The Potentials of Educational Data Mining for Researching Metacognition, Motivation and Self-Regulated Learning

PHILIP H. WINNE
Faculty of Education
Simon Fraser University
winne@sfu.ca

RYAN S.J.D. BAKER
Department of Human Development
Teachers College, Columbia University
baker2@exchange.tc.columbia.edu

Our article introduces the Journal of Educational Data Mining's Special Issue on Educational Data Mining on Motivation, Metacognition, and Self-Regulated Learning. We outline general research challenges for data mining researchers who conduct investigations in these areas, the potential of EDM to advance research in this area, and issues in validating findings generated by EDM.

Keywords: Educational Data Mining, Metacognition, Motivation, Self-Regulated Learning

1. INTRODUCTION

The triumvirate of motivation, metacognition and self-regulated learning (SRL) offers significant affordances to learning science at the same time as researching these constructs poses substantial challenges. On the side of affordances, these three factors, broadly construed, provide raw material for engineering the bulk of an account about *why* and *how* learners develop knowledge, beliefs, attitudes and interests. Add to these factors additional data about a learner's prior knowledge – its amount, cognitive configuration and retrievability – plus a short list of "primitive" capabilities such as working memory span and inhibition, and it is theoretically plausible to assemble a comprehensive model of learning.

On the side of challenges when researching these constructs lies a thicket of issues. In our view, three predominate. The first challenge is identifying indicators of motivation, metacognition and SRL. Part of this challenge is deciding how to divide motivation, metacognition, and SRL into categories that can be modeled. In making these decisions, related categories such as engagement and affect become relevant. In operationally defining indicators, it is important to obtain representations of learning as it unfolds, i.e., the *process* of learning, that are clearly and precisely matched to what theory describes. While the term "process" may imply a temporally continuous activity, as an arrow flies through the air, learning is not that. The process of learning is a temporal succession of states. The question of what marks the onset and offset of a state – what is a "just noticeable difference" – is the crux of this challenge.

Having defined a state in theory, the second challenge to investigating these constructs is gathering data that can be validly interpreted as instantiations of states. Choices of kinds of instruments for recording data and the

We would like to thank Belinda Yew for helpful comments and suggestions.

engineering of them introduce two weighty issues. First is the reliability with which the instrument is capable of gathering robust data, i.e., measurements that vary minimally when there is truly no change in a state. Second is the purity of the measurement. This is classically expressed in terms of a ratio of signal to noise. In empirical work, instruments always are subject to unreliability and never achieve a signal-to-noise ratio of 1. How much unreliability is tolerable? What is the magnitude and nature of error when interpreting a datum because the signal-to-noise ratio is imperfect, i.e., what is the validity of data? And what other factors influence the noise of measurement?

A third critical challenge to researching motivation, metacognition and SRL arises when learning scientists manipulate data – analyze it – in attempts to ameliorate unreliability, less than perfect signal-to-noise ratios and characterize how learners learn. The statistical correlation is a basic example of an aggregate index used in this way. We remind readers that such statistics inherently force simplifications that obscure the genuine phenomena expressed in each learner's behavior. How much of an error of prediction is error of unreliability or due to less than perfect signal-to-noise? The statistic is mute about an answer that has critical implications for theorizing.

We also note a potentially important constraint about the articles in this issue. We and the articles herein focus on students and on learning activities carried out in school or school-like contexts. Whether the theory and empirical findings presented here generalize to other learning contexts – work, pursuing a hobby or military training, to name a few – is yet to be determined.

1.1 Metacognition

It is common to introduce discussions of metacognition by characterizing it as "thinking about thinking." While not incorrect, that aphorism severely underrepresents the depth and scope of the construct. One window onto this complexity [Winne, 2011] is provided by jointly considering distinctions between (a) an object level versus a meta level of thought [Nelson & Narens, 1990], (b) forms of metacognitive knowledge and (c) monitoring and control.

