Learning Analytics Challenges: Trade-offs, Methodology, Scalability

Radek Pelánek
xpelanek@fi.muni.cz
Masaryk University
Brno, Czech Republic

ABSTRACT

Ryan Baker presented in a LAK 2019 keynote a list of six grand challenges for learning analytics research. The challenges are specified as problems with clearly defined success criteria. Education is, however, a domain full of ill-defined problems. I argue that learning analytics research should reflect this nature of the education domain and focus on less clearly defined, but practically essential issues. As an illustration, I discuss three important challenges of this type: addressing inherent trade-offs in learning environments, the clarification of methodological issues, and the scalability of system development.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; • **Applied computing → Education**.

KEYWORDS

trade-offs, scalability, methodology

ACM Reference Format:
Radek Pelánek. 2020. Learning Analytics Challenges: Trade-offs, Methodology, Scalability. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3375462.3375463

1 INTRODUCTION

At LAK 2019 keynote, Ryan Baker presented challenges for future learning analytics research [5]. The LAK 2020 has as a theme "Shaping the future of the field," and calls for papers that "explicitly address the theme of this year's conference by reflecting on past, present, and future research". In this context, I would like to present thoughts on the type of suitable challenges and research directions for learning analytics and to polemize with the approach proposed by Ryan Baker.

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03. . . $15.00
https://doi.org/10.1145/3375462.3375463

Ryan Baker presented six grand challenges (with symbolic prizes) [5]. The main point of the proposal, inspired by Hilbert's problems in mathematics, is to have the challenges defined as well-structured problems with clearly defined success criteria, e.g., the challenge of applicability of knowledge tracing beyond the screen is specified by requiring sufficient predictive ability with respect to a given metric and a threshold value.

I argue that such kind of challenges should not play the central role in learning analytics research. This kind of well-structured problems with clear goals can be clearly useful for moving research forward, getting attention and focus, e.g., as competitions associated with conferences and workshops. It is useful, however, mainly for short-term progress. It is less suitable as grand challenges and a long-term focus. The domain of education is fundamentally ill-structured and complex and includes inherent trade-offs [8]. Any learning environment must balance between many dimensions of the learning process (e.g., learning efficiency, student engagement, time spent) and needs to take into account the variability of conditions (e.g., differences between learning outcomes, learning processes, types of knowledge components, student predispositions). Too much focus on a single goal risks that we ignore other important aspects of learning.

It is useful to consider a well-known challenge in a closely related research field—Netflix prize in recommender systems [7]. The challenge was to improve the predictive accuracy of collaborative filtering techniques—10% improvement of RMSE with respect to a state-of-the-art algorithm. This is a typical example of a well-structured problem with a clear goal. The Netflix prize was definitely useful in bringing attention to the research field, but that may be, to a large degree, the effect of prize money (1 million dollars) associated with the challenge. The competition provided an impulse for research—several novel prediction approaches were developed for the competition. However, it is disputable how important the challenge was for the overall progress of the recommender systems field [18]. The field moves forward thanks to the clarification of methodology and widening its scope beyond simple criteria used in the challenge, e.g., by using novel data for recommendations (context), using other criteria beyond accuracy (e.g., diversity), or considering novel application domains for recommendations.

For the field of learning analytics, I believe that rather than focusing on specific challenges with clearly defined goals, we should focus more on hard-to-grasp areas that are, nevertheless, fundamental for the practical impact of learning analytics research. Specifically, I discuss three such areas: trade-offs, methodological issues, and scalability. This is not intended as a complete list or "the most important" list. I deliberately focus on issues that are close to my area of expertise—the development of online practice systems. The main purpose of the discussed issues is to serve an illustration of the kind of challenges that need to be tackled.

2 TRADE-OFFS

The development of learning systems (and education in general) faces many trade-offs—issues where the improvement in one aspect leads to deterioration in another aspect. These trade-offs typically do not have any "correct" solution and thus are hard to study. Yet, they are often very practically important, and their study offers interesting research challenges.

A typical example is the mastery criterion [35]. Mastery learning is a common element of personalized learning systems. The mastery criterion decides when to stop the practice of a given topic and move the student to a more advanced topic. Such a criterion typically has to balance between the risk of under-practice, which leads to problems in future learning as the student has not mastered the topic sufficiently, and redundant over-practice, which leads to a waste of students time and may lead to demotivation.

