British Journal of Educational Technology Vol 50 No 6 2019 2943–2958
doi:10.1111/bjet.12858

Explanatory learner models: Why machine learning (alone) is not the answer

Carolyn P. Rosé, Elizabeth A. McLaughlin, Ran Liu and Kenneth R. Koedinger

Carolyn P. Rosé is a professor of Language Technologies and Human–Computer Interaction in the School of Computer Science at Carnegie Mellon University. Her research focuses on technology support for learning through discussion and automated analysis of conversational process data. Elizabeth A. McLaughlin is a Scientific Technological Specialist at Carnegie Mellon University who is interested in the study of learning and learning analytics. Ran Liu, Chief Data Scientist at MARi, has worked on research in learning science and educational data science for over a decade, resulting in dozens of peer-reviewed publications and book chapters. She has extensive experience designing, implementing, and deploying education technology in classrooms and working with a variety of education stakeholders, particularly school district leaders and teachers. Kenneth R. Koedinger is the Hillman Professor of Computer Science at Carnegie Mellon. He investigates how people learn and provides learning engineering techniques and technology tools through his leadership of the LearnLab.org and LearnSphere.org efforts. Address for correspondence: Kenneth R. Koedinger, Human-Computer Interaction Institute and Psychology Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15201, USA. Email: koedinger@andrew.cmu.edu

Abstract

Using data to understand learning and improve education has great promise. However, the promise will not be achieved simply by AI and Machine Learning researchers developing innovative models that more accurately predict labeled data. As AI advances, modeling techniques and the models they produce are getting increasingly complex, often involving tens of thousands of parameters or more. Though strides towards interpretation of complex models are being made in core machine learning communities, it remains true in these cases of "black box" modeling that research teams may have little possibility to peer inside to try understand how, why, or even whether such models will work when applied beyond the data on which they were built. Rather than relying on AI expertise alone, we suggest that *learning engineering teams* bring interdisciplinary expertise to bear to develop *explanatory learner models* that provide interpretable and actionable insights in addition to accurate prediction. We describe examples that illustrate use of different kinds of data (eg, click stream and discourse data) in different course content (eg, math and writing) and toward different goals (eg, improving student models and generating actionable feedback). We recommend learning engineering teams, shared infrastructure and funder incentives toward better explanatory learner model development that advances learning science, produces better pedagogical practices and demonstrably improves student learning.

Introduction

This paper argues for a view of learning analytics we call *explanatory learner models*, the goal of which is to enable insight-driven use of such analytics in technology-enhanced education. Explanatory learner models do not just provide accurate prediction, but also offer actionable insights that may better advance both learning science and educational practice.

Practitioner Notes

What is already known about this topic

- Researchers in learning analytics and educational data mining have been successful in creating innovative models of data that optimize prediction.
- Some of these models produce scientific or practical insights and fewer have been put into use and demonstrated to enhance student learning.

What this paper adds

- We provide examples of development of explanatory models of learners that not only accurately predict data but also provide scientific insights and yield practical outcomes.
- In particular, researchers with expertise in cognitive science and math education content use AI-based data analytics to discover previously unrecognized barriers to geometry student learning. They use model-derived insights to redesign an online tutoring system and “close-the-loop” by experimentally demonstrating that the new system produces better student learning than the original.

Implications for practice and/or policy

- We define explanatory learning models and provide an articulation of a process for generating them that involves interdisciplinary teams employing human–computer interaction and learning engineering methods.
- Based on our experiences, we recommend learning engineering teams, shared infrastructure and funder incentives toward better explanatory learner model development that advances learning science, produces better pedagogical practices and demonstrably improves student learning.

Just as most science is driven by accepted forms of evaluation, such as proofs in mathematics or controlled experiments in psychology, learning analytics is no exception. In it, a primary form of evaluation in practice is the prediction accuracy of a proposed analytic model, usually measured by fit to data in a held-out test set not used to derive model parameters. Often it has been the case that prediction fit has served as a major driving force in educational data mining and data mining more generally, with researchers making and claiming advances by developing analytic models that yield better predictions than previous models. While accurate prediction facilitates scientific rigor, it does not, by itself, directly contribute to scientific understanding nor to actionable insights for practical applications. We argue for a greater emphasis on evaluating the explanatory power of a model in addition to its prediction accuracy. We define an *explanatory learner model* as a model for which insight about learners, the learning process, or the instructional context can be derived from interpretation of the structure or parameter estimates of learned models.

We begin by motivating some problems within the field of Learning Analytics under the “Politics, Pedagogy & Practices” theme of this special issue. We use a case study to motivate how pedagogical design practices need to be shifted with a goal of building explanatory learner models. Rather than AI researchers working to optimize prediction, a design process is needed that involves a broad set of interdisciplinary stakeholders in collaboratively interpreting AI-based models. These stakeholders may not only include AI researchers, but also cognitive, social or educational psychologists, psychometricians, discipline-based education researchers, educational equity researchers, etc. Interdisciplinary research team members bring distinct scholarly experience and theories to

