Using data-driven discovery of better student models to improve student learning

Kenneth R. Koedinger[1], John C. Stamper[1], Elizabeth A. McLaughlin[1], Tristan Nixon[2]

[1] Human-Computer Interaction Institute, Carnegie Mellon University
[2] Carnegie Learning, Inc.

Abstract. Deep analysis of domain content yields novel insights and can be used to produce better courses. Aspects of such analysis can be performed by applying AI and statistical algorithms to student data collected from educational technology and better cognitive models can be discovered and empirically validated in terms of more accurate predictions of student learning. However, can such improved models yield improved student learning? This paper reports positively on progress in closing this loop. We demonstrate that a tutor unit, redesigned based on data-driven cognitive model improvements, helped students reach mastery more efficiently. In particular, it produced better learning on the problem-decomposition planning skills that were the focus of the cognitive model improvements.

Keywords: data mining, machine learning, cognitive modeling

1 Introduction

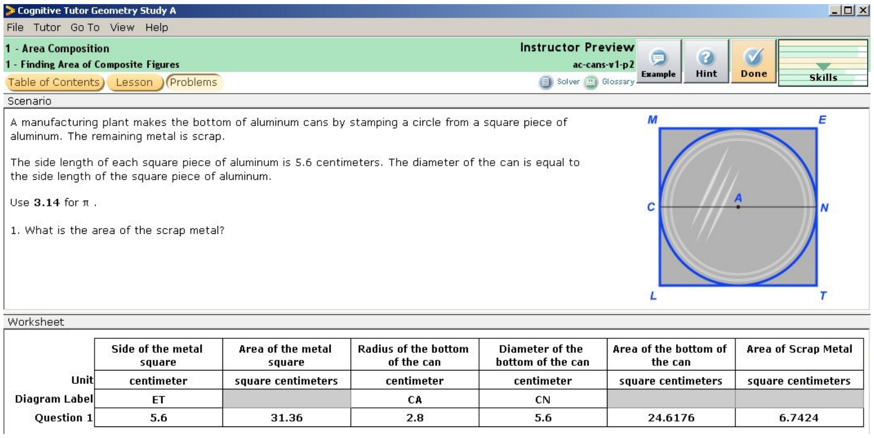
Much instruction is designed by intuition, drawing on the experiences and self-reflections of instructional designers or subject-matter experts. However, conscious access to our own knowledge is quite limited – estimated to be only about 30% of what we know [3]. The techniques of Cognitive Task Analysis (CTA), such as structured interviews of experts, can reveal such hidden knowledge. Furthermore, course redesign based on such analysis has been shown to improve student learning beyond that achieved by the original courses [3]. We have seen that greater levels of automation in CTA can be achieved by "mining" the log data from users of educational technology. By employing AI and statistical methods, better cognitive models have been discovered across multiple domains, and with student data from multiple technologies (intelligent tutors, online courses, games) [8]. This work is part of a related set of efforts to use data to discovery models of student knowledge and skill [1, 2]. One benefit of this data-driven approach to CTA is that it supplements human qualitative judgment with automated quantitative metrics that rigorously test purported cognitive model improvements. A critical next step is to the “close the loop” by using the improved cognitive models to redesign instruction and then to compare, in a controlled experimental study, whether the redesign produces better student learning than the original.

Past experiments testing the benefits for student learning of CTA-based course redesigns have had impressive results, but have typically taken a broad strokes approach to redesign [10; 3]. The redesigned “treatment” course usually differs from the original “control” course in many ways not all of which are clearly attributable to cognitive model improvements or to the insights obtained from CTA. One exception is a tightly controlled experiment within an algebra story problem symbolization tutor where the treatment differed from the control only in the replacement of one problem type (simpler story problems) with another (symbolic substitution problems) [6]. Prior CTA, employing the Difficulty Factors Assessment technique, had discovered the cognitive skills of composing symbolic expressions (e.g., if w=40x and y=800-w, then y=800-40x) as a particularly difficult component in learning to model story problems in algebraic symbols. The treatment was designed to isolate practice on these skills and led to improved learning over the control, including transfer from symbolic substitution to story problems [6].

The Difficulty Factors Assessment is a paper-based predecessor of our current educational technology data mining techniques for CTA; and while the symbolization study is a nice example of closing the loop, it does not provide direct evidence that data mining can be leveraged to produce better student learning. That is the goal of the current paper. Before presenting the experiment, we first review the CTA that led to the recommended improvements.

