning gains.

Corresponding with the discussion above regarding the unique challenges of solutions requiring parentheses, the paren-enhanced model (third row in Table 3) adds a parenthesis KC to the 2_step and substitution versions of the *trip, sisters,* and *children* problems. Surprisingly, this model does not improve the item generalization (0.428 > 0.426), though it does improve student generalization (0.473 < 0.477). The predictions of this model fail to account for the variance in difficulty of the non-parentheses problems.

As mentioned above, we were surprised that a couple of the non-parentheses problems posed great difficulty. In particular, the *class* (62+62-f) and *jackets* (d-1/3d) problems were quite hard (13% and 16% correct before substitution instruction). We hypothesized the difficulty of these problems was due to a quantity being referenced twice in the solution expression (i.e., 62 in the *class* problem and d in the *jackets* problem). To test this hypothesis we built the double-ref-enhanced model (fourth row in Table 3) by adding a double-ref KC to the paren-enhanced model on both of the 2_step and substitution versions of the *class* and *jackets* problems. The result is a substantially better prediction than the prior model on both item generalization (0.416 < 0.428) and student generalization (0.468 < 0.473).

Table 3. Knowledge component learning curve model comparison.

	KCs	Recursive grammar skill for 2_step & substitution	Paren skill	Double-ref skill	Item stratified CV (RMSE)	Student stratified CV (RMSE)
No-transfer	3	0	0	0	0.429	0.478
Substitution transfer	3	1	0	0	0.426	0.477
Paren-enhanced	4	1	1	0	0.428	0.473
Double-ref-enhanced	5	1	1	1	0.416	0.468

We have not yet modeled, but have recognized an alternative or additional explanation for the difficulty of the *class* and *jackets* problems. Right expanding forms, which require the "expression => quantity operator expression" rule, may be harder than left expanding forms, which require the "expression => expression operator quantity" rule. This idea garners plausibility from cognitive theory given that right expanding forms may require more cognitive load to maintain the subexpression to be written (e.g., 62-f) while the first part is planned and written (e.g, "62 +"). This analysis predicts that the *trip, class, jackets*, and *rowboat* problems should be more difficult and they are the most difficult 2_step problems.

Future analytic and modeling efforts should pursue these plausible new hidden skills hypotheses and, if confirmed, a close-the-loop study should test whether focused instruction on double reference problems and/or more practice on right expanding expressions yields better learning transfer.

4. SUMMARY AND CONCLUSION

It is worth noting that the control condition in this study is highly similar to the treatment. Many might say, if you practice algebra, you learn algebra. Under that simple analysis, no differences should be expected between the conditions. Further, this control condition is a highly plausible instructional approach supported by a straightforward rational task analysis and by many colleagues who predict it should work: To prepare for story problems involving two operators, practice story problems involving one operator. The detailed data-driven quantitative cognitive task analysis suggested otherwise, in particular, that an inherent difficulty for algebra students learning to symbolize complex story problems is not in the story problem comprehension but in the production of more complex symbolic forms. Isolated practice in producing such forms, as the substitution problems provide, should enhance this hidden cognitive skill and yield better transfer. In a large data pool (711 students) collected in middle school math classes across four school years, our close-the-loop experiment demonstrated strong support for this data-driven prediction.

Our analysis also provides support for cognitive and statistical models that use the same underlying latent constructs (e.g., knowledge components) to predict both task difficulty and task-to-task transfer. This result is not only important to the science of learning, but it has practical relevance to the goal of using data-driven discoveries about domain learning challenges to design instruction for learning transfer. Task difficulty data can be more easily collected than task-to-task transfer data. Ideal transfer data (i.e., comparing performance on task B when task A is or is not practiced before it) requires giving students curriculum sequences that may harm their learning, therefore, it is costly and ethically challenging. Task difficulty data, when appropriately modeled, provides promise that these cost and ethical challenges can be minimized.

Although this paper does not present new data mining methods, it does indicate that attempts to automatically discover cognitive models, such as LFA [2] and others like it (e.g., Rule Space [22], Knowledge Spaces [24], and matrix factorization [6; 7]) can be used to generate instructional designs that improve student learning and transfer. While innovation in data mining methods is a crucial part of EDM research, it is important to the health of the field and its relevance to society that we pursue more close-the-loop studies and keep the E in EDM!