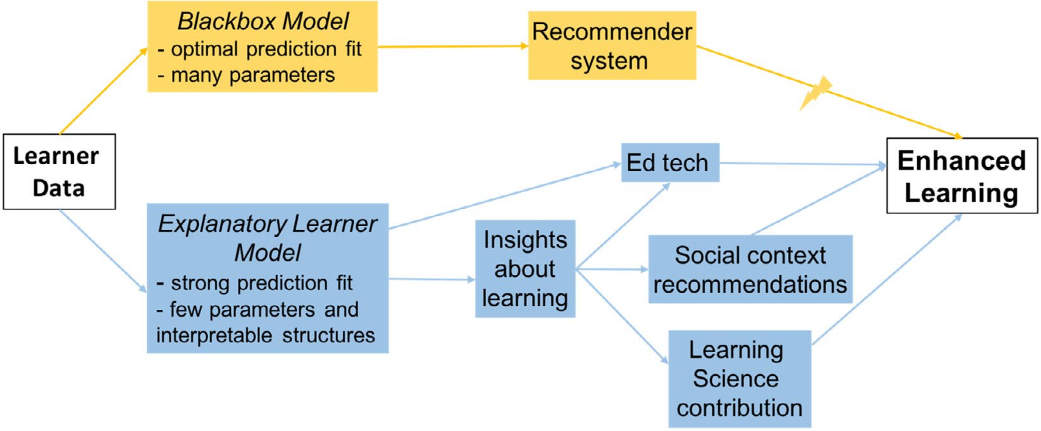


Figure 1: For AI researchers, a simple desirable path (top half) from learner data (leftmost box) to enhanced learning (rightmost box) involves machine learning of an optimal Blackbox Model that may be used inside applications such as recommender systems that improve student learning. This path is, unfortunately, most likely to fail (lightning bolt) because the recommender system will not be transparent enough to adapt in an insightful way to the individual needs of students nor to the context of use. We suggest an alternative path (bottom half) that uses learner data to create an Explanatory Learner Model and employs Human–Computer Interaction (HCI) methods to derive insights about student learning that drive educational technology development and recommendations for how the technology is best used in context. Together these produce enhanced learning [Colour figure can be viewed at wileyonlinelibrary.com]

bear in designing and interpreting models and in generating insights from models. An equally important set of stakeholders are the students and educational practitioners who serve them, including teachers, parents, school administrators, educational content/product developers, etc. These stakeholders aid the design and interpretation of models and are especially important in the process of turning insights into design ideas.

The examples we describe below provide a contrast with a simple, idealized approach to AI-based learning analytics. They illustrate problems that must be addressed and motivate making a distinction between two paths in engineering learning solutions, illustrated in Figure 1. One approach (top path in Figure 1) involves fitting learner data using machine learning methods with an emphasis on accuracy rather than explanatory power to create an optimal prediction from a complex model with many parameters. In this case, understanding the model is treated as non-critical as long as the model produces good prediction results. As an example, from such a blackbox prediction model, a recommender system may be built and used with students with the goal of enhancing their learning. Though the impact on learning may be positive in some cases, we argue this pathway is flawed (indicated by the lightning bolt in Figure 1). For recommender systems to be reliably and optimally successful in enhancing learning, developers need empirical methods for design and implementation of both the system itself and strategies for its successful deployment in context (eg, teaching, training). We recommend an alternative Explanatory Learner Model path (bottom path in Figure 1) that incorporates these methods and more, including applications other than recommender systems. This path is implemented by a *learning engineering* team (cf., Hess & Saxberg, 2014; Simon, 1967).

We use “learning engineering” to refer not only to the application of learning science in educational design but also to the use of data, both quantitative and qualitative, in an iterative process of design and improvement. Simon (1967) originated the term to suggest that college faculty should not only know their discipline, but also have "a professional knowledge of the laws of learning" (p. 73). Just as bridges are not designed by intuition but are engineered, Simon argued that learning environments should not be designed by intuition, but engineered. Since that paper, much progress has been made in the emergence of the learning sciences, in increasing adoptions of educational technologies and assessments that can provide detailed learning process and outcome data, and in the creation of associated new fields of learning analytics and educational data mining. In this context, we and others (eg, Hess & Saxberg, 2014) see value in reintroducing the notion of "learning engineering." We are motivated not only by increasing data available for learning analytics, but also by the fact that attempts to straightforwardly apply learning science often fail (Coalition for Evidence-Based Policy, 2013). For this reason, we include in the notion of learning engineering an emphasis on a process of data-driven iterative improvement.

In our view, developing explanatory learner models are a fundamental aspect of effective learning engineering. Rather than first emphasizing accuracy in model building and then attempting interpretation at a later phase as in much state-of-the-art work in neural model interpretation (eg, Fiacco & Rosé, 2019), from the beginning of the modeling process, a learning engineering team places as much attention on potential insights the model produces, through the inferred structures or key parameter estimates, as they place on its prediction fit. The team then uses the model and the insights to design a novel educational technology application. In the process, the team may employ various research methods from Human–Computer Interaction and learning science to support model interpretation, insight formation and application design. Note that model interpretation approaches used by AI researchers may still be of value at this interpretation stage (Fiacco, Cotos, & Rosé, 2019). These methods and model-based insights help the team not only develop the application but also develop recommended strategies for how stakeholders (eg, students, teachers, parents, administrators, etc.) should best implement the application in the desired social context of use. Importantly, this explanatory model path has the added benefit of potentially producing new contributions to learning science that may generalize the derived insights and social context recommendations.

Drilling down into the bottom path, we see explanatory learner models (cf., Liu & Koedinger, 2017b) as a key part of an iterative data-driven learning engineering approach (Hess & Saxberg, 2014; Simon, 1967). This approach is not just about making good use of quantitative data, but is also about making good use of qualitative data and theory. Collaborative, human-centered research practices, including educational design research (Barab & Squire, 2004; McKenney & Reeves, 2018), cognitive task analysis methods (Lovett, 1998) and Human–Computer Interaction (HCI) methods (Dabbs *et al.*, 2009) are a core part of model development and application.

In the body of this paper, we motivate the need for explanatory learning models through an example of the top path displayed in Figure 1 and then present some detailed cases of development and use of explanatory learner models that draw on the authors' experiences along the bottom path. These are, by no means, meant to be representative of the whole set of such work. There have of course been examples of opaque modeling approaches exemplified by the top path evaluated in A/B tests that have indeed resulted in positive effects, such as a series of studies in which shallow prediction models were used to detect opportunities for conversational agents to intervene in collaborative encounters (eg, Adamson, Dyke, Jang, & Rosé, 2014; Dyke, Adamson, Howley, & Rosé, 2013). In these cases, the intervention design was motivated by HCI practice, but the models used themselves were simply prediction models that do not really rise to the level of explanatory models. However, what is more telling is the number of examples of projects that stop at demonstrating prediction accuracy over competing models. In these cases, there is no follow-up "close-the-loop" effort to demonstrate consequences for enhanced student learning (eg, in an A/B experiment) or even, in some cases, little to no indication of scientific insights derived from the model or of potential applications of the model. Moreover, there are many exceptions besides our own (eg, Arnold, 2010; Arnold & Pistilli, 2012; Paquette, de Carvalho, & Baker, 2014; Paquette, Baker, de Carvalho, & Ocumpaugh, 2015), which we summarize in the recommendations section. The purpose of this article is to offer a reflection on the contrast between these approaches in order to raise a challenge for the field going forward to embrace the advantages of the bottom path.

