Perspectives on Computational Models of Learning and Forgetting

Florian Sense[1, 2, 3] (f.sense@rug.nl)
Tiffany S. Jastrzembski[4] (tiffany.jastrzembski@us.af.mil)
Michael C. Mozer[5] (mozer@colorado.edu)
Michael Krusmark[3] (michael.krusmark.ctr@us.af.mil)
Hedderik van Rijn[1, 2] (d.h.van.rijn@rug.nl)

[1] Department of Experimental Psychology, University of Groningen, The Netherlands
[2] Behavioral and Cognitive Neuroscience, University of Groningen, The Netherlands
[3] L-3 Technologies, Wright-Patterson Air Force Base, Dayton, OH, USA
[4] Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA
[5] Department of Computer Science & Institute of Cognitive Science,
University of Colorado, Boulder, CO, USA

Abstract

Technological developments have spawned a range of educational software that strives to enhance learning through personalized adaptation. The success of these systems depends on how accurate the knowledge state of individual learners is modeled over time. Computer scientists have been at the forefront of development for these kinds of distributed learning systems and have primarily relied on data-driven algorithms to trace knowledge acquisition in noisy and complex learning domains. Meanwhile, research psychologists have primarily relied on data collected in controlled laboratory settings to develop and validate theory-driven computational models, but have not devoted much exploration to learning in naturalistic environments. The two fields have largely operated in parallel despite considerable overlap in goals. We argue that mutual benefits would result from identifying and implementing more accurate methods to model the temporal dynamics of learning and forgetting for individual learners. Here we discuss recent efforts in developing adaptive learning technologies to highlight the strengths and weaknesses inherent in the typical approaches of both fields. We argue that a closer collaboration between the educational machine learning/data mining and cognitive psychology communities would be a productive and exciting direction for adaptive learning system application to move in.

Keywords: learning; forgetting; computational models; recurrent neural networks; process models; naturalistic data; educational application

Introduction

Imagine leading cognitive scientists came together for a conference—in Montreal, for example—and decided to build the best possible adaptive system to support student learning. A successful adaptive learning system would draw upon our theoretical understanding of human memory and its temporal dynamics: How does knowledge and skill develop with practice? How do memory traces decay over time? Which individual differences in these processes can be exploited to best adapt to individual learners?

Taking this hypothetical endeavor seriously is a productive thought experiment because it makes explicit the gap between our theoretical understanding—based primarily on research conducted in psychology laboratories—and practical applications—worked on primarily by computer scientists.

These two disciplines have largely operated in parallel, and both fields could benefit greatly from collaborating more closely. Mutual benefits for coming together will likely include an enhanced theoretical understanding of learning and memory through access to big, naturalistic data; and improved practical applications achieved through exploitation of robust and well-studied psychological principles.

Here, we will discuss a number of recent efforts to help bridge this interdisciplinary gap. We will present promising approaches to build adaptive learning systems from both the computer science and cognitive science/psychology fields, highlighting the strengths and weaknesses afforded by each type of approach. We will focus and structure the discussion around two recent reports stemming from real-world, educationally relevant use-cases: (1) the second language acquisition modeling (SLAM) challenge put forward by Duolingo, and (2) a comparison of the utility of different computational models to personalize review in a middle school classroom.

Duolingo's SLAM Challenge

The well-known online language-learning platform Duolingo[1] recently posed a challenge to the scientific community. They made data available from more than 6,000 users who independently studied English, Spanish, or French at their own pace, across a duration of 30 days on their platform. Using a corpus of 7+ million annotated words, Duolingo invited research teams to submit computational models to predict users' performance at a later point.[2] In their report of the competing models, they frame this approach—

[1] https://www.duolingo.com/

[2] Interestingly, Settles et al. describe the task as: "Given a history of errors made by learners of a second language, the task is to predict errors that they are likely to make at arbitrary points in the future" (2018, p. 56).

second language acquisition modeling (SLAM)—as a new computational task (Settles, Brust, Gustafson, Hagiwara, & Madnani, 2018). Settles et al. elaborate that educational software has made advances in simpler domains but that less is known about how beginners acquire second languages in realistic settings. As such, their challenge is a special case of "building the best possible adaptive learning system."

