Chapter 6: Going Beyond Better Data Prediction to Create Explanatory Models of Educational Data

Ran Liu, Kenneth R. Koedinger

School of Computer Science, Carnegie Mellon University, USA

DOI: 10.18608/hla17.006

ABSTRACT

In the statistical modelling of educational data, approaches vary depending on whether the goal is to build a predictive or an explanatory model. Predictive models aim to find a combination of features that best predict outcomes; they are typically assessed by their accuracy in predicting held-out data. Explanatory models seek to identify interpretable causal relationships between constructs that can be either observed or inferred from the data. The vast majority of educational data mining research has focused on achieving predictive accuracy, but we argue that the field could benefit from more focus on developing explanatory models. We review examples of educational data mining efforts that have produced explanatory models and led to improvements to learning outcomes and/or learning theory. We also summarize some of the common characteristics of explanatory models, such as having parameters that map to interpretable constructs, having fewer parameters overall, and involving human input early in the model development process.

Keywords: Explanatory models, model interpretability, educational data mining (EDM), closing the loop, cognitive models

Across the vast majority of educational data mining research, models are evaluated based on their predictive accuracy. Most often, this takes the form of assessing the model's ability to correctly predict successes and failures in a set of student response outcomes. Much less commonly, models may be validated based on their ability to predict post-test outcomes (e.g., Corbett & Anderson, 1995) or pre-test/post-test gains (e.g., Liu & Koedinger, 2015).

While predictive modelling has much to recommend it, the field of educational data mining could benefit from more emphasis on developing explanatory models. Explanatory models seek to identify interpretable constructs that are causally related to outcomes (Shmueli, 2010). In doing so, they provide an explanation of the data that can be connected to existing theory. The focus is on *why* a model fits the data well rather than only *that* it fits well. Often, explanatory models provide an interpretation of the data that has implications for theory, practice, or both. Here, we review educational data mining efforts that have produced explanatory models and, in turn, can lead to improvements to learning outcomes and/or learning theory.

Educational data mining research has largely focused on developing two types of models: the statistical model and the cognitive model. Statistical models drive the outer loop of intelligent tutoring systems (VanLehn, 2006) based on observable features of students' performance as they learn. Cognitive models are representations of the knowledge space (facts, concepts, skills, et cetera) underlying a particular educational domain. The majority of the research reviewed here concerns cognitive model refinement and discovery. We also briefly review other examples of explanatory models outside the realm of cognitive model discovery that educational data mining research has produced.

COGNITIVE MODEL DISCOVERY

Cognitive models map knowledge components (i.e., concepts, skills, and facts; Koedinger, Corbett, & Perfetti, 2012) to problem steps or tasks on which student performance can be observed. This mapping provides a way for statistical models to make inferences about students' underlying knowledge based on their observable performance on different problem steps. Thus, cognitive models are an important basis for the instructional design of automated tutors and are important for accurate assessment of learning and knowledge. Better cognitive models lead to better predictions of what a student knows, allowing adaptive learning to work more efficiently. Traditional ways of constructing cognitive models (Clark, Feldon, van Merriënboer, Yates, & Early, 2008) include structured interviews, think-aloud protocols, rational analysis, and labelling by domain experts. These methods, however, require human input and are often time consuming. They are also subjective, and previous research (Nathan, Koedinger, & Alibali, 2001; Koedinger & McLaughlin, 2010) has shown that expert-engineered cognitive models often ignore content distinctions that are important for novice learners. Here, we review three examples of efforts to discover and refine cognitive models based on data-driven techniques that alleviate expert bias while reducing the load on human input.

For statistical modelling purposes, the work described here uses a simplification of a cognitive model composed of hypothesized knowledge components. A knowledge component (KC) is a fact, concept, or skill required to succeed at a particular task or problem step. We refer to this specialized form of a cognitive model as a KC model or, alternatively, a Q-matrix (Barnes, 2005). The statistical model we used to evaluate the predictive fit of data-driven cognitive model discoveries is a logistic regression model called the additive factors model (AFM; Cen, Koedinger, & Junker, 2006), a generalization of item-response theory to accommodate learning effects.