2 Using Educational Technology Data for Cognitive Task Analysis

In [11], we presented a data-driven method for researchers to use data from educational technologies to identify and validate improvements in a cognitive model. For statistical modeling purposes, we used a simplification of a cognitive model made up of hypothesized components of knowledge or skills that students must acquire to be successful on target assessment tasks or activities. These knowledge components (KCs) identify latent variables in a logistic regression model called the Additive Factors Model (AFM) [11], which is a generalization of item-response theory [12]. The method involves a wash-rinse-repeat iteration: 1) inspect learning curve visualizations and best-fitting parameters of AFM for a given set of knowledge components (a KC model), 2) hypothesize changes to the KC model based on identified problematic KCs, and 3) refit AFM with the new KC model and return to step 1.



A manufacturing plant makes the bottom of aluminum cans by stamping a circle from a square piece of aluminum. The remaining metal is scrap.

The side length of each square piece of aluminum is 5.6 centimeters. The diameter of the can is equal to the side length of the square piece of aluminum.

Use **3.14** for π .

1. What is the area of the scrap metal?

	Side of the metal square	Area of the metal square	Radius of the bottom of the can	Diameter of the bottom of the can	Area of the bottom of the can	Area of Scrap Metal
Unit	centimeter	square centimeters	centimeter	centimeter	square centimeters	square centimeters
Diagram Label	ET		CA	CN		
Question 1	5.6	31.36	2.8	5.6	24.6176	6.7424

Fig. 1. A scaffolded "composite area" problem from the original Geometry Cognitive Tutor. In the lower table, the student fills in all cell values except the row and column labels. The columns for the areas of the metal square and the bottom of the can are given to scaffold student reasoning toward finding the composite area of scrap metal. These square and circle columns (2 and 5) are absent in an unscaffolded composite area problem.

This method was applied to a publicly available data set from DataShop [5] called "Geometry Area (1996-97)." This data was generated by students using a Cognitive Tutor for learning geometry. A screen shot from a newer version of the tutor can be seen in Fig.1. The data included 5,104 student steps completed by 59 students. Using the visualizations available in DataShop, we identified potential improvements to the best existing KC model at the time we started, called Textbook-New, had 10 KCs. Three of the learning curves for these KCs are shown in Fig. 2. The lines represent the error rate (y-axis) averaged over all students for the first 20 practice opportunities for each KC. Most of the KCs in this model have reasonably smooth learning curves, like circle-area (some roughness in the learning curve can result from noise rather than a bad KC and particularly so when there are fewer observations being averaged, which is common at higher opportunity numbers.) The compose-by-addition curve is particularly jagged with upward blips at opportunities 12 and 15-18 where the curve jumps from about 25% to about 50%. Assuming there are particular problem steps that are more likely to occur at these opportunities (which is the case in this data set), those steps appear to have some knowledge demand that the other steps do not. The compose-by-addition KC involves "composite area problems", that is, problems where the area of a composite shape must be found by combining (adding or subtracting) the areas of two constituent regular shapes (e.g., what's left when a circle is cut from a square). In addition to the bumpy curve, the AFM parameter estimates indicate that compose-by-addition has no apparent learning (the slope parameter estimate is 0), yet it is associated with difficult tasks (the intercept parameter is 1.04 in log-odds, corresponding to a 26% error rate). The rough curve, flat slope, and non-trivial error rate are indications of a poorly defined KC.

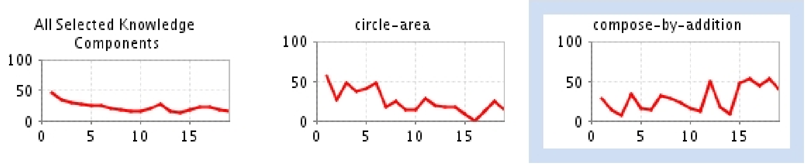


Fig. 2. Example learning curves where Y-axis is the error rate averaged across students (and KCs) and the X-axis is learning opportunities. Most curves, like the one for circle-area KC, are reasonably smooth and decreasing as indicated in the overall curve on the left. The curve for "compose-by-addition" is not smooth, with large jumps in the error rate particularly at opportunities 12 and 15.