Fifteen research teams responded to Duolingo's challenge, encompassing a multitude of approaches used to submit predictions. Most competing teams came from the field of natural language processing due to the fact that Duolingo posed the challenge in the context of a large computational linguistics conference.[3] An analysis of the types of algorithms used to power the predictions suggested that non-linear algorithms—recurrent neural networks (RNNs)—were especially successful, while linear models—item response theory variants—were least successful (Settles et al., 2018; Table 3). In fact, the top models demonstrating the highest predictive validity were all considered non-linear, suggesting that SLAM was mainly approached as deep knowledge tracing (Piech, Bassen, Huang, & Ganguli, 2015) in which RNNs are used to trace student performance over time.

It is interesting to note that none of the teams who submitted model predictions explicitly accounted for the cognitive processing mechanisms involved or how those processes unfold over time. These types of *process models* have been the focus of study in cognitive psychological research, but that research has remained largely in the realm of controlled, laboratory tasks.

In the following two subsections, we will discuss RNNs and process models respectively, to highlight the strengths and weaknesses of both types of models.

Recurrent neural networks (RNNs)

The dominance of RNNs in the Duolingo challenge is not surprising, given their flexibility in discovering useful representations from large amounts of data (LeCun, Bengio, & Hinton, 2015), enabling these models to leverage the rich meta-data available for each instance in the corpus (see Figure 3 in Settles et al., 2018). What is surprising, however, is that these models do not have a clear representation of time, which is of course a crucial dimension of learning (Bloom, 1974). Settles et al. state that none of the models explicitly considered that the passage of time affected acquisition and/or forgetting. This disregard for the temporal dynamics of learning and retention seems surprising given what is known about the spacing effect (e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993). It is further surprising to glean that Duolingo itself explicitly models time non-linearly, taking the shape of the forgetting curve into account (Settles & Meeder, 2016).

An analysis of the features that the different models used (see Section 5.2 and Table 4 in Settles et al., 2018) suggests that only the *response time* and *days in course* features had marginally significant effects on the quality of predictions—the modeling architecture (RNN or additive IRT) was the main driver of the differences between the teams. Notably, the *days in course* information for each entry could have been translated to the time that elapsed since the last encounter with an item (i.e., lag-time) in order to explicitly model forgetting as a non-linear function of lag-time. Instead, however, "forgetting was either modeled through engineered features (e.g., user/token histories), or opaquely handled by sequential RNN architectures" (Settles et al., 2018; Section 4).

Simply considering the *sequence* in which events occurred (rather than lag-time) is common in knowledge tracing models (Corbett & Anderson, 1995) and often works well because student behavior is usually modeled in a single session. Consequently, most Bayesian knowledge tracing models do not assume that forgetting takes place at all (see Khajah, Lindsey, & Mozer, 2016 for a BKT variant that does consider forgetting). The benefit of considering lag-time between (rather than the mere sequence of) events as input might only emerge if data are modeled on sufficiently long timescales, across which accurately modeling forgetting curves should be more important.

In a recent effort, Mozer, Kazakov, and Lindsey (2017) introduced an explicit representation of continuous time (CT) in a RNN that they trained on 11 different data sets. The hypothesis behind creating the CT-RNN variant was that including certain constraints might guide the model in its learning—essentially protecting it against its own flexibility (Mozer et al., 2017). Mozer et al. motivate their approach by drawing a helpful analogy with vision, in which models constrained to take known regularities into account outperform unconstrained models (in decyphering handwriting, for example: LeCun, Bottou, Bengio, & Haffner, 1998).To the surprise of the authors, their CT-RNN did not perform any better than the RNN that did not take CT into account, but was otherwise functionally identical. What is more: removing elapsed time from the input stream altogether did not impair the default RNN's performance by more than 5% at most, suggesting that that it did not incorporate temporal information to the extent one might expect.