Data-Driven Cognitive Model Improvement

Difficulty factors assessment (DFA; e.g., Koedinger & Nathan, 2004) moves beyond expert intuition by using a data-driven knowledge decomposition process to identify problematic elements of a defined task. In other words, when one task is much harder than a closely related task, the difference implies a knowledge demand of the harder task that is not present in the easier one. Stamper and Koedinger (2011) illustrated a method that uses DFA, along with freely accessible educational data and built-in visualization tools on DataShop[1] (Koedinger et al., 2010), to identify and validate cognitive model improvements. The method for cognitive model refinement iterates through the following steps: 1) inspect learning curve visualizations and fitted AFM coefficient estimates for a given KC model, 2) identify problematic KCs and hypothesize changes to the KC model, 3) re-fit the AFM with the revised KC model and investigate whether the new model fits the data better.

Through manual inspection of the visualizations of a geometry dataset (Koedinger, Dataset 76 in DataShop[2]), potential improvements to the best existing KC model at the time were identified (Stamper & Koedinger, 2011). Most of the KCs in this model exhibited relatively smooth learning curves with a consistent decline in error rate. One KC in the original model, *compose-by-addition*, exhibited a particularly noisy curve with large spikes in error rate at certain opportunity counts. In addition, the AFM parameter estimates for the *compose-by-addition* KC suggested no apparent learning (the slope parameter estimate was very close to zero, and not because the performance was at ceiling). A bumpy learning curve and low slope estimate are indications of a poorly defined KC. One common cause for a poorly defined KC is that some of its constituent items require some knowledge demand that other items do not. In other words, the original KC should really be split into two different KCs. To improve the KC model, all *compose-by-addition* problem steps were examined, and domain expertise was applied to hypothesize about additional knowledge that might be required on certain steps. As a result, the *compose-by-addition* KC was split into three distinct KCs, and each of the 20 steps previously labelled with the *compose-by-addition* KC were relabelled accordingly. The revised model resulted in smoother learning curves and, when fit with the AFM, yielded significantly better predictions of student performance than the original KC model did. Although this KC model improvement was aided by visualizations resulting from fitting a statistical model, the actual improvements were generated manually and thus were readily interpretable.

The discovered KC model improvements had clear implications for revising instruction. Koedinger, Stamper, McLaughlin, and Nixon (2013) used the data-driven KC model improvements to generate a revised version of the Geometry Area tutor unit. Revisions included adding the newly discovered skills to the KC model driving adaptive learning, resulting in changes to knowledge tracing, and the creation of new tasks to target the new skills. In an A/B experiment, half of the students completed the revised tutor unit and the other half

[1] http://pslcdatashop.org

[2] Geometry Area 1996–1997: https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=76

competed the original tutor unit. Students using the revised tutor reached mastery more efficiently and exhibited better learning on the skills targeted by the KC model improvement, based on pre- to post-test gains (Koedinger et al., 2013). These results show that the data-driven DFA technique lends itself to generating explanatory KC model refinements that can result in instructional modifications and improved learning outcomes.

Learning Factors Analysis

Learning factors analysis (LFA; Cen et al., 2006) was developed to automate the data-driven method of KC model refinement to further alleviate demands on human time. LFA searches across hypothesized knowledge components drawn from different existing KC models, evaluates different models based on their fit to data, and outputs the best-fitting KC model in the form of a symbolic model. As such, LFA greatly reduces demands on human effort while simultaneously easing the burden of interpretation, even if it does not automatically accomplish it.

