Opportunities for Human-AI Collaborative Tools to Advance Development of Motivation Analytics

Steven C. Dang
Carnegie Mellon University, Human Computer Interaction Institute
stevenda@cs.cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University, Human Computer Interaction Institute
koedinger@cmu.edu

ABSTRACT: Modern educational technology products must support learners' cognitive and motivational needs. Extending models of learner motivation to new products leads to difficulties in generalizing existing models to drastically different systems or operationalizing existing behavioral theories with sufficient construct validity. Overcoming these challenges requires teams with both learning science and data science skillsets. Recent advances in automated machine learning and interpretable machine learning have led to opportunities to empower learning scientists with the capabilities of an interdisciplinary data science team. Through a review of prior studies on motivation analytic development, we identify common data science challenges and review some successful algorithmic solutions. We also identify challenges to scaffolding these tasks to users without data science backgrounds and highlight some advances in automated machine learning and interpretable machine learning that may enable development of tools and services to fill this need.

Keywords: Motivation, Self-Regulated Learning, Automated Machine Learning, Interpretable Machine Learning, Measurement, Online learning environments

1 INTRODUCTION

There are many challenges in developing motivational analytics for online learning environments. Developers must go beyond analytics of easily observable interactions and utilize models that take into account the affective and self-regulated learning dynamics of learners in order to draw inferences on their motivations (Eccles & Wigfield, 2002). Developing high quality analytics requires accounting for a range of concerns around construct validity and reliability while also leveraging complex modeling methods that lie outside the realm of theories of learner motivation (Milligan, 2018). An open challenge in the measurement of motivation, like self-regulated learning (SRL), lies in the challenge of how to operationalize constructs on different systems (Roll & Winne, 2015). Overcoming these design challenges requires a mixture of learning science knowledge and data science skills. This can be a significant barrier for many educational technology product companies that do not have the resources to hire experts with such specialized skillsets. With recent advances in technologies such as interpretable and automated machine learning, there is an opportunity to make data science skills accessible to learning scientists and dramatically increase the pool of individuals capable of developing high quality motivational analytics. In this paper, we review prior work in analytic development for measurement of motivational constructs using strictly log data. In this work, we identify particular challenges to developing measures of motivational constructs and identify specific areas of opportunity

for tools and algorithmic development that can greatly lower the barriers to development of motivational analytics.

2 CHALLENGES IN MOTIVATION MEASUREMENT

Learners' motivation is a product of their SRL, and as highlighted by Winne (2010), SRL is contextual and context evolves as learners regulate their learning. The challenge of measuring motivation requires inferring many latent contextual influences such as learners' goals, metacognition, and task value (Winne & Hadwin, 2008). Many of these factors can be difficult to measure, but learner's displayed learning behaviors and strategies can be indicators of their latent motivational factors as evidenced in work by Dang & Koedinger (2019a). Prior work in learning analytics and educational data mining has elaborated models of affect, SRL strategies, and relevant learning behaviors (Lang et al, 2017). However, extending these models to new systems and datasets introduces a host of new challenges (Winne, 2014).

For instance, Rowe et al (2009) extended a model of off-task behavior (Baker et al, 2004) to a narrative-centered learning environment. This open-ended environment involved actions such as navigating a character around a virtual world, interacting with objects, and talking with non-player characters. In lieu of a usable measurement model, the authors developed an alternate operationalization of off-task behavior that required insight into the pedagogical value of possible interactions in the game. Performing this task required a degree of learning science knowledge that is typically not found in many data scientists in the work force.

Rowe et al were motivated to develop an off-task behavior measurement model by the findings of Baker et al, indicating that student learning is negatively impacted by such behaviors. To test their hypotheses, the authors collected additional motivational survey and achievement data. Exploratory analysis demonstrated that students engaged in off-task behavior about 15% (SD=8.9%) of the time, which differs from the 20% frequency found by Baker et al. The statistical analysis indicated that unlike Baker's off-task model, the Rowe model was not significantly related to pre-post learning. Likewise, the results found no relationship between off-task behavior and either achievement orientation or self-efficacy, contradicting the results of Dang & Koedinger (2019a). This analysis demonstrates two common challenges in extending theory to new systems. Performing such a validation requires a set of data science skills that are lacking in many industry learning scientists.