Another common trade-off is between engagement and learning. We want our learning systems to be both engaging and lead to efficient learning. But these two aspects are often in (at least partial) conflict. Features that increase engagement (e.g., gamification elements or decrease of difficulty of practice) often lead to lower efficiency of learning [32].

Some features of learning systems may be double-edged. Consider, for example, hints: these can be very useful for supporting student learning, but they may also lead to gaming-the-system behavior where students abuse the presence of hints [1, 4].

A specific example of a trade-off is the setting of thresholds. Many learning analytics techniques contain thresholds that influence their behavior and whose setting involves trade-offs between different aspects of their performance. A typical example is an affect detector [10], where the setting of thresholds varies precision and recall of techniques. Although research papers often summarize the performance by a single number (e.g., the AUC metric), a practical application needs to choose a specific threshold depending on the relative importance of false positive and false negatives in a particular application. This choice of a particular number is seldom addressed in research papers.

Some trade-offs concern the conflict between the interests of research and students [27]. Other trade-offs involve practical aspects that are often neglected in research: implementation costs, maintenance, technical dept [38]. Research papers (and Rayn Baker's challenges) typically focus on optimizing the predictive performance of learning analytics techniques. But from a practical perspective, it makes sense to sacrifice some predictive accuracy when we can use a simpler model. Research thus should focus not just on performance, but also on "implementation costs". This is not easy. What is a good measure for quantifying the implementation cost of a particular technique? What is a good balance between accuracy and cost? Such questions are difficult to grasp, but they are worthwhile challenges.

The trade-offs are hard to study but important. In addition to the study of specific trade-offs, it is also worthwhile to focus on general methods for studying trade-offs. They are an inherent part of education, so we better have tools for studying them. Many trade-offs involve the choice of thresholds. How do we compare thresholds? What are suitable ways of visualizing trade-offs and the impact of the choice of a particular threshold? How can we explore changes of a threshold using historical data? How can we perform experiments to optimize thresholds? Techniques like multiarmed-bandits [39] aim to optimize a single measure ("profit"). How do we use them to optimize several conflicting criteria?

Trade-offs are clearly not unique to education. Many other research areas have to deal with trade-offs. It is thus not meaningful to try to focus on completely general methods for studying trade-offs, but rather on transferring, clarifying, and improving techniques that have proved useful elsewhere (e.g., Pareto-efficient algorithms in recommender systems [26]).

3 METHODOLOGY

Baker's challenges [5] aim at developing techniques that achieve predictive accuracy with respect to a particular value of one particular metric (e.g., AUC > 0.65); the challenges do not provide any details on the experimental methodology that should be used (e.g., the use of stratification when dividing data into training and testing set). This corresponds to the common current practice in research in learning analytics and educational data mining. In many cases, researchers pick a single metric and report evaluation with respect to this metric, without specifying details of the used methodology [34]. Current tools like LearnSphere [21] try to facilitate the research by providing ready-made solutions for data analysis. In such cases, researchers may be actually unaware of the methodological details of the realization of their experiments.

This state of affairs is, however, problematic. Details of the methodology are important. Such important, but often neglected details are for example the choice of a performance metric (e.g., AUC versus RMSE), details of the computation of metrics (global, averaging per student, averaging per item), or the division of data into training and testing set [34].

Another important methodological issue in learning analytics is the presence of biases. In naturally occurring educational data (outside of small lab studies), biases are often strong and can significantly influence data analysis. Typical biases in data are mastery bias, where the length of student trace in data is (negatively) correlated with their skill, and self-selection bias, where some answers are provided only by a specific subset of students. Although biases in educational data are studied in the literature (e.g., [9, 15, 19, 30, 34]), there is currently no coherent methodology for dealing with biases, and many research works do not take this issue into consideration.

The clarification of methodological issues is particularly important in the context of the increasing popularity of the use of neural networks and other complex black-box models [37]. It is dangerous to start developing such kind of models when the methodological issues are not clearly settled; otherwise we risk that we train the black-box models in an unsuitable way and will not be able to recognize it. This problem is well illustrated by one of the first papers that tried to use deep learning techniques on educational data [36]. The authors of the paper claimed large improvements in predictive accuracy; later analysis showed that their comparison suffers from several methodological problems [20, 40, 41]. The need to be cautious about claims of neural networks performance is not specific to our domain—recent papers in recommender systems [11] and information retrival [25] make this point very clearly.