We applied the LFA search process across 11 datasets spanning different domains and different educational technologies, all publicly available from DataShop. Across all 11 datasets, this automated discovery process improved KC models' fit to data beyond the best existing human-tagged KC models (Koedinger, McLaughlin, & Stamper, 2012). Importantly, we demonstrated in an example dataset (Koedinger, Dataset 76 in DataShop) an interpretable explanation for the specific improvements made by the best LFA-discovered model. A manual KC model comparison between the best-fitting LFA model and the best-fitting human-tagged model revealed that the LFA model tagged separate KCs for forwards (i.e., find area given radius) and backwards (i.e., find radius given area) circle area problems, whereas these had been grouped together as a single "circle-area" KC in the human-tagged model. No such differences were found between the models for other shapes like rectangles, triangles, and parallelograms. Applying domain expertise to interpret the automated discovery, we hypothesized that LFA's model improvement may have captured the difficulty of knowing when and how to apply a square root operation for backwards circle-area problems, which is not required for forwards circle-area problems nor for the backwards area problems of other shapes.

We then assessed the external validity of this interpretation beyond the dataset from which the discoveries were made. We evaluated the presence of the square root difficulty in a novel dataset (Bernacki, Dataset 748 in DataShop[3]), one with a different structure from that used to make the discovery (Liu, Koedinger, & McLaughlin, 2014). Among other differences, the novel dataset contained more backwards circle-area problems and, importantly, forwards (i.e., find area given side length) and backwards (i.e., find side length given area) *square-area* problems. These square-area problems were not at all present in the original dataset from which the LFA-generated discovery was made. Applying our interpretation of the discovery, we constructed a KC model that tags separate forwards and backwards KCs only for shapes where backwards steps require computing a square root (squares, circles) but not for shapes where backwards steps don't (triangles, rectangles, parallelograms). When used in conjunction with the AFM, this KC model yielded the best fit to the novel dataset compared to several expert-tagged control KC models.

Since the novel dataset had a different structure from the original dataset, including differences relevant to the KC model discovery (i.e., existence of backwards square-area problems), it would not have been viable to apply directly the LFA-discovered KC model on this new dataset. Interpretation is necessary in order to test the generalizability of 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. Our current research is "closing the loop" on this LFA-generated discovery by assessing learning outcomes resulting from a tutor redesigned around the improved KC model (Liu & Koedinger, submitted).

Automated Cognitive Model Discovery Using SimStudent

An alternative automated approach uses a state-of-the-art machine-learning agent, SimStudent, to discover cognitive models automatically without requiring existing ones. SimStudent is an intelligent agent that inductively learns knowledge, in the form of rules, by observing a tutor solve sample problems and by solving problems on its own and receiving feedback (Li, Matsuda, Cohen, & Koedinger, 2015). One of the benefits of SimStudent is that it can simulate features of novices' learning trajectories of which domain experts may not even be aware. Real students entering a course do not usually have substantial domain-specific prior knowledge, so a realistic model of human learning ought not to assume this knowledge is given. In addition, SimStudent can be used to test alternative models of human learning to see which best predicts human behaviour (MacLellan, Harpstead, Patel, & Koedinger, 2016). For several datasets spanning various domains, SimStudent generated cognitive models that fit the data better than the best human-generated cognitive

[3] Motivation for learning HS geometry 2012 (geo-pa): https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=748

models (Li et al., 2011; MacLellan et al., 2016).

The output of the SimStudent's learning takes the form of production rules (Newell & Simon, 1972), and each production rule essentially corresponds to one knowledge component (KC) in a KC model. Using data from an Algebra dataset (Booth & Ritter, Dataset 293 in DataShop[4]) and in conjunction with the AFM, Li and colleagues (2011) compared a KC model generated by SimStudent to a KC model generated by hand-coding actual students' actions within the tutor. The SimStudent-generated model better fit the actual student performance data than the human-generated model did.