A visualization of the error rates on problem steps tagged with compose-by-addition revealed that some steps are much harder than others. These steps may involve additional knowledge-demands that make them harder. By inspecting the problem content, we found that some of the composite problems were "scaffolded" such that they included columns that cued students to find the component areas first (see the square and circle columns in Fig. 1) [4]. Other problems were "unscaffolded" and did not start with such columns, thus students had to pose these sub-goals themselves. Indeed the blips in error rate for compose-by-addition (seen in the learning curve in Fig. 2) correspond with a high frequency of these more difficult unscaffolded problems. This analysis suggested that the compose-by-addition KC was not at a fine enough level to accurately explain the student data and that an alternative KC decomposition is needed. To improve the model, we split compose-by-addition into three KCs, one representing "*compose-by-addition*" with scaffolding present, a second where the student had to "*decompose"* a composite area without scaffolding, and a third where the student needs simply to "*subtract"* in order to execute a decomposition plan (formulated in a prior question within the same problem). In the new "DecomposeArith" KC model, the 20 steps that were previously labeled with the compose-by-addition KC are relabeled -- six with the new decompose KC, eight with the new subtract KC, and six keep the compose-by-addition KC label. The DecomposeArith model results in smoother, declining learning curves and, when fit with AFM, yields a significantly better prediction of student performance than the original.

To further validate the hypothesized model improvements, we performed a parallel analysis on a second Geometry Area data set also available in DataShop called "Geometry Area Hampton 2005-2006 Unit 34." The original Textbook student model associated with this data set had 13 KCs and when the steps for compose-by-addition were split into the three KCs as suggested above, a new DecomposeArith model was created with 15 KCs. Using AFM, we confirmed that this new model better predicts student data, reducing BIC (15,375 to 15,176) and root mean square error (RMSE) on test set fit in cross validation (.408 to .404) and thus supporting the existence of the new KCs.

The next step was to use the discovered model to improve the instruction in the cognitive tutor unit.

3 Redesigning the Geometry Cognitive Tutor

An improved cognitive model can be used in multiple possible ways to redesign a tutor:

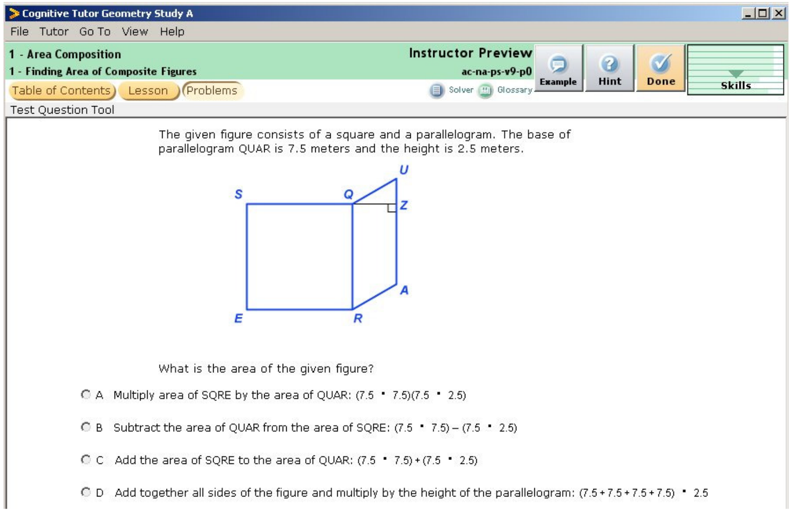
1) Resequencing – position problems requiring fewer KCs before ones needing more earlier
2) Knowledge tracing – add/delete skill bars for better cognitive mastery
3) Creating new tasks – add problems to focus practice on new KCs
4) Changing instructional messages, feedback or hint messages

We applied the improved model to the Geometry area unit of a high school geometry course. The improved model's new KCs are related to the planning of problem decomposition. We added three new skills to the tutor that differentiate unscaffolded decomposition, scaffolded, and simple addition/subtraction. These new skills resulted in changes to knowledge tracing and led to the creation of new tasks. In particular, students in the new version are not given credit for the difficult decomposition planning step via success on simpler scaffolded or subtraction steps, but only through success on unscaffolded composition steps.

We also added new problems to better target these newly identified skills. In our first attempt at redesign (briefly described in [11]), we identified four types of problems: unscaffolded, table scaffolded, area scaffolded, and problem statement scaffolded. Table scaffolded problems reflect the current setup in the tutor and include columns for intermediate areas (as in Fig. 1). Unscaffolded problems remove the columns for intermediate areas. Area scaffolded problems give the areas of the component shapes. Problem statement scaffolded problems have the same table as the unscaffolded problems but provide an explicit hint in the problem statement directing the student to first find the component areas. During the implementation of this first redesign attempt [11], we experienced some issues with the parameter settings and knowledge tracing algorithm which resulted in students never mastering all skills. We also found that the problem statement scaffolded problems did not seem to help the students learn the KCs, so we removed this type of problem in the next design iteration.