Fundamentally, memory is a repository for knowledge. A learner can have knowledge of an object, e.g., the quadratic equation or the architecture and functions of the human circulatory system. A learner also can have knowledge about that knowledge – metaknowledge – such as it's form(s) for representing the knowledge (symbolic, graphical), affordances of those different representations' affordances, the generality of the knowledge, etc. Metaknowledge is not distinguished by its greater complexity or abstraction. Rather, metaknowledge stands in relation to knowledge that is its object.

Three fundamental forms of knowledge and meta-knowledge are widely recognized. Declarative knowledge is descriptive; e.g., the quadratic equation is graphed as a parabola. Procedural knowledge operates on or translates knowledge; e.g., re-presenting the algebraic representation of an instantiation of the quadratic equation as a graph. Conditional knowledge sets boundaries on knowledge; e.g., if the root of the quadratic is taken for a negative number, the quadratic cannot be resolved using real numbers.

Knowledge in memory can be used in two ways. First, knowledge can be used as a basis for examining experience (including thoughts). The process of examination is the cognitive operation of monitoring. Second, knowledge can encode a basic or a complex operation that is applied to information. For example, computing the roots of a quadratic expression can be carried out algebraically or graphically. In choosing one method, a learner exercises cognitive control. Monitoring depends on conditional knowledge. Cognitive control depends on procedural knowledge.

Metacognition, thus, is cognition that stands in relation to other cognition that is its object and that accomplishes classifications of experience and transformations of information. In exercising metacognitive monitoring and metacognitive control, learners actively engage in thinking about their learning and factors that bear on learning.

1.2 Motivation

Theories of motivation strive to account for why people initiate behavior, including thoughts, and why they continue to behave in a particular way or their change behavior. Implicit in this view of motivation are four key features. First, people have goals (preferences, ambitions, aspirations). Without goals, there would be no incentive to initiate, persist in or change behavior. Second, people hold beliefs that a particular behavior (or organizations of behavior: plans and activities) has a greater chance to reach a goal than other forms of behavior. Otherwise, random choice would suffice. Third, people perceive they have choice as well as efficacy to enact their chosen approach on a path to achieving their goal. If choice were unavailable, motives could not be expressed. If efficacy were lacking, intentions embodied in exercising choice would be irrational. Underlying these three features is the fourth: People are metacognitive in considering how they intend to behave; motivation is considered in relation to the object of motivation, and the behaviors that may be engaged in.

Among various contentious issues in the field of motivation research, one is the relationship of motives to affect (emotions in context). Affect can be forecast as a state that is likely to arise under particular conditions. In this case, an affective state plays the role of a goal that a learner is motivated to approach or to avoid. Affect also emerges inherently as people engage in activities. In this role, affect updates the conditions that contribute to shaping learners' thoughts and behavior. Recent work has examined affective states theorized to emerge during academic experiences, the so-called academic emotions [Pekrun et al., 2002], of boredom, confusion, frustration, and engaged concentration [cf. Baker, D'Mello, Rodrigo, & Graesser, 2010].

1.3 Self-Regulated Learning

Self-regulated learning (SRL) is a behavioral expression of metacognitively guided motivation. What marks SRL from other forms of regulation and complex information processing is that the goal a learner seeks has two integrally linked facets. One facet is to optimize achievement. The second facet is to optimize how achievement is constructed. This involves navigating paths through a space with dimensions that range over processes of learning and choices about types of information on which those processes operate.

With Hadwin, Winne proposed a 4-phase model of SRL [Winne & Hadwin, 1998; Winne, 2001, 2011]. We recap it here because we see this model as a lens for identifying and articulating potential foci for modeling SRL.

Before sketching the phases, three significant qualities of Winne and Hadwin's model merit highlighting. First, learners are agents. This means *they* choose which information to include in cognitive processing and which operations they will apply to it. This assumption entails that the effects of changes in a learner's environment – an experimenter's treatment, a peer's feedback, an avatar's recommendation or a computer program's request for input – are never deterministic. Nor are they stochastic because a human being cannot behave in truly random ways. Importantly, this assumption affords opportunity for learners to metacognitively monitor and exercise metacognitive control at any state of work on a task. Second, although Winne and Hadwin describe SRL as unfolding over phases, phases are weakly sequenced. Once learners accept a task (As agents, learners can reject and modify tasks set for them externally.), they typically proceed in a linear manner across phases 1-2-3-4. But this model explicitly allows any progression. Third, products of work a learner generates within a phase can be used recursively, thereby providing input to that same phase, as the cognitive system crystallizes a successive state at the next tick of information processing time.