What progress do we need with respect to methodology? What are the challenges for future research? The methodological issues need further clarification with clear identification of potential problems in the current research. Such clarification can be often done using simulated data, which enable us to illustrate specific issues in a simplified setting [15, 31]. This approach is often used in psychometrics research [13], whereas in learning analytics it is quite marginal. We also need clear terminology to describe methodological choices. A clear terminology would enable researches to specify their experiments concisely and thus facilitate the understanding of results and future replication of studies.

Replication and reproduction of past studies do not sound like fancy challenges but are very important. The goal should be not just to test the generalizability of previously presented results to new settings (a different learning system, a different student population), but also check whether the reported results are not artifacts of some methodological choices (e.g., the specific choice of a performance metric).

Many learning analytics techniques are not completely generalizable. Educational data from different sources (e.g., learning of facts versus learning of rules, elementary school students versus university students) differ in many aspects and often require different treatment. Instead of focusing on evaluation which tries to find which technique is "better" [16], it may be more fruitful to focus on studying the "What works when?" question, e.g., by mapping techniques to taxonomies and educational frameworks like the Bloom taxonomy [3] or the Knowledge-Learning-Instruction framework [22].

With respect to biases in data, many challenges lie ahead. We need to better identify and describe different types of biases present in educational data. We need to develop methods for checking whether a particular data set contains a specific bias (a specific example of such analysis is provided in [9]). And finally, since it is practically impossible to obtain realistic data without biases, we need methods for dealing with biases.

4 SCALABILITY

Scalability is important from several perspectives. The most straightforward is the technical aspect. We need our learning analytics techniques to scale to data sets that occur in practice. This means that more focus should be paid to the computational efficiency of techniques. For example, research on student modeling typically focuses only on predictive accuracy and completely ignores the issue of computational efficiency. However, for methods like Bayesian modeling, the computational complexity of parameter fitting can be a major obstacle to practical adoption.

Moreover, it is useful to consider the scalability issue more generally. If we want our research to scale from lab studies to practical adoption, we need to address also scalability of the development of systems and their educational content. A large portion of learning analytics research currently happens in the United States or other rich countries with a large population. This places the research in a specific context: a large and rich target audience for educational products and consequently high competition among educational products; only products with very good features can compete; it is feasible to pay relatively large teams of developers and content authors to develop products. This context implicitly shapes the research carried by research teams.

However, a large portion of the Earth's population lives and learns in other settings—in poorer economies or countries with languages spoken by a much smaller number of people than English. In these settings, there is much smaller competition in educational products, but the market also much smaller and thus it is not feasible to pay large teams. Whereas in the US setting it may be economically feasible to develop optimized tutoring system for learning equations, in other settings it is necessary to have a much larger scope of content to gain an audience for the sustainable operation of a learning system.

In these settings, scalability of techniques and ease of development become central issues. How do we develop (intelligent) learning systems that scale? How do we efficiently develop and maintain content for a learning system? These are practical questions, which can, however, lead to interesting research challenges. This kind of development could (and should) be still based on research. We just need research with slightly different priorities—not optimizing just their performance, but also taking into account their scalability and ease of application. As a specific example, consider models of student knowledge [33]. Such models are typically evaluated only with respect to their predictive accuracy. Complex models often achieve only slightly better performance than simpler models [6]. When is the difference practically important to justify the use of a complex model in the implementation?

As another example, consider the "adaptive design" [2] approach to development, where the analysis of data provides an impulse for the redesign of a learning system. Such kind of "closing the loop" studies [23] provides a nice example of a practical learning analytics research that leads to improvement in learning. But such analysis is currently rather time-consuming—if we need to invest many hours of expert research in improving one aspect of geometry learning, it is not feasible to apply the approach in resource-limited settings. A worthwhile challenge is to develop this kind of methods in a more scalable fashion.

It may also be useful to consider a "debugging perspective" that is common in computer science and software engineering. This perspective takes into account that people make mistakes, i.e., we start by assuming that systems contain errors and that it is necessary to develop techniques for automated detection of these errors. Similar perspective could be useful for inspiring learning analytics research. We should admit that our learning system contains errors and try to find them efficiently. The errors could be in the software implementation—these should be left for software engineers to deal with. Other errors, however, concern the learning content (typos, wrong answers, misleading formulations), meta-data (mapping of items to knowledge components), and models (poor parametrization, unsatisfied assumptions). How can we automatize the detection of such kinds of errors? Recent research considered such questions, e.g., in content analytics methods [24] or Q-matrix refinement [12], but more research in this direction would be useful.