Also, despite appearing to match the Baker et al construct on its face, the evidence indicates that the model developed by Rowe et al was not measuring the same construct. This construct validity problem challenge is an open problem in scaling the research of the learning analytics and educational data mining communities to more online learning environments (Huggins-Manley et al, 2019). One common process for validating a construct on a system is to collect ground truth data, as done by Rowe et al, and to leverage this information to validate a proposed model by applying appropriate statistical tests for agreement with expectations. As online learning environments expand their available content and are deployed to broader audiences, the challenge becomes how to sample a representative dataset to train sufficiently general models, an increasingly cost prohibitive task.

3 OPPORTUNITIES FOR HUMAN-AI COLLABORATION

3.1 Leveraging Multiple Facets to Bootstrap Construct Validity

One challenge in using observed behaviors to measure some latent motivational construct is that, unlike in experimental contexts, multiple constructs are likely implicated in any given behavior. Huggins-Manley et al (2019) discuss several relevant threats to construct validity. Construct confounding occurs when "inferences are drawn on one construct even though indicators reflect more than one construct". Confounding constructs with facets of constructs occurs when "only some facets of a construct are measured, invalidating inferences about the full construct". Mono-operation bias occurs because "a single indicator of a construct underrepresents the inferred construct, which is more complex than a single indicator.

In our prior work, we attempted to tackle these threats to construct validity by leveraging multiple indicators of the target construct (Dang & Koedinger, 2019b). In this work, we operationalized our latent construct, diligence, through a series of metrics defined around how readily learners start work and how long learners can maintain focus. Analysis demonstrated that combined factors yielded both better predictions, reliability, and alignment with motivational factors as measured through correlation with survey-based motivation instruments. These results point to an interesting foundation for tools and services that support a construct operationalization process by leveraging multiple facets of a construct defined in the available behavior data in lieu of ground truth labels to perform construct validation and iteration.

3.2 Fitting Parameters using a Multi-faceted Latent

Operationalizing a measurement model for classifying a target behavior is a complex process involving a combination of expert learning science knowledge to understand the types of constructs to target in the data and data science knowledge to identify how to measure such targets in the data. Aleven et al (2006) demonstrate an a-priori thresholding process for operationalizing SRL theory into a measurement model on fine-grained data. The model consists of a set of if-then-else rules representing a decision tree model for a pattern of help seeking hypothesized by SRL theory. In order to apply this model to the data, the model operationalized concepts such as "Familiar at all?" and "Sense of what to do?" using a set of calculated values in the data and thresholds that are set to values that the authors describe as "intuitively plausible, given our past experience". This a-priori heuristic is difficult to reproduce and requires an intuitive sense of how users interact with the system, which is not necessarily experience many product development teams possess.

Baker et al (2004) demonstrate a typical approach in the machine learning community, treating the model definition problem as a supervised machine learning problem. Ground truth labels of gaming the system behavior were collected from classroom observations and were used to train a machine learning model on the data. This model fitting process requires a degree of data science experience to both identify the correct algorithm to apply to the data given an understanding of the deeper structure of the problem and to elaborate a set of raw features from the raw data that can improve the ability the algorithm to find a well-fitting model. New toolkits in the automated machine learning (auto-ml) community have simplified this process of feature engineering. For instance, Kanter & Veeramachaneni (2015)

developed the Featuretools framework that leverages deep machine learning algorithms to automatically elaborate meaningful features over the raw data and takes a user-defined goal to automatically identify appropriate algorithms to fit a model that solves the target problem.

Another challenge in applying machine learning algorithms is evident in how Kuvalja et al (2014) leverage a pattern recognition algorithm to identify patterns of behavior that were indicative of SRL processes. In order to apply the algorithm, the authors defined three parameters: the minimum number of occurrences of a pattern, the probability of observing the pattern, and a threshold for how often a pattern must be observed in some time interval. Setting values for these algorithm parameters appears similar to the a-priori threshold setting demonstrated by Aleven et al (2006) but requires both knowledge relevant to the occurrence of SRL behaviors in practice and an understanding of the algorithm. Work in auto-ml also tackles this problem of automated parameter selection. For instance, Kandasamy et al (2019) leverage bayesian optimization to automatically identify the optimal parameter values to use for a particular machine learning algorithm to fit a model to a given data set.