Their null findings are surprising in light of earlier work that demonstrated the power of taking statistical regularities in the temporal dynamics of forgetting into account. For example, Khajah, Lindsey, and Mozer (2016) extended a Bayesian knowledge tracing model and showed that it performed as well as a RNN knowledge tracing model. Their extensions were based on psychological principles—such as exponential decay of knowledge over time, which is usually not assumed in Bayesian knowledge tracing—that constrained the potential patterns that their model could learn from the data relative to the deep knowledge tracing model.

[3] Specifically, the "13th Workshop on Innovative Use of NLP for Building Educational Applications" held at NAACL-HLT 2018 (http://naacl2018.org/).

More importantly, they highlight the fact that the processes that are assumed to influence learning and forgetting are explicitly expressed in the model's specification: The model parameters correspond to psychological concepts of theoretical relevance. For example, how quickly skill X decays for student Y, or how much students vary in their abilities.

RNNs are currently the preferred choice of computer scientists because of their flexibility to learn arbitrary representations from copious amounts of data. The very architecture guaranteeing this flexibility, however, poses a risk to overfitting the data and makes it extremely difficult to interrogate the model. For adaptive learning systems to be used in practice, systems powered by RNNs may preclude the ability of the system to understand what the learner may optimally require or why the learner is struggling. For these reasons, researchers in psychology—whose main goal is to describe underlying cognitive processes—have not embraced RNNs. Instead, they have developed process models.

Process models

In process models, theoretical assumptions regarding underlying cognitive processes are hard-coded in the model itself. A prime example of an overarching architecture of process models is the Adaptive Control of Thought–Rational (ACT-R; Anderson, 2007) framework[4], which implements testable theories of human memory processing, and supports the creation of cognitive models that are capable of predicting and explaining human behavior. ACT-R has been used to successfully account for a depth and breadth of phenomena, including language comprehension, learning and memory, problem solving and decision, and even interpretation of fMRI data.

With regards to adaptive learning systems, ACT-R has been leveraged by the intelligent tutoring community to minimize the distance between student and expert models. In the case of algebra tutors, for example, ACT-R models each step for solving a problem explicitly, and functions by identifying the root cause for student errors. It then provides the appropriate assistance and mentoring for the individual student to remediate the identified error. These cognitive tutors are highly successful for helping students *acquire* knowledge (Anderson, Corbett, Koedinger, & Pelletier, 1995). Practically speaking, however, they fail to include decay mechanisms, so they lack the ability to account for maintenance or sustainment needs long-term.

A number of process models have focused on and extended ACT-R's declarative memory module to model the temporal dynamics of learning and forgetting in greater detail. Pavlik and Anderson (2005) extended ACT-R to account for effects of spacing using an activation-based decay mechanism. They applied this model iteratively and demonstrated success in making real-time predictions for a language learning task, nicely pushing the bounds of computational modeling application for real-world educational use. More recent extensions incorporated response latencies for each learning event to better trace memory strengths over time, showing promise in both laboratory (Sense, Behrens, Meijer, & van Rijn, 2016) and real-life learning tasks (Sense, van der Velde, & van Rijn, 2018; van Rijn, van Maanen, & van Woudenberg, 2009). In addition, this model fared well when evaluated against a range of *theoretical* criteria (see Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018), however, gaps were noted in its ability to make out-of-sample predictions, particularly at long temporal horizons, or to account for the speeded benefit of relearning when initial practice was initially more spaced.

The Predictive Performance Equation (PPE) is another model that explicitly captures the spacing effect, motivated in its development to remediate limitations of existing models. PPE leverages and combines elements of the General Performance Equation (Anderson & Schunn, 2000), ACT-R, and the New Theory of Disuse (Bjork & Bjork, 1992). This novel computational account of the spacing effect has demonstrated its theoretical and applied validity across a breadth of empirical data (see Walsh et al., 2018). PPE has built upon the shoulders of giants previously described and pushed into the *prescriptive* realm for real-world applications. This means that real-time predictions are iteratively made and successive, optimal training schedules are immediately delivered to the individual learner. PPE has successfully been applied to the domain of cardiopulmonary resuscitation (CPR), demonstrating greater performance effectiveness and minimized training time to acquire and sustain proficiency through personalized, precision learning capability (Jastrzembski et al., 2017).