A motivating case study: Machine learning needs HCI and learning engineering to make analytics useful

If AI-based analytics are to have an impact on education, developers must attend to the human interaction and social context of using the technology. This attention to Human–Computer Interaction issues surrounding technology use and adoption have long been emphasized in educational technology research (eg, Corbett, Koedinger, & Hadley, 2001; Koedinger, Anderson, Hadley, & Mark, 1997). As new players, like AI researchers, come into the educational technology space, it is important for them to partner with others with expertise in addressing these context issues that touch on the general issues of politics, pedagogy and practices. These issues start with, but do not stop with, how effectively individual users (students, teachers, parents, administrators) interact with the technology. In addition to these usability and usefulness issues, there are issues of desirability that determine whether and how a technology is adopted, accepted, trusted and engaged with over multiple years. Techniques for user-experience design are relevant (eg, Beyer & Holtzblatt, 1998; Karapanos, Zimmerman, Forlizzi, & Martens, 2009).

While issues of usability, usefulness and desirability are relevant in technology applications of all kinds, issues related to long-term impacts on student learning and engagement are of particular importance in the educational technology realm. These are issues for which explanatory learner models can have real benefit, as illustrated below (see "Explanatory knowledge component model…" section), in that insights derived from models can help insure that learning outcomes are achieved. As noted above, it is certainly possible for AI researchers to develop applications without taking the explanatory model path. They can use machine learning and prediction accuracy to create a “blackbox” model (a model where parameter interpretation is not sought) and use that model in a recommender system. However, to be sure, that system is unlikely to be effortlessly useful, usable, desirable and produce long-term learning outcomes. Further, a team may find, as application efforts are pursued, that retrospective attempts at deriving explanatory characteristics of the model are valuable.

Case study: Challenges in Achieving positive impact of AI-based recommendations when uptake is a choice

Advice-giving is an area where the “technology is not enough” message speaks loudly and clearly. The best advice, if not heeded, will produce nothing of value. Here, we reflect on lessons learned from two past efforts to build recommender systems to support learning at scale, in each case optimizing for advice that was high quality in accordance with some objective measure, but each of which was thwarted in part because the advice did not take into consideration some important aspect of human choice.

The first example was an intervention to support help exchange deployed in a MOOC study (Rosé & Ferschke, 2016). Earlier research on MOOC discussion forums had indicated that the experience of confusion, as well as exposure to other students’ confusion, are both associated with elevated attrition in MOOCs (Yang, Wen, Howley, Kraut, & Rosé, 2015). Attempts at resolving confusion by making help requests in the threaded discussions were frequently left without a satisfactory response (Yang, Piergallinin, Howley, & Rosé, 2014). In response to these two problems, a discussion-focused intervention, called the Quick Helper, was integrated with a Coursera MOOC to support help-seeking as well as increase the probability that help requests will be met with a satisfactory response. This help-seeking intervention connected students with student peers restimated to be able to answer their question. The QuickHelper was continuously available to students via a button. When they clicked it, they were guided to formulate a help request, which was posted to the MOOC discussion board, and the text and metadata were forwarded to the Quick Helper system (Yang *et al.*, 2014). A corpus-based evaluation demonstrated that the optimization was successful in balancing a number of concerns related to user capabilities and load balancing. A study of the interface through which learners selected potential help providers also supported the design with respect to its intended purpose regarding facilitation of help seeking (Howley & Rosé, 2018).

Though overall the intervention was successful, as a follow up, we investigated the reasons why students sometimes did not respond to invitations to join discussion threads. We used data from participation in the course in earlier weeks to make predictions about which students showed evidence of being appropriate to recommend as help providers. But sometimes Quick Helper recommended students who had demonstrated this ability in earlier participation, but had dropped out between that demonstration and the subsequent recommendation. This occurred easily since students dropped out of the course without any official declaration. Students who dropped out of the course already were, not surprisingly, reluctant to spend time offering help to students still in the course. Thus, the process broke down if those students were invited to participate. The take home message here is that it would have been valuable in hindsight to have been more aware of the "invisible" assumptions behind the model that might thwart its success in practice despite its apparent good performance in a corpus-based evaluation, hence an argument in favor of more explanatory models.

A related study addressed a separate question that is important for managing group learning at scale, namely assignment of students to learning groups. The history of this work has involved a series of lab studies (Wen, Maki, Dow, Herbsleb, & Rosé, 2018), a MOOC deployment (Wen *et al.*, 2018) and a series of deployments in a large online course offered by a major university in the mid-Atlantic region of the US. The lab studies and MOOC study offer substantial evidence that the approach to team assignment resulted in teams where students worked together better, both in terms of process measures and group productivity measures, with effect sizes around three standard deviations on both measures in one lab study. Based on this success, two large classroom studies were conducted to evaluate how the paradigm would play out if the recommendations were generated in the same way as before, but students had the opportunity to decide whether to take the recommendation or not. The paradigm for team formation we investigate here works as follows. In the course, students complete 10 individual weekly projects and a 7-week team project. At the end of each individual project, each student authors a short reflection piece which other students then provide feedback on, serving as a collaborative exchange. We automatically analyze this feedback for transactivity. Transactivity is known to be linked to effective learning in collaborative groups as it captures learners' building off the contributions of their learning partners (Berkowitz & Gibbs, 1983; Teasley, 1997; Wen, Maki, Wang, & Rosé, 2016). Based on the transactivity observed in the exchanges that already occurred between students on their reflection pieces, a constraint satisfaction algorithm (Wen, 2016) maximizes the predicted transactivity across all teams in the course in order to suggest teams.

We evaluated whether students about to participate in a team project in this online course would adopt teams suggested by the course staff instead of forming their own. In an initial study, students almost universally chose not to take the automated recommendations. In many cases, students chose to work with friends, though some used more objective observations of teamwork, work ethic and technical skills in their choice of team mates. A post hoc analysis at the end of the course revealed that students who chose to work with friends had significantly more reported problems with communication and division of labor. Nevertheless, in a second semester where we shared these results with the students at the time when the recommendations were given, they still largely chose not to take the recommendations. Again, what thwarted the success in practice was an assumption (in this case, that students were *required* to take the recommendation) that did not generalize from the earlier context of evaluation to a new context for practice.