On a high-level perspective, a key scalability question is the identification of priorities: What are the most cost-effective places to improve a system? Where should the attention of system developers go? This kind of question (specifically formulated as "What's most broken?") has recently started to be studied [28].

The issue of scalability is clearly not important "just for the rest of the world". Resources are always restricted, and scalability issues are being addressed even by US-based authors, e.g., in the mentioned "What's most broken?" study [28] or in the exploration of crowdsourcing for content development [17]. But the relative research attention given to the topic is quite low, whereas in many settings it is an issue of central importance.

5 CONCLUSIONS

In the presented discussion, I have provided examples of challenges in three specific directions: trade-offs, methodology, scalability. I do not claim that the discussed directions and challenges are the most important for the learning analytics field. There are many other important, ill-defined, tricky issues like privacy [14], user interface (e.g., the design of dashboards) [29], or more focus on long-term behavior (e.g., taking forgetting into account more systematically) and long-term goals. Concerning the long term, [18] makes a point of explicitly including timeframe into the formulation of the recommender problem; their argument is closely relevant to learning analytics as well.

The main point of my argument is not about specific challenges. The main point is to argue that this *type* of open-ended research challenges is in learning analytics more important than the well-structured challenges with clearly specified goals. Challenges with clear goals and clear winners have their place. But let us use them mainly as devices for making short-term progress. For long-term progress, we need to accept that our field is complex and faces ill-defined problems with many trade-offs and that our research must necessarily reflect its nature.

REFERENCES

[1] Vincent Aleven and Kenneth R Koedinger. 2000. Limitations of student control: Do students know when they need help?. In *Proceedings of Intelligent Tutoring Systems*. Springer, 292–303.
[2] Vincent Aleven, Elizabeth A McLaughlin, R Amos Glenn, and Kenneth R Koedinger. 2016. *Handbook of research on learning and instruction*. Routledge, Chapter Instruction based on adaptive learning technologies, 522–559.
[3] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2000. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, abridged edition*. Pearson.
[4] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in gaming the system behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.
[5] Ryan S Baker. 2019. Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *JEDM—Journal of Educational Data Mining* 11, 1 (2019), 1–17.
[6] Joseph Beck and Xiaolu Xiong. 2013. Limits to accuracy: How well can we do at student modeling?. In *Proceedings of Educational Data Mining*.
[7] Robert M Bell and Yehuda Koren. 2007. Lessons from the Netflix prize challenge. *SiGKDD Explorations* 9, 2 (2007), 75–79.
[8] David C Berliner. 2002. Educational research: The hardest science of all. *Educational researcher* 31, 8 (2002), 18–20.
[9] Jaroslav Čechák and Radek Pelánek. 2019. Item Ordering Biases in Educational Data. In *Proceedings of Artificial Intelligence in Education*. Springer International Publishing, 48–58.
[10] Ryan SJ d Baker, Sujith M Gowda, Michael Wixon, Jessica Kalka, Angela Z Wagner, Aatish Salvi, Vincent Aleven, Gail W Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra.. In *Proceedings of Educational Data Mining*. ERIC.
[11] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 101–109.
[12] Michel Desmarais, Behzad Beheshti, and Peng Xu. 2014. The refinement of a Q-matrix: Assessing methods to validate tasks to skills mapping. In *Proceedings of Educational Data Mining*.
[13] Michel C Desmarais and Ildiko Pelczer. 2010. On the faithfulness of simulated student performance data. In *Proceedings of Educational Data Mining*.
[14] Hendrik Drachsler and Wolfgang Greller. 2016. Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of Learning Analytics & Knowledge*. ACM, 89–98.
[15] Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. 2013. Simulated Students, Mastery Learning, and Improved Learning Curves for Real-World Cognitive Tutors.. In *AIED Workshops Proceedings*.
[16] Yue Gong, Joseph E Beck, and Neil T Heffernan. 2011. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 27–46.
[17] Neil T Heffernan, Korinn S Ostrow, Kim Kelly, Douglas Selent, Eric G Van Inwegen, Xiaolu Xiong, and Joseph Jay Williams. 2016. The Future of Adaptive Learning: Does the Crowd Hold the Key? *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 615–644.
[18] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systemsbeyond matrix completion. *Commun. ACM* 59, 11 (2016), 94–102.
[19] Tanja Käser, Kenneth R Koedinger, and Markus Gross. 2014. Different parameters – same prediction: An analysis of learning

curves. In *Proceedings of Educational Data Mining*. 52–59.