PPE is unusual in its focus on *prescribing* training schedules in real-life tasks and conditions (but also see, e.g.: Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009), as most process models are primarily developed and evaluated for theoretical purposes. PPE exemplifies the capabilities and limitations of process models more generally: When the relevant processes in a particular domain are mapped onto the mechanics of a model, those models can extrapolate from the available data to make cognitively-plausible prescriptions. Model parameters directly map onto concepts relevant to the modelled domain and can be interpreted and communicated meaningfully (e.g., "Your ability is very high, but this is an unusually difficult fact to learn. You should rehearse this item four hours sooner than the other facts in this set.")

The downside, however, is that process models do not readily translate to new domains or even similar tasks within the same domain. Model parameters that capture individual learning and forgetting signatures often vary across domains and tasks (e.g., Sense et al., 2016). Therefore, using the parameters estimated for a person in one domain, does not mean their performance profile can automatically be accurately predicted in another domain. However, recent work with PPE showed that prior data may be used to *inform* free parameters (Collins, Gluck, Walsh, & Krusmark, 2017; Collins, Gluck, Walsh, Krusmark, & Gunzelmann, 2016),

[4] http://act-r.psy.cmu.edu/about/

indicating that the model does not have to start from scratch in every domain.

Another issue is that most process models with potential for adaptive learning are based on very sparse inputs: lag-time, sometimes response latency, and accuracy—which is often aggregated to reduce noise. Thus, the models are not inherently equipped to leverage the rich meta-data available in, for example, the Duolingo data in the way that RNNs are.

In the final section, we will discuss potential ways of "moving forward" but before, we turn from an online learning platform to the classroom in order to discuss a recent effort to deploy adaptive learning software in realistic educational settings.

Personalized Review in the Classroom

Duolingo's challenge to the scientific community is instructive because it reflects a clearly defined task that an adaptive learning system must perform: modeling second language acquisition (Settles et al., 2018), i.e., predicting future performance given a corpus of learning history. The preceding discussion of how well a number of computational models might be able to perform this task is a productive way to compare the models' theoretical assumptions. If we take the goal of *building the best possible adaptive learning system* seriously, however, we must also keep the end users in mind: the learners.

Today, learners increasingly engage with study materials in distributed learning environments and the culture of learning is changing. While lectures will be scheduled at fixed times, more and more aspects of learning are now self-directed, self-paced, and available on demand in online learning environments. Traditional, structured classroom settings, which are different from Duolingo's learning environment, progressively move towards incorporating distributed learning approaches to aid face-to-face interactions (e.g., Sense et al., 2018). The *best* adaptive learning system would function in realistic, modern educational settings, in which learners follow courses that expose them to materials in a prescribed sequence; in which there might be regular quizzes on subsets of the material; and in which the goal is to perform well on a (cumulative) exam at the end of the course. The ideal system would be able to inform each learner about their progress, the current state of their knowledge, which elements of the course they should focus on, and assist them in their self-regulated learning decisions (Bjork, Dunlosky, & Kornell, 2013).

One elucidatory effort deployed retrieval-practice software as part of the curriculum in a middle school (Lindsey, Shroyer, Pashler, & Mozer, 2014). In a semester-long Spanish course, 179 students engaged with a flashcard tutoring system during class time. Each week, they completed three 20- to 30-minute sessions: In the first and second, the week's new materials were studied to proficiency before reviewing old materials; in the third, a test of the week's new materials was administered. The authors tested three different algorithms that scheduled items during review. The personalized spacing algorithm resulted in the highest performance on the cumulative end-of-semester exam, with especially high performance for items that were introduced early in the semester (Lindsey et al., 2014). The algorithm was dubbed DASH—because it incorporated information regarding item difficulty, student ability, and study history—and the authors argue that their model is in principle agnostic with regards to the domain that is modeled as long as knowledge in that domain can be deconstructed into "primitive knowledge components" (Lindsey et al., 2014, p. 643), which is comparable to the assumptions made by ACT-R in general and PPE in particular (see above).