More importantly, inspired by related work [6], we realized there was an opportunity to better support students' learning of the hardest skill, the decomposition planning skill that recognizes a composite area is being sought and sets sub-goals to find it by first finding the component areas. We called this the "know to pose" skill and it always appeared with other skills on problem steps in the first redesign. The design challenge was to create a problem (or step) that makes visible and isolates just this "know to pose" skill. Our solution, shown in Fig. 3, was to ask students to come up with a plan to solve an unscaffolded composite area problem and recognize a correct description of such a plan.

In general, changes in skills can lead to changes in the feedback and hint messages the tutor provides. Thus, the new problems also come with new, more focused, context-sensitive instruction that follows directly from the cognitive model improvements.



The given figure consists of a square and a parallelogram. The base of parallelogram QUAR is 7.5 meters and the height is 2.5 meters.

What is the area of the given figure?

○ A Multiply area of SQRE by the area of QUAR: (7.5 · 7.5)(7.5 · 2.5)

○ B Subtract the area of QUAR from the area of SQRE: (7.5 · 7.5) – (7.5 · 2.5)

○ C Add the area of SQRE to the area of QUAR: (7.5 · 7.5) + (7.5 · 2.5)

○ D Add together all sides of the figure and multiply by the height of the parallelogram: (7.5 + 7.5 + 7.5 + 7.5) · 2.5

Fig. 3. Example of new problem type to isolate the know-to-pose KC. Students need to perceive the desired irregular area as being composed of areas of regular shapes and then devise a decomposition plan for solving for the irregular area. They do not need to execute the plan, but rather recognize a description of it.

To implement the new tutor, we needed to set the Bayesian Knowledge Tracing parameters for the new KCs. We set them by hand based on the available data, while recognizing the possibility of introducing differences between the experimental conditions. Given the introduction of more KCs, we wanted to avoid students in the treatment spending more time than the control, so we tried to err in the direction of more lenient settings (i.e., a higher initial probability of knowing a new KC). As it turned out, these settings were not too low as treatment students better learned decomposition skills than control students.

We also implemented a "minimizing" problem-selection algorithm which would help focus student practice by selecting problems with the fewest unmastered skills. This new algorithm is in contrast with the standard algorithm which selects problems that maximize a student's opportunity to practice unmastered skills.

4 Experiment

We performed an *in vivo* experiment comparing the redesigned tutor ("treatment") with the existing tutor ("control"). The study was run with 103 students (52 control, 51 treatment) as part of regular geometry classes in a local suburban high school in the Fall of 2011. Due to absenteeism, seven students did not complete the posttest and were excluded from our analyses leaving 96 students (48 control, 48 treatment).

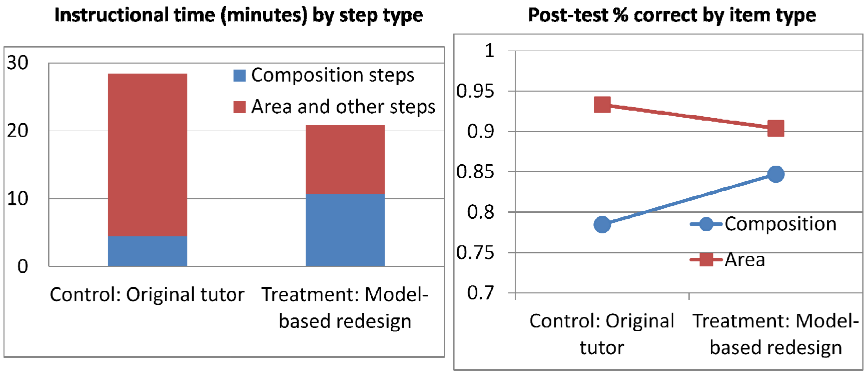
Pre- and post-test measures were paper and pencil and included two versions (A and B) and two orders (four forms) with 12 problems each (5 area, 6 composition, and 1 compare - a qualitative judgment of the relative area of two related figures). The forms (A1, A2, B1, and B2) were randomly assigned for both pre and posttest. For each version, the cover stories, constants and sequence of problems varied but the shapes remained the same.