To summarize, research in machine learning has yielded a plethora of techniques for optimization of objective functions, and in both of these cases, optimization led to insights into advice that was high quality, in one case because the selected helpers were identified as having needed expertise, and in the other case because the automated grouping showed success in user studies where the students were required to take the recommendation. In both of these efforts, the results are a mix of victory and defeat. They suggest a clear need for system development and implementation to make the assumptions behind the model more explicit and transparent so they can at least be considered in advance of application of the approach in practice in another context.

Explanatory models provide a path to scientific insight and better application

We next describe two extended examples that illustrate explanatory modeling efforts: (1) knowledge component model discovery and (2) instructor and student use of model results.

Explanatory knowledge component model discovery produces insight and better applications

The most firmly grounded and rigorous evaluation of a learning analytics result is whether it yields better student learning when applied. We have referred to such evaluation as "closing the loop" (Koedinger, Stamper, McLaughlin, & Nixon, 2013), as it completes a cycle of system design, deployment, data analysis and discovery leading back to redesign. Past successes in closing the loop have benefited from the interpretability or explanatory character of the initial model discovery (Koedinger *et al.*, 2013), and the example we present here is no exception. It uses an educational data mining approach, called Learning Factors Analysis (Cen, Koedinger, & Junker, 2006), that provides an automated method of improving cognitive models that underlie adaptive learning technologies (cf., Koedinger & Corbett, 2006).

An abstracted form of a cognitive model convenient for statistical analysis is a so-called knowledge component (KC) model, (Koedinger *et al.*, 2010). Unlike a cognitive model, a KC model does not attempt to fully represent the cognitive processes that combine to produce the richness of human task performance. Instead, a KC model specifies, with labels (eg, "add like-denominator fractions"), a set of hypothesized component processes needed for a set of tasks in a domain (eg, various fraction arithmetic problems). These component processes could be facts, skills, or principles, collectively called "knowledge components" or KCs. For any set of tasks, there are multiple possible hypotheses for what the KCs are and these hypotheses imply how tasks cluster in terms of student performance across tasks (through difficulty predictions) and over time (through learning transfer predictions). For example, does learning to add fractions like 1/4 + 1/6 involve the same KC or different KCs from learning to add fractions like 1/4 + 1/8? We can determine from data which of these KC hypotheses is correct, the merged KC (first one) or the split KCs (second one). The data favor the split KC hypothesis because the second set of tasks is much easier (ie, one of the two denominators, 8, is the common denominator) than the first (ie, a new denominator, 12, needs to be determined).

Learning Factors Analysis (LFA) is an automated method to use data to discover which of many possible KC models perform best (Cen *et al.*, 2006). LFA performs a combinatorial search process across hypothesized KC models. These KC models are automatically derived from existing, human-labeled KC models using simple operations that split, merge or add KCs to generate hypothesized KC models. Each generated KC model is evaluated for its predictive fit to data on student task performance over time. Because LFA starts with human-labeled KCs and uses simple operations to produce new KCs, it facilitates reasonably straightforward interpretation and explanation of the best-fitting KC model(s) that are generated. Koedinger and colleagues applied LFA on 11 datasets spanning different domains and different educational technologies, all publicly available from DataShop (http://learnlab.org/datashop). For all 11 datasets, the new KC models that LFA discovered improved cross-validated prediction fit in comparison to the previous best existing human-generated KC models (Koedinger, McLaughlin, & Stamper, 2012).

Following an interdisciplinary learning engineering approach, the research team brought to bear expertise in cognitive science and math education to interpret the LFA-discovered model from one of these datasets (DataShop Dataset #76), produced from an online geometry tutor. Their investigation into why the discovered model was better than the best human-generated model revealed one critical difference. The LFA-discovered model split circle area items into two separate KCs, one involving forward calculations (ie, find the area, given the radius) and the other involving backward calculations (ie, find the radius, given the area). In contrast, the original human-generated model labeled both types of items as a single KC (ie, circle-area). The LFA-discovered model did not find benefits for this type of KC split for other shapes in the dataset (rectangles, triangles and parallelograms). The data indicate that while for most shapes, backward application of geometry area formulas is generally not harder to do or learn than forward application, it is in the case of circle area. Applying domain expertise, the researchers hypothesized that the automated model improvement had identified a hidden skill: knowing when and how to apply a square root operation for backward circle area items. This hidden skill is not needed in forward circle area items, or for backward application for other area formulas.

The researchers validated the interpretation of the model discovery on a novel dataset whose structure (ie, problems) and properties (ie, school district, students) differed from those of the original dataset on which the LFA discovery was made (Liu, Koedinger, & McLaughlin, 2014). The researchers *did not* apply the improved model directly to new data (eg, as in Feng, Heffernan, & Koedinger, 2009) or run an exact replication of the study. Rather, they validated that their *interpretation* of the prior analysis resulted in a new KC model that fit a novel dataset containing problems and skills (eg, square area backward problems) that did not even exist in the original dataset on which the model discovery was based. This kind of generalization test would not have been possible if the discovery were not interpreted and explained.

Liu and Koedinger (2017a) then *closed the loop* on the LFA model discovery by first creating an intelligent tutoring system covering the shape-area unit and resembling the geometry tutoring system that had produced the original “model discovery” dataset. This version was the control condition tutoring system. Based on the interpreted KC model improvements discovered by LFA, the researchers created a redesigned tutoring system that utilized the improved KC model, leading to different amounts of relative practice on forward versus backward items compared to the control tutoring system. The derived insight (that learning when to employ square root is hard) also led the team to make principled modifications to the interface and hints given on backward circle and square area items to scaffold application of the square root. They found significantly higher learning gains in the redesigned tutor condition. Thus, the insight from the LFA model discovery led to a pedagogical modification that, in turn, led to increased learning outcomes. This case study underscores the importance of considering the interpretability and actionability of educational data science results. The interpretation of the LFA discovery was critical to a number of the redesign changes that led to improved learning outcomes.