Mozer and Lindsey (2016) discuss the DASH framework more generally in a recent book chapter—aptly subtitled "psychological theory matters in the big data era"— in which they argue that theory-inspired models such as ACT-R and the multiscale context model (Mozer et al., 2009) can inform theory-agnostic machine learning approaches, specifically collaborative filtering. In this framework, collaborative filtering is used to estimate difficulty and ability from the study history (again: DASH) to infer a student's knowledge state. The generalized power-law of forgetting (Wixted & Carpenter, 2007) can then be used to project the decay of knowledge into the future. Mozer and Lindsey discuss variations of their DASH framework that vary with regards to the information that is considered when instantiating forgetting curves. Their simulation results suggest that for the tested scenarios, individual differences in both learning and forgetting should be considered and that models do much worse if they do not take forgetting into account at all. In two experiments, the authors provide strong empirical evidence that personalized review is more effective than other forms of spacing, which is in line with other research rejecting one-size-fits-all approaches to spacing (Mettler, Massey, & Kellman, 2016).

Conducting experiments of this kind in schools imposes additional administrative and logistic costs on a research project that are not required if large online learning platforms make their data available to researchers (e.g., Ridgeway, Mozer, & Bowles, 2017). A more accessible, educationally relevant context for most researchers might be provided by the classrooms of the universities they work at (e.g., Sense et al., 2018). Ultimately, the best possible adaptive learning system must be tested for effectiveness and usability by real learners, not on historical data.

Moving Forward

Moving forward, we believe it is crucial that cognitive scientists engage with the educational data mining community in order to test their process models with naturalistic data. This will allow cognitive scientists to demonstrate the usefulness of formulating relevant cognitive processes explicitly and to learn from approaches commonly used to model learning in computer science. A productive way forward might be to formally evaluate models of different types against each other to map out the boundary conditions for which the strengths and weaknesses of each class apply. For example: In which domains does each type of model fare best? What types of data does each type of model optimally function with? And perhaps most critically, can the strengths of one model alleviate the weaknesses of another through integration?

One potential path towards leveraging the strengths of both process models *and* RNNs is to have the models collaborate when making predictions. The DASH model proposed by Lindsey et al (2014), for example, could be simultaneously fit with a RNN using gradient descent. Instead of making independent predictions, the models would sum the two model predictions to make a single prediction. With predictions thus combined, the RNN will learn the residual between the restricted but interpretable DASH and the actual data. This would maintain the interpretable parameters of DASH and exploit the flexibility of the RNN at the same time. The specific implementation of DASH proposed by Lindsey et al. could be replaced with any other process model, of course, and the collaborative predictions could be weighted to give preferential treatment to either the process model or the RNN.

Although significant progress has been made to close the gap between computational models and educational or training practice application, it is important to realize that literature is sparse or nonexistent for timescales and contexts most keenly relevant to formal educational institutions where typical summer breaks invoke an inherent acceptance of knowledge decay each year (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; McCombs, Augustine, & Schwartz, 2011); or for military training, where irregular delays between training and use is common and maintenance of readiness for high-risk, low-volume skills is a significant challenge. Thus, additional research must be conducted to evaluate the applied utility of any computational model that could be of practical use.

We argue that a multidisciplinary, collaborative approach bringing the power of neural network and process modeling approaches together, would be an exciting direction for adaptive learning system application to move in (also see Mozer, Wiseheart, & Novikoff, 2019). It would acknowledge the value of the human-in-the-loop by integrating our theoretical understanding of the human memory system with RNNs' ability to make sense of large data; thereby pulling their affordances together in a unified task to build the best adaptive learning system possible.

Acknowledgements

This research was supported by NSF grants SES-1461535 and DR-L1631428.

References

Anderson, J. R. (2007). *How can the human mind exist in the physical universe? Oxford Series on Cognitive Models and Architectures*. New York, NY: Oxford University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of Learning Sciences*, *4*(2), 167–207.