[20] Mohammad Khajah, Robert V Lindsey, and Michael C Mozer. 2016. How deep is knowledge tracing?. In *Proceedings of Educational Data Mining*.

[21] Ken Koedinger, Ran Liu, John Stamper, Candace Thille, and Phil Pavlik. 2017. Community based educational data repositories and analysis tools. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 524–525.

[22] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[23] Kenneth R Koedinger, John C Stamper, Elizabeth A McLaughlin, and Tristan Nixon. 2013. Using data-driven discovery of better student models to improve student learning. In *Proceedings of Artificial intelligence in education*. Springer, 421–430.

[24] Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, Marek Hatala, and George Siemens. 2015. Content analytics: The definition, scope, and an overview of published research. *Handbook of Learning Analyitcs* (2015), 77–92.

[25] Jimmy Lin. 2019. The neural hype and comparisons against weak baselines. In *ACM SIGIR Forum*, Vol. 52. ACM, 40–51.

[26] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A Pareto-efficient Algorithm for Multiple Objective Optimization in e-Commerce Recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 20–28.

[27] Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. 2014. Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits.. In *Proceedings of Educational Data Mining*. 161–168.

[28] Shiven Mian, Mononito Goswami, and Jack Mostow. 2019. What's Most Broken? Design and Evaluation of a Tool to Guide Improvement of an Intelligent Tutor. In *Processings of Artificial Intelligence in Education*. Springer, 283–295.

[29] Inge Molenaar and Carolien Knoop-van Campen. 2017. Teacher dashboards in practice: Usage and impact. In *Proceedings of European Conference on Technology Enhanced Learning*. Springer, 125–138.

[30] R Charles Murray, Steven Ritter, Tristan Nixon, Ryan Schwiebert, Robert GM Hausmann, Brendon Towle, Stephen E Fancsali, and Annalies Vuong. 2013. Revealing the Learning in Learning Curves. In *Proceedings of Artificial Intelligence in Education*. Springer, 473–482.

[31] Juraj Nižnan, Radek Pelánek, and Jan Papoušek. 2015. Exploring the Role of Small Differences in Predictive Accuracy using Simulated Data. In *Proceedings of AIED Workshop on Simulated Learners*.

[32] Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Impact of question difficulty on engagement and learning. In *Proceedings of Intelligent Tutoring Systems*. Springer, 267–272.

[33] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 313–350.

[34] Radek Pelánek. 2018. The details matter: methodological nuances in the evaluation of student models. *User Modeling and User-Adapted Interaction* 28 (2018), 207–235. Issue 3. https://doi.org/10.1007/s11257-018-9204-y

[35] Radek Pelánek and Jiří Řihák. 2018. Analysis and design of mastery learning criteria. *New Review of Hypermedia and Multimedia* 24 (2018), 133–159. Issue 3.

[36] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Proceedings of Advances in Neural Information Processing Systems*. 505–513.

[37] Patrick Riley. 2019. Three pitfalls to avoid in machine learning. *Nature* 572 (2019), 27–29.

[38] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*. 2503–2511.

[39] Joannes Vermorel and Mehryar Mohri. 2005. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of European conference on machine learning*. Springer, 437–448.

[40] Kevin H Wilson, Xiaolu Xiong, Mohammad Khajah, Robert V Lindsey, Siyuan Zhao, Yan Karklin, Eric G Van Inwegen, Bojian Han, Chaitanya Ekanadham, Joseph E Beck, et al. 2016. Estimating student proficiency: Deep learning is not the panacea. In *Proceedings of Neural Information Processing Systems, Workshop on Machine Learning for Education*.

[41] Xiaolu Xiong, Siyuan Zhao, Eric Van Inwegen, and Joseph Beck. 2016. Going Deeper with Deep Knowledge Tracing.. In *Proceedings of Educational Data Mining*. 545–550.