5. ACKNOWLEDGEMENTS

This work was supported in part by IES award #R305C100024 and NSF award #SBE-0836012.

6. REFERENCES

[1] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006). Adapting to When Students Game an Intelligent Tutoring System. In *Proc Int Conf Intelligent Tutoring Systems*, 392-401. Jhongli, Taiwan.

[2] 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.) *Proc 8th Int Conf ITS*, 164-175.

[3] Clark, R.E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2007). Cognitive task analysis. In J.M. Spector, M.D. Merrill, J.J.G. van Merriënboer, & M.P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593).

[4] Corbett, A.T. and Anderson, J.R. (1995). Knowledge decomposition and subgoal reification in the ACT programming tutor. In *Proc Artificial Intelligence and Education,1995*. Charlottesville, VA: AACE.

[5] Corbett, A.T., Anderson, J.R., Carver, V.H. and Brancolini, S.A. (1994). Individual differences and predictive validity in student modeling. In A. Ram & K. Eiselt (eds.) In *Proc Sixteenth Annual Conference of the Cog Sci Soc.*

[6] Desmarais MC. (2011). Mapping question items to skills with non-negative matrix factorization. *SIGKDD Explor, 13,* 30–36.

[7] Desmarais M.C. & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert- based Q-matrices. In *Proc Artificial Intelligence and Education*, *2013.* Memphis, TN, 441–450.

[8] Feldon, D. F., Timmerman, B. C., Stowe, K. A., & Showman, R. (2010). Translating expertise into effective instruction: The impacts of cognitive task analysis (CTA) on lab report quality and student retention in the biological sciences. *J Research in Sci Teaching, 47*(10), 1165–1185.

[9] Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy, *Cognitive Science, 7*, 155- 170.

[10] Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

[11] Heffernan, N. & Koedinger, K. R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In Shafto, M. G. & Langley, P. (Eds.) *Proc of the 19th Annual Conf Cog Sci Soc*, (pp. 307-312).

[12] Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. In *Proceedings of PME-NA*, pp. 21-49.

[13] Koedinger, K. R., & Anderson, J. R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments*, *5*, 161-180.

[14] Koedinger, K.R. & McLaughlin, E.A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.). *Proc 32nd Annual Conf Cog Sci Soc* (pp. 471-476.)

[15] Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated Student Model Improvement. In Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. (Eds.) *Proc 5th Int Conf on EDM*. (pp. 17-24)

[16] Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. In *Proc Int Conf on Artificial Intelligence in Education,* pp 421-430.

[17] Koedinger, K.R., & Tabachneck, H.J.M. (1995). Verbal reasoning as a critical component in early algebra. Paper presented at the annual meeting of the *American Educational Research Association*, San Francisco, CA.

[18] Koedinger, K. R., Yudelson, M., & Pavlik, P.I. (in press). Testing Theories of Transfer Using Error Rate Learning Curves. *Topics in Cognitive Science Special Issue.*

[19] Liu, R., Koedinger, K. R., & McLaughlin, E. A. (2014). Interpreting Model Discovery and Testing Generalization to a New Dataset. In Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds.) *Proc 7th International Conference on Educational Data Mining* (pp.107-113).

[20] Seamster, T.L., Redding, R.E., Cannon, J.R., Ryder, J.M., & Purcell, J.A. (1993). Cognitive task analysis of expertise in air traffic control. *Int J Aviat Psy, 3,* 257–283.

[21] Stamper, J. & Koedinger, K.R. (2011). Human-machine student model discovery and improvement using data. In Biswas, G., Bull, S., Kay, J. & Mitrovic, A. (Eds) *Proc 15th Int Conf, AIED 2011* (pp.353-360).

[22] Tatsuoka KK. (1983).Rule space: an approach for dealing with misconceptions based on item response theory. *J Educ Meas, 20,* 345–354.

[23] Velmahos, G. C., Toutouzas, K. G., Sillin, L. F., Chan, L., Clark, R. E., Theodorou, D., & Maupin, F. (2004). Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. *The American Journal of Surgery, 18*, 114-119

[24] Villano M. (1992). Probabilistic student models: Bayesian belief networks and knowledge space theory. *Proc 2nd Int Conf Intelligent Tutoring Systems,* NewYork: Springer-Verlag.