In phase 1, learners scan their environment to identify attributes of a task as well as the resources and constraints they perceive as factors that may affect that task. The environment is explicitly divided to refer to external factors – e.g., access to Google, time limits, noisy distractions – and internal factors – e.g., self efficacy, incentives forecast for achievements of different sorts, feelings about the topic.

In phase 2, learners set goals. These play an important role because goals equate to standards learners use when metacognitively monitoring operations (e.g., rate, effort) and products (e.g., achievements, feelings, desires to attain or avoid particular results such as peers' positive or negative judgments of one's competence). As goals are formed, learners devise plans for reaching them. Metaphorically, they develop a production system of IF–THEN–ELSE elements that are forecast to realize goals.

Phase 3 is where learners carry out the task per se, "running" the production system previously devised. This is the phase where comprehension is developed, problems are solved and answers checked, and applications are designed and tested.

In phase 4, which Winne and Hadwin consider optional, learners review the mosaic of their work on the task. The purpose in this phase is "large scale" tuning or what Saloman and Perkins labeled forward reaching transfer [Salomon & Perkins, 1989].

Phase 3 of Winne and Hadwin's model highlights that a critical construct in modeling and researching SRL and motivation is engagement. Whether a student engages in learning actively or chooses to disengage from the activity, completely or in part, is governed by motivation and, in this context, is a self-regulated choice. Indeed, it can be challenging to disentangle "simple" engagement from SRL. Positive engagement can often only be clearly inferred from overt expressions of SRL, and many behaviors classified as disengagement actually express self-regulation choices. For instance, Baker et al. [2011] suggested that off-task behavior can be a positive emotional self-regulation strategy for alleviating negative aspects of boredom during learning, thereby actually helping to re-engage students. Additionally, in many cases only subtle differences discriminate sophisticated SRL strategies and "apparent" disengaged behaviors. For example, rapidly clicking through hints can indicate that the student is trying to "game the system" and bypass known but time-consuming material [Baker et al., 2004], or can indicate that the student is treating the hints as worked examples that afford self-explanation [Shih et al., 2008].

2. THE STATE OF THE FIELDS

Research programs investigating motivation, metacognition and self-regulated learning have been ongoing for approximately 100, 45 and 30 years, respectively. As might be predicted, the range of theories and models is proportional to each field's age. And, while multiple research programs within and spanning these fields have been progressive, there remain important gaps in theory. Particularly noticeable is the lack of robust findings that have empirically demonstrated benefits in promoting each *individual* student's learning.

Winne and Nesbit's [2010] recent review of research on academic achievement posited two lenses that characterize this overall body of work. By far the predominant lens focuses on studies of the psychology about "the way things are." By this, they meant phenomena in learning science "that, in principle, are universal among learners and across subject areas and are not likely under learners' control" [p. 654]. The upshot of these quests for findings about the way things are is concisely captured by Roediger [2008]:

> "The most fundamental principle of learning and memory, perhaps its only sort of general law, is that in making any generalization about memory one must add that 'it depends'" [p. 247].

As a contrast to that paradigm and in accord with the assumption that learners, as agents, will regulate their learning, Winne and Nesbit recommended an approach that investigates "the way learners make things." This paradigm has three main components. First, instruments should be developed to gather data that trace over time which information each learner operates on and how each learner operates on each of those selections. Second, from those data, temporally extended trajectories of engagement should be assembled for each learner. Third, in exploring those trajectories and their relations to other "snapshot" data (e.g., responses to self-report questionnaires, measures of aptitude or end-of-course achievement), learners should be grouped into homogeneous groups based on data rather than pinning hope on random assignment to neutralize sources of unidentified variance [see also Winne, 2006].