More importantly, inspecting the differences between the SimStudent model and the human-generated model revealed interpretable features that explained the advantages of the SimStudent model. One example of such a difference is that SimStudent created distinct production rules (KCs) for division-based algebra problems of the form Ax=B, where both A and B are signed numbers, and for the form −x=A, where only A is a signed number. To solve Ax=B, SimStudent learns to simply divide both sides by the signed number A. But, since −x does not represent its coefficient (−1) explicitly, SimStudent must first recognize that −x translates to −1x, and then it can divide both sides by −1. The human-generated model predicts that both forms of division problems should have the same error rates. In fact, real students have greater difficulty making the correct move on steps like −x = 6 than on steps like −3x = 6. Within the same Algebra dataset, problems of the form Ax=B (average error rate = 0.28) are easier than problems of the form −x=A (average error rate = 0.72). SimStudent's split of division problems into two distinct KCs suggests that students should be tutored on two subsets of problems, one subset corresponding to the form Ax=B and one subset specifically for the form −x=A. Explicit instruction that highlights for students that −x is the same as −1x may be beneficial (Li et al., 2011).

We hypothesized that the interpretation of this particular SimStudent KC model discovery would generalize to novel problem types, just as the LFA-generated model discovery did. In a novel equation-solving dataset (Ritter, Dataset 317 in DataShop[5]), we tested whether the explicit vs. implicit coefficient distinction similarly applied to *combine like terms* problems. We looked at differences in performance for items of the form Ax + Bx = C, where both A, B, and C are signed numbers (explicit-coefficient items), and items where either A or B were equal to 1 or −1 with the coefficient perceptually absent (implicit-coefficient items). This analysis confirmed that explicit-coefficient items (average error rate = 0.35) are easier than implicit-coefficient items (average error rate = 0.45) among *combine like terms* problems. This new dataset not only replicated the original finding that SimStudent made on divide problems, but it also revealed that the finding generalizes to a separate procedural skill, *combine like terms*.

Fitting a KC model with separate KCs for the explicit- vs. implicit-coefficient forms of *combine like terms* items revealed a large improvement in predictive fit relative to a KC model with a single *combine like terms* KC. Furthermore, although the learning curves for both the explicit-coefficient divide and combine like terms KCs reflected smooth and decreasing error rates, the respective learning curves for implicit-coefficient *divide* and *combine like terms* items were both flat, with slopes close to zero. This suggests that students would benefit greatly from more practice on, and more explicit attention to problem steps involving implicit coefficients. Here, again, the explanatory power of the SimStudent KC model discovery made it possible to generalize the explanation to distinct problem types on which SimStudent was never trained.

Comparison to Other Work

Both LFA and SimStudent are capable of producing cognitive model discoveries that not only significantly improve predictive accuracy but are readily interpretable and, thus, explanatory. We have demonstrated that the interpretations yielded by these cognitive model discoveries generalize to novel problem types not present in the data from which the discoveries were made. Finally, they produce clear recommendations for revising instruction, even in contexts that are very different from those in which the original data were collected. These are all hallmarks of explanatory modelling efforts that move beyond simply improving predictive accuracy to have meaningful impact on learning theory and instruction.

The fact that methods like LFA are "human-in-the-loop" – that is, requiring input from a domain expert – has been cited as a limitation. In the case of LFA, one or more expert-tagged cognitive models are required initially in order to produce new model discoveries. We argue, however, that this "human-in-the-loop" feature leads the results of such modelling efforts to be explanatory. There have been a number of recent efforts to fully automate the process of discovering and/or improving cognitive models (González-Brenes & Mostow, 2012; Lindsey, Khajah, & Mozer, 2014). These methods have much to recommend, as they dramatically reduce demands on human time and produce competitive results in predictive accuracy. However, the resulting cognitive models of these efforts have

[4] Improving skill at solving equations via better encoding of algebraic concepts (2006–2008): https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=293

[5] Algebra I 2007–2008 (Equation Solving Units): https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=317

not been interpreted or acted upon with respect to improving instruction.