That LFA initially requires human input has been cited as a limitation (eg, González-Brenes & Mostow, 2012) in arguments favoring purely automated methods of discovering KC models. We argue, however, that it is precisely this "human-in-the-loop" feature that leads the results of such modeling efforts to be interpretable. Some researchers (eg, González-Brenes & Mostow, 2012; Lindsey, Khajah, & Mozer, 2014) have attempted to fully automate the process of discovering KC models. These methods have much to recommend them, reducing demands on human time and producing competitive results in predictive accuracy. However, the resulting KC models of these efforts have not been interpreted or acted upon with respect to improving instruction. These models are arguably blackbox models because of the huge number of parameters and remain non-explanatory until they are interpreted.

Other modeling efforts that have included a "human-in-the-loop" component, like Ordinal SPARFA-Tag (Lan, Studer, Waters, & Baraniuk, 2013), which incorporates domain expert concept tagging in the model development process up front, have yielded considerably more interpretable cognitive models than alternative methods. Although any final interpretations of modeling efforts are necessarily made by humans, methods like LFA and Ordinal SPARFA-Tag greatly improve the likelihood of generating sensible resulting models by incorporating the human effort up front (Lan, Waters, Studer, & Baraniuk, 2014).

Teacher & student use of explanatory learner models

Many data mining techniques have evolved for developing sophisticated learner models and pedagogical support, yet a gap continues to exist between modeling, usage and practical application. This need for manageable and understandable learner models in teaching systems that can be utilized by both teachers and students has been known for years. Early research included the design and development of "scrutable models" with a focus on giving students feedback on model inferences (Holden & Kay, 1999; Kay, 2000) and that line of research is updated in a companion piece in this special issue (Kay & Kummerfeld, 2019). One early fielded example of such is the "Skillometer" in Cognitive Tutors (Koedinger *et al.*, 1997), which used a dynamically-updating bar chart to show the model's estimates of the probability that students have mastered each target skill or Knowledge Component. Cognitive Tutor development and use has always emphasized teacher involvement (eg, Corbett *et al.*, 2001) and even early versions of the high school mathematics cognitive tutors had a "teacher toolkit" to provide data summaries to teachers about their students (eg, Ritter, Blessing, & Wheeler, 2003).The evidence for these early limited explanatory versions of data models is limited and mixed, especially with respect to student use of such models. We next provide some examples of promise toward enhanced support, particularly for instructors and course developers, but certainly these are only a small step toward wide-scale successful use of explanatory models by teachers (much less students).

Building on these early efforts, recent design research has begun to better clarify teacher needs and design data dashboards that are more responsive to them. Holstein, McLaren, and Aleven (2017) interviewed 10 middle school teachers and found they want to know *why* a student is struggling (eg, careless mistakes vs. misconceptions), not simply that they are having difficulty. They also want support and resources for how to best help their students with content and affective states. In a follow-up study with 286 middle school students, across 18 classrooms and 8 teachers, Holstein, McLaren, and Aleven (2018) found higher student learning gains in classes with teachers receiving real-time data analytics, through a headset-displayed dashboard, than in classes where the same teachers did not have access to the data dashboard.

An instructor dashboard is a key part of online college-level courses from the Open Learning Initiative (OLI). An evaluation of the OLI Statistics course compared blended use of these highly interactive course materials, in a half-semester, with a full semester existing version of a face-to-face course. Students learned more from the blended course (eg, 18% vs. 3% gain on a standardized statistics test) despite spending only half the time (50 hours vs. 100 hours; Lovett, Meyer, & Thille, 2010). One of the key success factors, highlighted by the researchers and cited by the instructor, was the ability to use the dashboard before class to better understand what topics were worth focussing on in class discussion.

Another method for improving instruction and learning is to provide instructors and developers with easy-to-use tools for analyzing student data. We provide an example that demonstrates how DataShop tools (eg, learnings curves, performance profiler, error reports) were used by a course designer to quickly assess problems in a college course on causal modeling (inspired by early learning analytics by Corbett & Anderson, 1995). This example involves an exploration of interaction data from undergraduates using the online OLI course on Causal Discovery (DataShop dataset #2500).

The instructor and a learning engineer scanned a page of DataShop visualizations of the learning curves of each hypothesized knowledge component (a topic, concept or skill). Some curves did not show a smooth decline in error rate suggesting an opportunity for improvement. For example, the learning curve shown in Figure 2, for a knowledge component about recognizing "colliders" on a path of variables in a causal graph, shows an upward error rate spike at opportunity four. The instructor knew that an error rate blip suggests a likely issue with the student tasks associated with that opportunity—in other words, there may a *hidden knowledge component* (an unknown student challenge) that the current labeling of this topic was not indicating.

By clicking on this point in the learning curve, the instructor is able to access the problem producing this error blip, shown in Figure 3, and inspects a DataShop error rate report, shown in Figure 4. It shows that the step in this problem about the variable X1 is much easier for students,