Beyond the issue of lack of ground truth labels discussed in section 3.1, building high quality models of behavior for an online learning environment requires a number of other data science skills not typical of the training for many learning scientists. Advances in auto-ml have demonstrated a capacity for intelligent algorithms to tackle many of these tasks with minimal input from users, making such tasks more accessible to non-data science users. While a broader survey of auto-ml is beyond the scope of this work, we point to Zöller & Huber (2020) for a more comprehensive survey of available auto-ml frameworks and their capabilities.

3.3 Identifying Heterogeneity

Many online learning environments leverage fine-grained moment-by-moment behavior data to inform analytics. However, the contribution of Aleven et al (2006) highlights how many learning science theories do not make strong predictions about exactly how motivational factors influence learner's decision-making given some specific combination of contextual factors that we can observe in the data and bridging this gap is a not trivial. Data scientists aggregate learner behaviors to test theoretically predicted relationships that should be evident across contexts. Such aggregation techniques make it difficult to identify where a model may be inadequately capturing the target construct. For instance, a disengaged learner is expected to be lower performing than a more engaged learner because learning necessarily requires completion of some work to engage with concepts. Engagement might reasonably be operationalized as total time on task. Shih et al (2011) demonstrate that the quality of a student's engagement is evident in the speed of a student's response to a problem immediately following a request for help. Thus, two students might appear very similar in their total time working, but within that time, students' varying levels of cognitive engagement is evident in the differences in response time following instructional assistance provided by the learning environment. Using a simple operationalization of engagement would miss this source of variation in the data. Available model performance metrics and model interpretation tools lack adequate support for learning scientists to think critically about the performance of their current models and identify these shortcomings (Kaur et al, 2019).

3.3.1 Scaffolding Model Iteration with Qualitative Data Analysis

We believe that data programming (Ratner et al, 2016) offers an interesting approach to allowing users to encode expert knowledge onto datasets. In this paradigm, users are asked to define rules, similar to the if-then-else rules defined by Aleven et al (2006), that encode some heuristic that experts might leverage to make classification judgements when reviewing learner behavior. These rules can be collected across multiple users and be overlapping and partially disagree with each other. The set of rules are used to infer labels for the data. While this is an interesting approach for enabling users to more naturally encode their knowledge onto the data, there remains the question of how to support users in realizing what knowledge might be relevant?

Baker & de Carvalho (2008) demonstrated that users with learning science knowledge and product familiarity are capable of reviewing segments of learner behavior data and drawing valid inferences on what that learner may be doing. This is a viable method to leverage learning scientists to identify instances of learner behavior that may be incorrectly classified and activating relevant knowledge that could be used to define heuristic rules to describe the model shortcoming. However, there are several barriers to supporting users in searching for segments of learning actions that may reflect some unknown but currently unaccounted for adjustment to the current model. We review some of the challenges we have faced in applying qualitative analysis to iterate on models of motivation and highlight opportunities for intelligent algorithms and tools to improve this workflow.

3.3.2 Prioritizing Data for Qualitative Review

As discussed previously, learning science theory typically can only make predictions about learner behavior when aggregated to student-level analytics. However, the goal of the qualitative analysis process is to support experts by reviewing cases of fine-grained behavior. This implies a two-step process where users must first identify a student to review and from that student's data, specific instances of behavior must be selected for review. Datasets can involve hundreds if not thousands of students, and just an hour of student data can translate to several hundred data points. There is an opportunity to improve the efficiency of this exploration process by leveraging algorithms to prioritize data to review.

In the first stage of the process, users need support in identifying which student's data to analyze first. Learning science theory informs users expectations for relationships between student-level aggregated variables in the data. These relationships can be used by anomaly detection algorithms to associate each student with some degree of non-fit to expectation and prioritize students accordingly. Anomaly detection can be as simple as defining a linear regression predicting some relationship between the available data and using the size of the prediction error as a data ranking. More complex anomaly detection methods are available, and we reference Chandola et al (2009) for a more comprehensive review of this literature. We believe tools that can provide such a ranking mechanism in addition to an ability to review metadata for each student would greatly improve users' ability to more effectively decide how to explore available student data.