Other modelling efforts, including a "human-in-the-loop" component like Ordinal SPARFA-Tag (Lan, Studer, Waters, & Baraniuk, 2013), have yielded considerably more interpretable cognitive models than many alternative methods. Although humans must do any final interpretation of modelling efforts, methods like LFA and Ordinal SPARFA-Tag greatly improve the likelihood of generating sensible resulting models by incorporating the human effort up front. In fact, comparing the original SPARFA model (Lan, Studer, Waters, & Baraniuk, 2014), which only incorporates concept tags post-hoc, to Ordinal SPARFA-Tag, which incorporates domain-expert concept tags in the model development process up front, shows that the latter model results in much more interpretable cognitive models.

More attention and effort towards generating interpretable cognitive models is, in our view, progress in the right direction. Nevertheless, as we have argued, expert labelling is still subject to biases and does not offer much to advance learning theory using the rich educational data available. Human involvement improves interpretability, whereas the data-driven component offers ways to alleviate subjective biases and advance our understanding of how novices learn. Methods such as LFA leverage both the unique strengths of human involvement and of automation towards creating models that are more predictive *and* explanatory.

STUDENT GROUPING

A growing body of research suggests that modelling student-specific variability in statistical models of educational data can yield better predictive accuracies and potentially inform instruction. Prior attempts to group students based on features available in educational datasets have focused on techniques such as K-means and spectral clustering. These techniques have been used to generate student clusters predictive of post-test performance (Trivedi, Pardos, & Heffernan, 2011) and that yield predictive accuracy improvements when clusters are fit with different sets of parameters (Pardos, Trivedi, Heffernan, & Sárközy, 2012). Many clustering techniques, however, tend to result in student groupings that are difficult to interpret. Yet, interpretation is critical if the results of clustering are to eventually inform improvements in instructional policy (e.g., individualizing instruction appropriately to different groups of students).

In recent research (Liu & Koedinger, 2015), we developed a method for grouping students that not only dramatically improves the predictive accuracy of the AFM but inherently lends itself to producing meaningful student groups. By doing a first-pass fit of the AFM to the data and examining systematic patterns in the residuals (differences between predicted and actual data) across different practice opportunities, we consistently found students belonging to one of three learning rate groups: 1) those who exhibit flatter learning curves than the AFM predicts, 2) those who exhibit steeper learning curves, and 3) those whose learning curves are on par with the model's predictions. Introducing a parameter that individualizes learning rates to each of these learning rate groups substantially improves model predictive accuracy, beyond that of the regular AFM, across a variety of datasets spanning multiple educational domains. Across datasets, the slope parameter estimates for each of the three groups were consistent with our interpretation of the groups (i.e., the estimated group-level slopes were always lowest for the flat-curve group, and highest for the steep-curve group). Furthermore, in a subset of datasets for which there exist paper pre- and post-test data, we observed a systematic relationship between learning-curve group and the degree of pre- to post-test improvement (Liu & Koedinger, 2015).

Unlike other, more "bottom-up" methods of creating stereotyped groups of students, this method yielded student groups that are readily interpretable and potentially actionable. For example, it is clear that the flat-curve student group represents either students who are already performing at ceiling when they start the unit or curriculum (and thus do not have much room for improvement) or students who are starting anywhere below ceiling but struggling to progress with the material. In either case, there are clear instructional implications for students classified into this group. The explanatory power of the resulting model again benefitted from doing some up-front interpretation and developing the model with an eye towards interpretability.

TOWARDS BUILDING EXPLANATORY MODELS

We argue for the importance of considering the interpretability and actionability of educational data mining efforts in producing more explanatory models. For a model to be explanatory, one should be able to understand *why* the model achieves better predictive accuracy than alternatives. In addition, the understanding of this *why* should either advance our understanding of how learners learn the relevant material or have clear implications for instructional improvements, or both. We summarize by outlining some of the features that tend to characterize explanatory models.