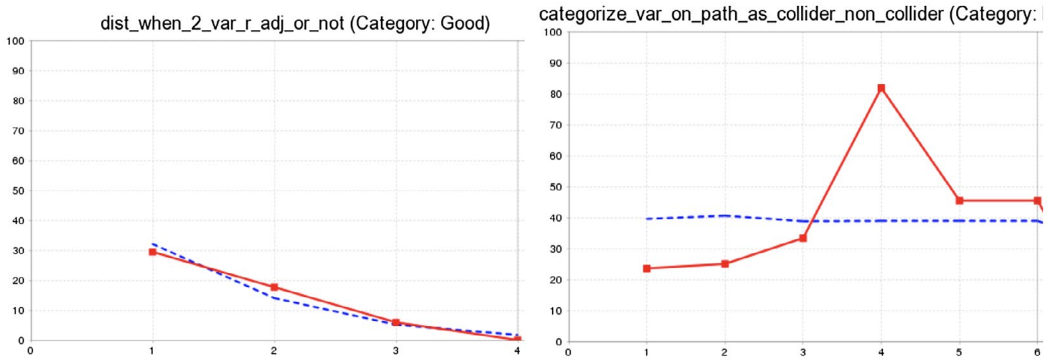
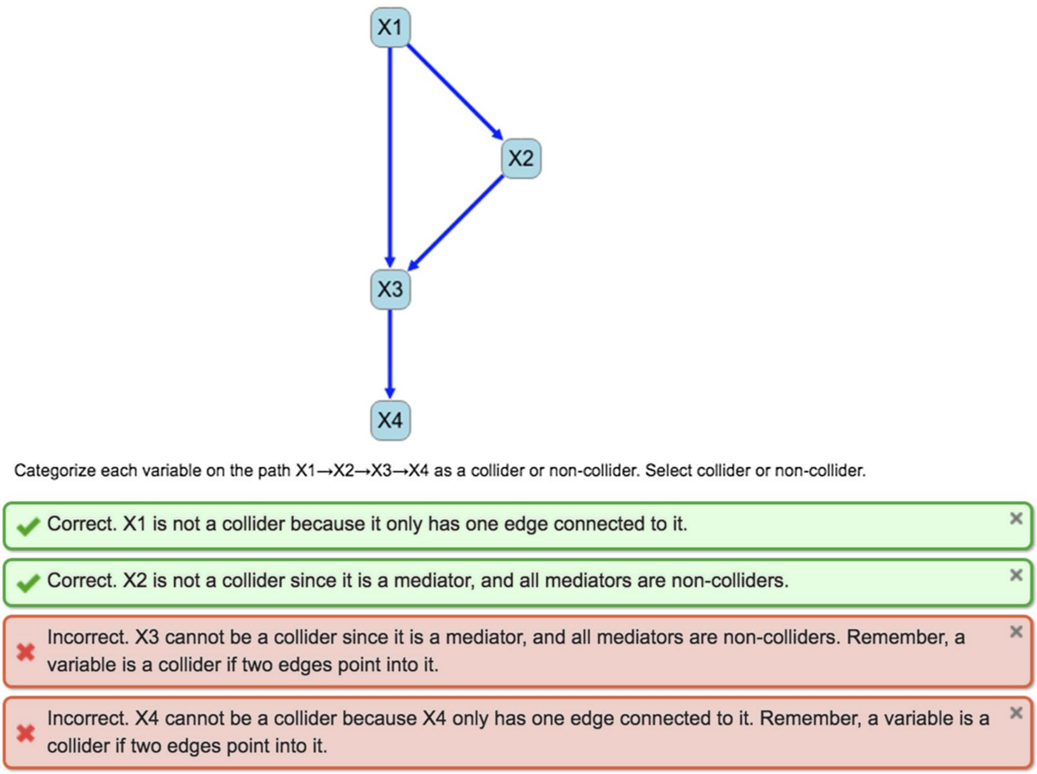


Figure 2: Datashop learning curves for two Knowledge Components. The x-axis shows the number of opportunities and the y-axis is the error rate. The solid line represents real student data and the dashed line is the prediction. On the left is an example of a good learning curve (smooth and declining) for the knowledge component "distance when 2 variables are adjacent or not", with a good model fit to data. The right shows a flat learning curve of the knowledge component "categorize variable on path as collider or non-collider", with poor model fit to data. The spike at opportunity four suggests an issue with the steps associated with that data point [Colour figure can be viewed at wileyonlinelibrary.com]



Categorize each variable on the path X1→X2→X3→X4 as a collider or non-collider. Select collider or non-collider.

✓ Correct. X1 is not a collider because it only has one edge connected to it. ×

✓ Correct. X2 is not a collider since it is a mediator, and all mediators are non-colliders. ×

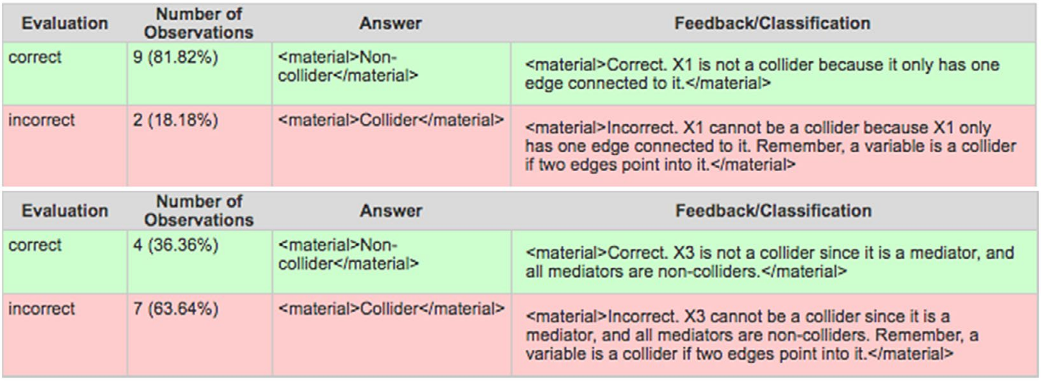
✗ Incorrect. X3 cannot be a collider since it is a mediator, and all mediators are non-colliders. Remember, a variable is a collider if two edges point into it. ×

✗ Incorrect. X4 cannot be a collider because X4 only has one edge connected to it. Remember, a variable is a collider if two edges point into it. ×

Figure 3: (a) The top of the figure shows an example problem in the OLI platform for a KC about identifying a variable as a collider or non-collider on a causal path and (b) the bottom of the figure shows the immediate feedback given after each response. Note that X1 and X4 give different feedback than X2 and X3 (one edge vs. mediator, respectively) indicating different knowledge is required [Colour figure can be viewed at wileyonlinelibrary.com]

at 81.8% correct (top "correct" row in Figure 4), than the step about the variable X3, at only 36.3% correct (second "correct" row in Figure 4).

The instructor tries to figure out why identifying whether X3 is a collider on the path X1 → X2 → X3 → X4 is harder than identifying X1 as such. One reason is that X1 has only one edge connected to it (as the error feedback accurately indicates), but a more nuanced reason is that while X3 is not a collider on this path (where the X1-X3 edge is not included), X3 *is* a collider on the path X1 → X3 ← X2. In this case, the error feedback message does note this important nuance. As a consequence, the instructor and learning engineer decide to relabel these steps as different knowledge components and, more importantly, to make two associated improvements to the course. First, after this relabeling, there are not enough practice opportunities for both KCs, especially for the harder one. Thus, the redesign will involve creation of more problems (eg, another one with the same image, but asking to identify colliders on the X1 → X3 ← X2 path). Second, they will change the feedback messages in cases where a variable is a collider on some path, but not on the specified path. The next use of this revised course in the coming semester will provide data to see whether these changes enhance student learning as indicated by smooth declining error rate learning curves for the new KCs identified.