The treatment had one problem type, unscaffolded problems, that are harder than the table scaffolded problems used in the control and are more genuinely representative of the desired problem solving. The treatment also had two other problem types, area scaffolded and decomposition planning (as in Fig. 3), that are less complex, involving fewer steps but better isolating the critical decomposition skills. The intention was that these problems would more efficiently focus student learning on these skills, minimize distraction from and time spent on other skills, and better prepare students for unscaffolded problem solving practice. Thus, we hypothesized students would learn decomposition skills more effectively and more efficiently, that is, at a faster rate.

As shown in Fig. 4a, indeed, the treatment students mastered the required skills in much less time on average (20.9 minutes) than the control (28.4 minutes; see Fig. 1a). An ANCOVA with pre-test as a covariate found this difference to be statistically reliable (F (1, 93) = 4.6, p = .03) and an effect size (Cohen's *d)* of .6 indicates that it is substantial. Interestingly, despite taking 26% less time, the treatment students solved more problems (14.0 per student) than control students (10.4). We discuss later the reasons behind the treatment's faster completion of problems. We confirmed that all students mastered all knowledge components (8 in the treatment and 6 in the control) according to the Cognitive Tutor's Bayesian Knowledge Tracer (p_{known} > .95).

We must be cautious in using the tutor data alone to conclude that treatment students learned at a faster rate. The mastery criteria employed by the two tutors was different, based on different cognitive models. The post-tests, however, were the same and provide a more clearly comparable assessment of student achievement and its transfer from the computer environment to paper. We find, indeed, that the treatment did just as well on the posttest (M = 86.6% correct) as the control (M = 85.5%). An ANCOVA with pre-test as a covariate finds no reliable post-test difference by condition (F(1, 93) = 1.03, p=.31). The cognitive model differences in the two tutors suggest we should see a different pattern of performance on the post-test, with better performance of the treatment on composition problems. As Fig. 4b shows we find just such a pattern. We performed a MANOVA with condition as a factor and two separate post-test sub-scores, one for the decomposition problems and one for the pure area problems, as the dependent variables. Indeed the condition by problem-type interaction apparent in Fig. 4 is significant (F (1, 94) = 4.05, p = .047).

In fact, treatment students better performance on the composition items on the post-test may be underestimated in that many of the items were easier scaffolded composition problems. One of the problems in particular (the PIZZA problem) was an unscaffolded composition problem (it seeks the area after removing a circle inscribed in a square). We expected it to be the hardest problem on the test and indeed it was (pretest = 59%, average all pretest = 80%). The pre to post results are striking: the control shows little difference, a 5% gain (.50 to .55), whereas the treatment has an 18% gain (.67 to .85). This difference is consistent with the hypothesis that the redesigned tutor enables better learning of the challenging problem decomposition skills.



(a) (b)

Fig. 4. Students using the redesigned tutor reached mastery a) in significantly less time (21 vs. 28 minutes) while actually spending more time on the critical decomposition skills and b) better learned these decomposition skills as demonstrated by better post-test performance on composition problems.

Toward better explaining the faster learning rate in the treatment, we also disaggregated the instructional time into time spent on composition steps versus other steps (e.g., finding area, entering givens, doing algebra). On average, treatment students spent less time on other steps (10.2 minutes) than control students (24.0 minutes). However, treatment students actually spent more time on composition steps (10.7 minutes) than the control students (4.5 minutes). A MANCOVA with pretest as a covariate and instructional time on decompose steps and other steps as the dependent measures confirmed the condition by step-type interaction to be significant, F (1, 93) = 140, p < .0001. These time differences are largely a consequence of different numbers of assigned steps. In particular, treatment students did fewer other steps on average than the control (173 vs. 224) and more composition steps (40.8 vs. 29.1). These differences reflect the cognitive model differences in the two tutors and, in particular, the model-based design of problems in the treatment to efficiently isolate decomposition skills and to minimize time spent on other skills.

5 Discussion and Conclusion

Following our past demonstrations that better cognitive models of students can be discovered from data [8; 11], we have tested the hypothesis that using an improved model to redesign an adaptive tutor yields better student learning. The evidence supports the hypothesis. In particular, we found students using the redesigned tutor reached mastery (as demonstrated within the tutor and on a post-test) in significantly less time than students using the original tutor. Despite needing less overall time, the redesigned tutor had treatment students spending more time than control students on the difficult problem decomposition planning skills that were identified by way of a semi-automated Cognitive Task Analysis process. These students performed better on the targeted composition problems on the post-test.