Winne and Nesbit argue that work that shines light on the way learners make things will have the double benefit of advancing learning science at the same time as it tailors interventions based on a learner's similarity to close neighbors. A challenge that will face this work, which data mining and other learning analytics are very well suited to support, will be resolving or rejecting the question of whether a learning science of "the way things are" can prosper.

3. THE POTENTIAL CONTRIBUTION OF EDUCATIONAL DATA MINING

One foundational premise of this special issue is that educational data mining, the field concerned with automated analysis of educational data, can play a signal role advancing research on motivation, metacognition and self-regulated learning. A full review of types of EDM method is outside the scope of this special issue [see Baker & Yacef, 2009; Romero & Ventura, 2010 for reviews along these lines] but we will discuss some key areas of application for this research area.

One key benefit of EDM methods is that they can be used to concretely identify and operationalize specific behaviors, theorized to represent motivation, metacognition, and/or SRL. EDM methods have been used to identify disengaged behaviors such as gaming the system [Baker et al., 2004] and behaviors that are irrelevant to the learning task [Sabourin et al., this issue]. They have also been used to identify student affective states [D'Mello et al., 2008], and to identify SRL behaviors within learning systems based on problem-solving [Aleven et al., 2006], concept-mapping [Rebolledo-Mendez et al., this issue], adaptive hypermedia [Bouchet et al., this issue], and exploratory learning environments [Amershi & Conati, 2009]. Models can be developed through prediction modeling (where human labels of a construct are obtained, and then a model is developed which can infer the human labels on new data), knowledge engineering (where models are developed rationally), and clustering (where models are developed in a bottom-up fashion from data), among other methods.

Once such a model is developed, it can be studied in fine-grained detail, as well as applied to new sets of data to study relationships between a behavior and other "snapshot" data, as discussed by Winne and Nesbit [2010]. This process has been termed "discovery with models" [Baker & Yacef, 2009]. For example, Sabourin et al. [this issue] studied the relationship between behaviors that are irrelevant to the learning task and the trajectory of affect over time. Rebolledo-Mendez and colleagues [this issue] examined the relationship between patterns characterizing how a student interacted with a pedagogical agent (which reflected SRL strategies) and student learning. Forsyth and colleagues [this issue] researched how student reading behaviors influenced dialogue with a pedagogical agent (which reflected SRL strategies) and student learning and affect. Bouchet and colleagues [this issue] investigated how adaptive versions of their learning system responded differently to students clustered by their SRL behavior.

Research that applies EDM to support discovery with models is one of the most popular methods in recent years, and a considerable amount of this research investigates SRL, metacognition, or motivation. For example, Amershi and Conati [2009] studied relationships between SRL behaviors in exploratory learning environments and learning outcomes. Beck et al. [2008] explored relationships between help-seeking behavior in intelligent tutoring systems and learning. Arroyo and Woolf [2005] examined the relationships between a range of SRL behaviors and motivations in an intelligent tutoring system.

An important thrust in EDM is analyzing student trajectories as they develop over time. Articles in this special issue illustrate the state-of-the-art in this respect. Sabourin et al. [this issue] looked at single steps in the interplay between student affect and disengagement/SRL behaviors. Kinnebrew et al. [this issue] went further to examine complex sequences over time using sequential pattern mining methods to find particularly common sequences of behavior, and then determine which sequences characterized generally successful vs. unsuccessful learners. Pavlik and colleagues [this issue] developed a dynamical system model that explicitly illuminated interlocking relationships between students' selections over time among multiple strategies and performance. Analyses over time in these papers are relatively micro in scope, involving single learning systems used over a course of hours; but similar methods have been used to study student trajectories over entire semesters [cf. Arnold, 2010] and even to link disengagement and affect in middle school to the eventual decision to attend university [San Pedro et al., in press].

One key constraint affecting EDM research of this type is the need to carefully consider model generalizability. For basic research within a single data set, it may be sufficient to validate a model's appropriateness for the current data set. But if the goal is to eventually apply the model to new students in similar or different contexts, it is critical to re-validate within those new data sets using a multi-level cross-validation approach that directly examines the model's generalizability to new students, new content and new populations).