Explanatory modelling efforts tend to start with "clean" independent variables that have either simple functions or map to clearly defined constructs. For example, LFA derives new variables from existing, expert-labelled variables using simple split, merge, or add operators. Another example comes from automated analyses of verbal data in education, a branch of educational data mining that includes automated essay scoring, producing tutorial dialogue, and computer-supported collaborative learning. A major consideration in this area is how to transform raw text or transcriptions into features that can be used in a machine-learning algorithm. Approaches to this issue range from simple "bag of words" methods, which counts the frequency of each word present in the text, to much more sophisticated linguistic analyses. One consistent theme across findings is that feature representations motivated by interpretable, theoretical frameworks have been among the most promising (Rosé & Tovares, in press; Rosé & VanLehn, 2005). Thus, incorporating some human time and thought into defining and labelling these independent variables up front can greatly improve the explanatory power of the resulting model.

Another feature of explanatory models, one that relates most to actionability, is that the dependent variable maps to a well-defined construct. The work on learning rate groups is an example of this: since the groups to which students are classified are defined up front, it is clear what it means for a student to be in the "flat" learning curve group, as opposed to the "steep" one. This makes the results from modelling readily actionable. Another body of research in which the dependent variable tends to be well mapped to an interpretable construct is the modelling of affect and motivation using features of tutor log data. These techniques use pre-defined psychological or behavioural constructs, measured through questionnaires or expert observations, to develop and refine "detectors" that can identify those constructs within tutor log data activity (e.g., Winne & Baker, 2013; San Pedro, Baker, Bowers, & Heffernan, 2013; D'Mello, Blanchard, Baker, Ocumpaugh, & Brawner, 2014). The "detectors" are developed specifically to identify pre-determined constructs and, thus, the results of these algorithms are readily actionable. For example, Affective AutoTutor is an intelligent tutoring system for computer literacy that automatically models students' confusion, frustration, and boredom in real time. Detection of these affective states is then used to adapt the tutor actions in a manner that responds accordingly. An experimental study "closing the loop" on this affective detector showed higher learning gains for low-domain knowledge students who interacted with the Affective AutoTutor compared to a non-affective version (D'Mello et al., 2010). For these modelling efforts to be fully explanatory though, interpretations of the independent variables driving the affective outcomes are also needed.

Finally, explanatory models tend to be characterized by fewer estimated parameters (independent variables, or features). For example, the AFM has only one parameter for each student and two parameters for each knowledge component. Adding learning rate groups extends the model by only one additional parameter, group membership. This makes the contribution of the added parameter easy to attribute and interpret. Having fewer parameters also allows each parameter's estimates to have more explanatory power, alleviating issues of indeterminacy. Because the AFM has only one difficulty parameter and one learning parameter for each KC, one can, for example, meaningfully interpret a low learning parameter estimate as suggesting that KC needs either refinement or instructional improvement.

We have illustrated some ways in which concrete steps in the design of educational data modelling efforts can yield more explanatory models. The relationships between the fields of educational data mining, learning theory, and the practice of education could be greatly strengthened with increased attention to the explanatory power of models and their ability to influence future learning outcomes.

REFERENCES

Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. *Proceedings of* AAAI 2005: *Educational Data Mining Workshop* (pp. 39–46). Technical Report WS-05-02. Menlo Park, CA: AAAI Press. http://www.aaai.org/Library/Workshops/ws05-02.php

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashlay, T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (ITS 2006), 26–30 June 2006, Jhongli, Taiwan (pp. 164–175). Berlin: Springer-Verlag.

Clark, R. E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling & User-Adapted Interaction*, 4, 253–278.

D'Mello, S., Blanchard, N., Baker, R., Ocumpaugh, J., & Brawner, K. (2014). I feel your pain: A selective review of affect sensitive instructional strategies. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design recommendations for adaptive intelligent tutoring systems: Adaptive instructional strategies* (Vol. 2). Orlando, FL: US Army Research Laboratory.

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., & Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (ITS 2010), 14–18 June 2010, Pittsburgh, PA, USA (pp. 245–254). Springer.

González-Brenes, J. P., & Mostow, J. (2012). Dynamic cognitive tracing: Towards unified discovery of student and cognitive models. In K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (EDM2012), 19–21 June, 2012, Chania, Greece (pp. 49–56). International Educational Data Mining Society.