It appears from the post-test results, that the treatment may not have gotten optimal practice on some area skills. For example, the treatment did not do as well on trapezoid area problems on the post-test. Unlike immediately prior units that differentiate individual area skills (e.g., rectangle vs. circle vs. trapezoid), this composite area unit had a single "individual area" KC for all regular shapes. We know from prior model search that this merged KC is too coarse and would benefit from being split into more specific KCs. Doing so, we suspect, would yield further improvements in student learning from this composition unit. Students using such a further redesigned unit should still do many fewer area steps overall than in the current control, but would get more as-needed practice on harder area skills, like trapezoid area, than the current treatment.

A related limitation of the current "close-the-loop" demonstration is that the redesigns follow from a KC model that, while validated statistically, was proposed from human inspection of learning curve data [13]. It would further strengthen the argument for this approach to have other demonstrations of close-the-loop success in other domains where LFA has achieved KC model discoveries through more automatic methods [8].

It may be tempting to conclude that "students learn what they spend time on", but this simple statement is dangerously misleading. It depends critically on how we categorize student activities. *All* of the problems that both groups solved in this study were composition problems, and the control group spent more time on these problems overall. Thus, by the simple statement, they should have learned the decomposition skills better. They did not. A finer grained cognitive analysis of student activity tells a different story -- one that matches the data! We need to categorize problem-solving steps, not problems, and we need to do so with respect to their cognitive demands, recognizing that different contexts for the same action require students to acquire different knowledge [13]. Our prior model discovery revealed a different skill is needed for unscaffolded composition steps than for scaffolded ones.

The phrase "how we categorize student activities" is another way of saying "cognitive model". Students learn the elements (the knowledge components) of the cognitive model they spend time practicing. However, the structure of that model is not obvious. Knowledge components are not directly observable and most are not open to conscious reflection, despite our strong feelings of self-awareness of our own cognition [3]. They can, however, be inferred and discovered from student performance data across multiple tasks [cf., 7] via a statistical comparison of alternative categorizations, that is, of alternative cognitive models.

Thus, it is a great opportunity for AI and Education not only in mining educational technology data to discover better cognitive models, but in closing the loop by redesigning systems based on the resulting insights and testing them toward achieving better student learning.

Acknowledgement. This work was supported by LearnLab, the Pittsburgh Science of Learning Center (NSF award 0836012). We thank Gail Kusbit, Steve Ritter, and the rest of LearnLab for their help on this project.

References

1. Barnes, T., The Q-matrix Method: Mining Student Response Data for Knowledge. In J Beck, (ed) Proceedings of AAAI 2005: Educational Data Mining Workshop.
2. Beheshti, B., Desmarais, M. & Naceur, R. (2012). Methods to find the number of latent skills. Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. (eds.) *Proceedings of the 5th International Conference on Educational Data Mining.* (pp.81-86) Chania, Greece.
3. Clark, R. E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2007). Cognitive task analysis. In J.Spector, M.Merrill, J.vanMerriënboer, & M.Driscoll (Eds.), *Handbook of research on educational communications and technology,*pp. 577–593. Mahwah, NJ.
4. Corbett, A.T. & Anderson, J.R. (1995) Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, pp. 253-278.
5. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010) A Data Repository for the EDM commuity: The PSLC DataShop. In Romero, Ventura, Pechenizkiy, Baker, (Eds.) Handbook of Educational Data Mining. CRC Press. (http://learnlab.org/datashop)
6. Koedinger, K.R. & McLaughlin, E.A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. (pp. 471-476.) Austin, TX: Cognitive Science Society.
7. Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36* (5), 757-798.
8. Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated Student 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.
9. Lee, R. L. (2003). Cognitive task analysis: A meta-analysis of comparative studies. Unpublished doctoral dissertation, University of Southern California, Los Angeles.
10. Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*. http://jime.open.ac.uk/2008/14
11. Stamper, J., & Koedinger, K.R. (2011) Human-machine student model discovery and improvement using DataShop. In Kay, J., Bull, S. and Biswas, G. (eds.) *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED2011). pp. 353-360. Berlin Germany:Springer.
12. Wilson, M. & de Boeck, P. (2004). Descriptive and explanatory item response models. In P. de Boeck & M. Wilson (eds.), *Explanatory item response models* (pp. 43-74). Springer.
13. Zhu, X. & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137-166.