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R Learning Theory : No Magic Bullets. In R. Glaser (Ed.), *Advances in Instructional Psychology: Educational Design and Cognitive Science* (Vol. 5, pp. 1–34). Mahwah, NJ: Lawrence Erlbaum.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of Foreign Language Vocabulary and the Spacing Effect. *Psychological Science*, *4*(5), 316–321. http://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*, 417–44. http://doi.org/10.1146/annurev-psych-113011-143823

Bloom, B. S. (1974). Time and Learning. *American Psychologist*, *29*(9), 682–688. http://doi.org/10.1037/h0037632

Collins, M. G., Gluck, K. A., Walsh, M., & Krusmark, M. (2017). Using Prior Data to Inform Initial Performance Predictions of Individual Students. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1800–1805). Madison, WI.

Collins, M. G., Gluck, K. A., Walsh, M., Krusmark, M., & Gunzelmann, G. (2016). Using Prior Data to Inform Model Parameters in the Predictive Performance Equation. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 75–80). Philadelphia, PA.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, *66*(3), 227–268. http://doi.org/10.3102/00346543066003227

Corbett, A. T., & Anderson, J. R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Jastrzembski, T., Walsh, M. M., Krusmark, M., Kardong-Edgren, S., Oermann, M., Dufour, K., … Stefanidis, D. (2017). Personalizing training to acquire and sustain competence through use of a cognitive model. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented cognition. Enhancing cognition and behavior in complex environments* (pp. 148–161). Switzerland: Sprinter International Publishing AG.

Khajah, M. M., Lindsey, R. V, & Mozer, M. C. (2016). How Deep is Knowledge Tracing? *ArXiv Preprint ArXiv:1604.02416v2*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*. http://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998).

Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lindsey, R., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving Students' Long-Term Knowledge Retention Through Personalized Review. *Psychological Science*, *25*(3), 639–647. http://doi.org/10.1177/0956797613504302

McCombs, J. S., Augustine, C. H., & Schwartz, H. L. (2011). *Making summer count: How summer programs can boost children's learning*. Rand Corporation.

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A Comparison of Adaptive and Fixed Schedules of Practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. http://doi.org/10.1037/xge0000170

Mozer, M. C., Kazakov, D., & Lindsey, R. V. (2017). Discrete-Event Continuous-Time Recurrent Nets. *ArXiv Preprint ArXiv:1710.04110*.

Mozer, M. C., & Lindsey, R. (2016). Predicting and Improving Memory Retention: Psychological Theory Matters in the Big Data Era. In M. N. Jones (Ed.), *Big Data in Cognitive Science.* (pp. 43–73). Psychology Press.

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1321–1329). La Jolla, CA: NIPS Foundation.

Mozer, M. C., Wiseheart, M., & Novikoff, T. P. (2019). Artificial intelligence to support human instruction. *Proceedings of the National Academy of Sciences*, *116*(10), 3953–3955. http://doi.org/10.1073/pnas.1900370116

Pavlik, P. I., & Anderson, J. R. (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-based Model of the Spacing Effect. *Cognitive Science*, *29*(4), 559–86. http://doi.org/10.1207/s15516709cog0000_14

Piech, C., Bassen, J., Huang, J., & Ganguli, S. (2015). Deep Knowledge Tracing. *Advances in Neural Information Processing Systems*, 505–513.

Ridgeway, K., Mozer, M. C., & Bowles, A. R. (2017). Forgetting of Foreign-Language Skills: A Corpus-Based Analysis of Online Tutoring Software. *Cognitive Science*, *41*, 924–949. http://doi.org/10.1111/cogs.12385

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. http://doi.org/10.1111/tops.12183

Sense, F., van der Velde, M., & van Rijn, H. (2018). Deploying a Model-based Adaptive Fact-Learning System in a University Course. In *Proceedings of the 16th International Conference on Cognitive Modeling* (p. 138). Madison, WI.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second Language Acquisition Modeling. In *Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 56–65).

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Association for Computational Linguistic (ACL)*, 1848–1858.

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. In *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115). Manchester, UK.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect. *Cognitive Science*, *42*, 644–691. http://doi.org/10.1111/cogs.12602

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren Power Law and the Ebbinghaus Savings Function, *18*(2), 133–134.