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 C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798.

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* (CogSci 2010), 11–14 August 2010, Portland, OR, USA (pp. 471–476). Austin, TX: Cognitive Science Society.

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated cognitive model improvement. In K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (EDM2012), 19–21 June, 2012, Chania, Greece (pp. 17–24). International Educational Data Mining Society.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences*, 13(2), 129–164.

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better cognitive models to improve student learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 421–430). Springer.

Lan, A. S., Studer, C., Waters, A. E., & Baraniuk, R. G. (2013). Tag-aware ordinal sparse factor analysis for learning and content analytics. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July, Memphis, TN, USA (pp. 90–97). International Educational Data Mining Society/Springer.

Lan, A. S., Studer, C., Waters, A. E., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15, 1959–2008.

Li, N., Cohen, W., Koedinger, K. R., & Matsuda, N. (2011). A machine learning approach for automatic student model discovery. In M. Pechenizkiy et al. (Eds.), *Proceedings of the 4th International Conference on Education Data Mining* (EDM2011), 6–8 July 11, Eindhoven, Netherlands (pp. 31–40). International Educational Data Mining Society.

Li, N., Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2015). Integrating representation learning and skill learning in a human-like intelligent agent. *Artificial Intelligence*, 219, 67–91.

Lindsey, R. V., Khajah, M., & Mozer, M. C. (2014). Automatic discovery of cognitive skills to improve the prediction of student learning. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberge (Eds.), *Advances in Neural Information Processing Systems*, 27, 1386–1394. La Jolla, CA: Curran Associates Inc.

Liu, R., & Koedinger, K. R. (submitted). Closing the loop: Automated data-driven skill model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*.

Liu, R., & Koedinger, K. R. (2015). Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Education Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 420–423). International Educational Data Mining Society.

Liu, R., Koedinger, K. R., & McLaughlin, E. A. (2014). Interpreting model discovery and testing generalization to a new dataset. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK (pp. 107–113). International Educational Data Mining Society.

MacLellan, C. J., Harpstead, E., Patel, R., & Koedinger, K. R. (2016). The apprentice learner architecture: Closing the loop between learning theory and educational data. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 151–158). International Educational Data Mining Society.

Nathan, M. J., Koedinger, K. R., & Alibali, M. W. (2001). Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In L. Chen et al. (Eds.), *Proceedings of the 3rd International Conference on Cognitive Science* (pp. 644–648). Beijing, China: USTC Press. http://pact.cs.cmu.edu/pubs/2001_NathanEtAl_ICCS_EBS.pdf

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Pardos, Z. A., Trivedi, S., Heffernan, N. T., & Sárközy, G. N. (2012). Clustered knowledge tracing. In S. A. Cerri, W. J. Clancey, G. Papadourakis, K.-K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (ITS 2012), 14–18 June 2012, Chania, Greece (pp. 405–410). Springer.

Rosé, C. P., & Tovares, A. (in press). What sociolinguistics and machine learning have to say to one another about interaction analysis. In L. Resnick, C. Asterhan, & S. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association.

Rosé, C. P, & VanLehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *International Journal of Artificial Intelligence in Education*, 15, 325–355.

San Pedro, M., Baker, R. S., Bowers, A. J., & Heffernan, N. T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July, Memphis, TN, USA (pp. 177–184). International Educational Data Mining Society/Springer.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. doi:10.1214/10-STS330

Stamper, J., & Koedinger, K. R. (2011). Human-machine student model discovery and improvement using data. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED '11), 28 June–2 July, Auckland, New Zealand (pp. 353–360). Springer.

Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2011). Clustering students to generate an ensemble to improve standard test score predictions. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED '11), 28 June–2 July, Auckland, New Zealand (pp. 377–384). Springer.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16, 227–265.

Winne, P., & Baker, R. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, 5, 1–8.