Evaluation	Number of Observations	Answer	Feedback/Classification
correct	9 (81.82%)	<material>Non-collider</material>	<material>Correct. X1 is not a collider because it only has one edge connected to it.</material>
incorrect	2 (18.18%)	<material>Collider</material>	<material>Incorrect. X1 cannot be a collider because X1 only has one edge connected to it. Remember, a variable is a collider if two edges point into it.</material>

Evaluation	Number of Observations	Answer	Feedback/Classification
correct	4 (36.36%)	<material>Non-collider</material>	<material>Correct. X3 is not a collider since it is a mediator, and all mediators are non-colliders.</material>
incorrect	7 (63.64%)	<material>Collider</material>	<material>Incorrect. X3 cannot be a collider since it is a mediator, and all mediators are non-colliders. Remember, a variable is a collider if two edges point into it.</material>

Figure 4: An instructor inspecting this DataShop Error Report focuses on two steps labeled as the same topic (the KC in Figure 2) and finds they have drastically different error rates (18.18% for first step and 63.64% for the second). Thus, the instructor inspects this problem (see Figure 3) to try to explain why the second step (involving variable X3) is harder than first (involving variable X1)
[Colour figure can be viewed at wileyonlinelibrary.com]

Recommendations

A running theme of this article is that machine learning results are not enough. A model may produce great predictions on previously collected data, but then be ineffective when employed in an application—not just because the model may not generalize well to the next context, but also because the design of either the associated system or surrounding plan for use are flawed.

Consider the case with the QuickHelper MOOC (see "Case study..." section), the model performed well in the original corpus-based evaluation and was published at a top tier recommender system conference. Nevertheless, the recommender system failed to work in an optimal way when it was subsequently used and evaluated in a MOOC. The in situ evaluation in this case revealed the importance of accounting for the status of recipients in terms of their continued activity in the course, which highlights the need for a development methodology that anticipates and avoids such problems. The team can now redesign the algorithm, by including this status variable, and then evaluate it with the new and old data and with a modified system application. Model building, system integration and in situ testing should be interleaved. Sources of model and application error should be sought outside of the limited notion of error that applies solely to the model decontextualized from authentic use cases.

There are existing examples, outside of our own, of the application of machine learning and AI—in the context of learning engineering—to improve learning processes. Paquette and colleagues (2014) combined machine learning and knowledge engineering techniques to build detectors of gaming the system behaviors. Their knowledge engineering efforts enhanced the explanatory character of their models as compared to pure machine learning approaches, leading to increased predictive power and enhanced generalization across data sets (Paquette *et al.*, 2015). Another example is the Course Signals project at Purdue, which used student interaction data from a virtual learning environment (eg, Blackboard, Canvas) and predictive modeling to provide an early alert to students as to their progress (Arnold, 2010), with some evidence for improving student outcomes (Arnold & Pistilli, 2012). Similarly, Huberth, Chen, Tritz, and McKay (2015) used analytics of student survey, performance and background data to develop E^{2}Coach, which tailors communication with introductory physics students and yields improved student learning outcomes.

In general, we recommend that academic, industry and funding leaders actively attempt to *foster collaborative learning engineering* where interdisciplinary teams work toward producing explanatory learner models, develop applications based on them and on HCI and Learning Engineering methods and pursue close-the-loop A/B learning experiments. The interdisciplinary teams we envision essentially come in four broad buckets: (1) *technical* disciplines, including AI, machine learning, language technologies, statistics, etc., (2) *social science* disciplines, including cognitive, educational and social psychology, linguistics, economics, etc., (3) *education* research disciplines including discipline-based education research (literacy ed, math ed, physics ed, etc.), curriculum and instruction, psychometrics, etc., and (4) *design* disciplines including interaction design, user experience design, ethnography, art, etc. There are also relevant existing and emergent combination disciplines like cognitive science, human-computer interaction, AI in education, learning sciences and learning engineering. But, a key point of collaborative learning engineering is not that individuals have expertise in all these areas, but that the team as a whole does when these different areas are bridged within teams.

Moreover, we suggest that learning engineering teams should put more emphasis on the interpretability and actionability of educational data mining and learning analytic efforts, to produce more explanatory models (as in the bottom path in Figure 1). Teams should try to understand *why* the model achieves better predictive accuracy than alternatives. An explanation for why can yield insight into how students succeed or struggle to learn the relevant material, as was illustrated in the examples from the Geometry course analysis and redesign (see "Explanatory knowledge component model…" section) and the OLI Causal Reasoning course (see "Teacher & Student use …" section). This insight should, in turn, be used in a redesign of course materials or practices, as indicated in both examples. Ideally, a follow-up “close-the-loop” A/B experiment is run, as in the Geometry example, to confirm that the redesign (B) does indeed produce better student learning outcomes than the original design (A). Finally, interpretable insights derived from data-driven discoveries can be understood by the broader learning sciences community and thus contribute to advancing the science of learning.

Acknowledgements

This work was supported by a National Science Foundation grant (ACI-1443068) toward developing LearnSphere.org.

Statements on open data, ethics and conflict of interest

The analyses and data described in this paper are available, openingly or by request, through LearnSphere.org or, in some cases as noted in the body more directly through LearnLab.org/DataShop or discoursedb.github.io.

Research was carried out after approval of the CMU IRB.

The authors have no conflicts of interest.

References

Adamson, D., Dyke, G., Jang, H. J., & Rosé, C. P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of AI in Education*, *24*(1), 91–121.

Arnold, K. E. (2010). Signals: Applying academic analytics. *Educause Quarterly*, *33*(1), 10. Retrieved from the Educause Review Online website: http://www.educause.edu/ero/article/signals-applying-academic-analytics

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics & Knowledge*. New York, NY: ACM.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, *13*(1), 1–14.

Berkowitz, M. W., & Gibbs, J. C. (1983). Measuring the developmental features of moral discussion. *Merrill-Palmer Quarterly*, *1982*, 399–410.

Beyer, H., & Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. San Francisco, CA: Elsevier.

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

Coalition for Evidence-Based Policy. (2013). *Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects*. Retrieved from http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf

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

Corbett, A. T., Koedinger, K. R., & Hadley, W. H. (2001). Cognitive tutors: From the research classroom to all classrooms. In P. S. Goodman (Ed.), *Technology-enhanced learning: Opportunities for change* (pp. 235–263). Mahwah, NJ: Erlbaum.