Within the subset of data from an individual learner, the next step of the qualitative analysis process is to select which behavioral data to review first. The greatest challenge in this space is in supporting a search process where the target of the search is unknown. Anomaly detection algorithms can be applied to the raw data and then the data could be prioritized

within student based on how anomalous the data appears to be. However, such methods do not leverage unencoded expert knowledge to support the ranking process. Recent SRL research has applied pattern-mining methods to identify relevant behaviors because SRL behaviors are driven by learning processes and which causes observable reoccurring patterns in the data (Molenaar & Järvelä, 2014). We believe it would be valuable to cluster and summarize students' behaviors and then to leverage interpretable machine learning frameworks such as interpret ML (Nori et al, 2019). These frameworks can enable users to understand the type of behavior encapsulated by a cluster in terms of the key contextual and behavioral features in the data. Similar to the method proposed by Baker & de Carvalho, expressing the data summaries in terms of these low-level details can allow experts to infer what may be happening and identify possible unexpected relationships. Together, these anomaly detection and interpretable machine learning techniques offer new avenues to surface relevant structure in the data that can inform learning scientists while searching through the vast quantity of learner data.

4 DISCUSSION

The major challenge in motivational measurement lies in identifying developers with both the data science and learning science knowledge to competently build measurement models from existing prior work in the learning analytics and educational data mining fields. We have demonstrated through examples in our prior work as well as others that there are opportunities for applying advances in autonomous machine learning and interpretable machine learning to empower learning science experts without experience in data science to be able to perform the same construct operationalization processes for building motivational analytics. We believe this is a promising open area of research for tools and algorithm development that can greatly accelerate the adoption of motivational analytics in online learning environments throughout the marketplace.

REFERENCES

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. International Journal of Artificial Intelligence in Education, 16(2), 101-128.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004, April). Off-task behavior in the cognitive tutor classroom: when students" game the system". In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383-390).

Baker, R., & de Carvalho, A. (2008, June). Labeling student behavior faster and more precisely with text replays. In *Educational Data Mining 2008*.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185-224.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.

Dang, S., & Koedinger, K. R. (2019a). Exploring the Link Between Motivations and Gaming. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 276-281).

Dang, S., & Koedinger, K. R. (2019b). Towards a Behavior-Based Psychometric Instrument for Unobtrusive Measurement of Diligence. Manuscript submitted for publication.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. Annual review of psychology, 53(1), 109-132.

Huggins-Manley, A. C., Beal, C. R., D'Mello, S. K., Leite, W. L., Cetin-Berber, D. D., Kim, D., & McNamara, D. S. (2019). A commentary on construct validity when using operational virtual learning environment data in effectiveness studies. *Journal of Research on Educational Effectiveness*, *12*(4), 750-759.

Kandasamy, K., Vysyaraju, K. R., Neiswanger, W., Paria, B., Collins, C. R., Schneider, J., ... & Xing, E. P. (2019). Tuning hyperparameters without grad students: Scalable and robust Bayesian optimisation with Dragonfly. *arXiv preprint arXiv:1903.06694*.

Kanter, J. M., & Veeramachaneni, K. (2015, October). Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1-10). IEEE.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Vaughan, J. W. (2019). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Conditionally accepted at the 2020 ACM CHI Conference on Human Factors in Computing Systems (CHI 2020)

Kuvalja, M., Verma, M., & Whitebread, D. (2014). Patterns of co-occurring non-verbal behaviour and self-directed speech; a comparison of three methodological approaches. *Metacognition and learning*, *9*(2), 87-111.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Milligan, S. K. (2018, March). Methodological foundations for the measurement of learning in learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 466-470).

Molenaar, I., & Järvelä, S. (2014). Sequential and temporal characteristics of self and socially regulated learning. *Metacognition and Learning*, *9*(2), 75-85.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*.

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014, July). Towards Understanding Expert Coding of Student Disengagement in Online Learning. In *CogSci*.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems* (pp. 3567-3575).

Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, *2*(1), 7-12.

Rowe, J. P., McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2009, July). Off-Task Behavior in Narrative-Centered Learning Environments. In *AIED* (pp. 99-106).

Shih, B., Koedinger, K. R., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of educational data mining*, 201-212.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational psychologist*, *45*(4), 267-276.

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, *9*(2), 229-237.

Winne, P. H., & Hadwin, A. F. (2012). The weave of motivation and self-regulated learning. In *Motivation and self-regulated learning* (pp. 309-326). Routledge.

Zöller, M. A., & Huber, M. F. (2019). Survey on automated machine learning. *arXiv preprint arXiv:1904.12054*.