Dabbs, A. D., Myers, B. A., Mc Curry, K. R., Dunbar-Jacob, J., Hawkins, R. P., Begey, A., & Dew, M. A. (2009). User-centered design and interactive health technologies for patients. *Computers, Informatics, Nursing*, *27*, 175–183.

Dyke, G., Adamson, A., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and discussion with conversational agents. *IEEE Transactions on Learning Technologies*, *6*(3), special issue on Science Teaching, 240–247.

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

Fiacco, J., & Rosé, C. P. (2019). Deep neural model inspection and comparison via functional neuron pathways. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*.

Fiacco, J., Cotos, E., & Rosé, C. P. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of Learning Analytics and Knowledge (LAK'19), Tempe, AZ, USA* (pp. 310–319).

González-Brenes, J. P., & Mostow, J. (2012). Dynamic cognitive tracing: Towards unified discovery of student and cognitive models. In *Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece* (pp. 49–56).

Hess, F., & Saxberg, B. (2014). *Breakthrough leadership in the digital age: Using learning science to reboot schooling*. Thousand Oaks, CA: Corwin.

Holden, S., & Kay, J. (1999). The scrutable user model and beyond. In R. Morales (Ed.), *AIED, artificial intelligence and education, 99 Workshop W7: Open, interactive, and other overt approaches to learner modelling* (pp. 51–62). Retrieved from http://www.dai.ed.ac.uk/groups/aied/Conferences/Ovalm/Papers

Holstein, K., McLaren, B. M., & Aleven, V. (2017). Intelligent tutors as teachers' aides: Exploring teacher needs for real-time analytics in blended classrooms. *LAK*, *2017*, 257–266.

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, … B. du Boulay (Eds.), *Artificial intelligence in education. AIED 2018. Lecture Notes in Computer Science* (Vol. 10947, pp. 154–168). Berlin: Springer.

Howley, I., & Rosé, C. P. (2018). Empirical evidence for evaluation anxiety and expectancy-value theory for help sources. *Proceedings of the International Conference of the Learning Sciences (ICLS'18)*. London, UK.

Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-tailored student support in introductory physics. *PLoS ONE*, *10*(9), e0137001.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. (2009). User experience over time: An initial framework. In *Proceedings of CHI09* (pp. 729–738). New York, NY: ACM Press.

Kay, J. (2000). Stereotypes, student models and scrutability. In C. Cauthier & G. Van Frasson (Eds.), *Intelligent tutoring systems: 5th International Conference ITS 2000, LNCS* (Vol. 1839, pp. 19–30). Montreal, Canada: Springer.

Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, this issue [50th Anniversary Special issue].

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). New York, NY: Cambridge University Press.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30–43.

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* (pp. 43–56). Boca Raton, FL: CRC Press.

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

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 *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 421–430).

Lan, A. S., Studer, C., Waters, A. E., & Baraniuk, R. G. (2013). Tag-aware ordinal sparse factor analysis for learning and content analytics. In *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, TN, USA* (pp. 90–97).

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

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. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 1386–1394). Red Hook, NY: Curran Associates, Inc.

Liu, R., & Koedinger, K. R. (2017a). Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, *9*(1), 25–41.

Liu, R., & Koedinger, K. R. (2017b). Going beyond better data prediction to create explanatory models of educational data. In Lang, C., Siemens, G., Wise, A. F., & Gaevic, D. (Eds.), *The Handbook of learning analytics* (1 ed., pp. 69–76). Alberta, Canada: Society for Learning Analytics Research (SoLAR).

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* (pp. 107–113). London, UK.

Lovett, M. C. (1998). Cognitive task analysis in service of intelligent tutoring systems design: A case study in statistics. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Intelligent tutoring systems, lecture notes in computer science* (Vol. 1452, pp. 234–243), New York, NY: Springer.

Lovett, M., Meyer, O., & Thille, C. (2010). JIME-The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media Education*, *2008*(1), Art-13.

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research*. London: Routledge.

Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. In *Proceedings of the 36th Annual Cognitive Science Conference* (pp. 1126–1131).

Paquette, L., Baker, R. S., de Carvalho, A., & Ocumpaugh, J. (2015). Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In *Proceedings of the 23rd Conference on User Modelling, Adaptation and Personalization, Dublin, Ireland* (pp. 183–194).

Ritter, S., Blessing, S. B., & Wheeler, L. (2003). Tools for component-based learning environments. In T. Murray, S. B. Blessing, & S. Ainsworth (Eds.), *Authoring tools for advanced learning environments* (pp. 467–489). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Rosé, C. P., & Ferschke, O. (2016). Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses, *International Journal of AI in Education*, 25th Anniversary Edition, *26*(2), 660–678.

Simon, H. A. (1967). The job of a college president. *Educational Record*, *48*(Winter), 68–78.

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaboration? *Discourse, tools and reasoning* (pp. 361–384). Berlin, Heidelberg: Springer.

Wen, M. (2016). *Investigating virtual teams in massive open online courses: Deliberation-based virtual team formation, discussion mining and support* (Submitted in partial fulfillment of the PhD degree). Carnegie Mellon University, School of Computer Science, Language Technologies Institute, Pittsburgh, PA.

Wen, M., Maki, K., Dow, S. P., Herbsleb, J., & Rosé, C. P. (2018). Supporting virtual team formation through community-wide deliberation. In *Proceedings of the 21st ACM Conference on Computer-Supported Cooperative Work and Social Computing, New York, NY, USA*.

Wen, M., Maki, K., Wang, X., & Rosé, C. P. (2016). Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. In *Proceedings of Educational Data Mining (EDM 2016), Raleigh, NC, USA* (pp. 533–538).

Yang, D., Piergallinin, M., Howley, I., & Rosé, C. P. (2014). Forum thread recommendation for massive open online courses. In *Proceedings of the 7th International Conference on Educational Data Mining, London, UK* (pp. 257–260).

Yang, D., Wen, M., Howley, I., Kraut, R., & Rosé, C. P. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second ACM Conference on Learning @ Scale (L@S '15), Vancouver, BC, Cananda* (pp. 121–130).