It is also important to validate models with appropriate criteria. For instance, if the eventual use of a model (for discovery with models or intervention) depends on binary assignment to classes ({0,1}, signifying that a construct is absent or present), models should be evaluated using statistics appropriate to binary classifications, such as Cohen's k, precision, or recall. By contrast, if model confidence levels will be used in discovery with models or intervention, models should instead be evaluated using statistics that take model confidence into account, such as A′, the probability of discrimination. (A′ closely approximates AUC ROC, the area under the Receiver Operating Characteristic Curve). In general, if confidence-based analysis or intervention is possible, we recommend it, as it leverages more of the available information.

When models discovered using EDM are appropriately validated, as illustrated by articles in this special issue, research on metacognition, motivation, and self-regulated learning is advanced in a fashion that is both concrete and respects the short- as well as long-term nature of learning as a process.

4. CONCLUSIONS

A goal of this special issue is to stimulate interest in elaborating and advancing research in metacognition, motivation, and self-regulated learning by applying educational data mining methods. As articles in this special issue show, a generative literature is emerging that brings EDM into these fields and opens up exciting new perspectives on identifying and modeling events central to understanding learning processes. We alert readers to cognate work in the pages of previous issues of this journal and the EDM conference, as well as work that emanates in the proceedings of related conferences such as Artificial Intelligence in Education, Intelligent Tutoring Systems, and Learning Analytics and Knowledge.

We hold the view that educational data mining helps to operationalize constructs that are intrinsically slippery, to investigate multivariate relationships among constructs in this complex space, to model and understand student trajectories, and to clarify how far we can trust conclusions and findings. In these ways, EDM will help to boost the richness, concreteness, and usefulness of research in metacognition, motivation, and self-regulated learning.

REFERENCES

ALEVEN, V., MCLAREN, B., ROLL, I., and KOEDINGER, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *Proceedings of the International Conference on Artificial Intelligence and Education* 16, 101-128.

AMERSHI, S. and CONATI, C. 2009. Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining* 1, 1, 18-71.

ARNOLD, K.E. 2010. Signals: Applying Academic Analytics. *Educause Quarterly* 33, 1 (2010).

Ivon Arroyo and Beverly Park Woolf. 2005. Inferring learning and attitudes from a Bayesian Network of log file data. *Proceedings of the 12th International Conference on Artificial Intelligence and Education*, 33-40.

BAKER, R.S., CORBETT, A.T., and KOEDINGER, K.R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems.*Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.
BAKER, R.S.J.D., D'MELLO, S., RODRIGO, M.M.T., and GRAESSER, A.C. 2010. Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies* 68, 4, 223-241.
BAKER, R.S.J.D. and YACEF, K.. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining* 1, 1, 3-17.
BECK, J.E., CHANG, K., MOSTOW, J., and CORBETT, A. 2008. Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* 383-394.
D'MELLO, S.K., CRAIG, S.D., WITHERSPOON, A., MCDANIEL, B., and GRAESSER, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User Adapted Interaction* 18, 45-80.
NELSON, T.O. and NARENS, L. 1990. Metamemory: A theoretical framework and new findings. In G.H. BOWER (Ed.), *The psychology of learning and motivation*, Vol. 26, Academic Press, New York, 125-173.
PEKRUN, R., GOETZ, T., and TITZ, W. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research, *Educational Psychologist* 37, 2, 91-105.
ROMERO, C. and VENTURA, S. 2010. Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40*, 6, 601-618.
SALOMAN, G. and PERKINS, D.N. 1989. Rocky Roads to Transfer: Rethinking Mechanism of a Neglected Phenomenon. *Educational Psychologist 24,* 2, 113-142.
SAN PEDRO, M.O.C.Z., BAKER, R.S.J.D., BOWERS, A.J., and HEFFERNAN, N.T. In press. Predicting college enrollment from student interaction with an Intelligent Tutoring System in middle school. To appear in *Proceedings of the 5th International Conference on Educational Data Mining.*
SHIH, B., KOEDINGER, K.R., and SCHEINES, R. 2008. A response time model for bottom-out hints as worked examples. *Proceedings of the 1st International Conference on Educational Data Mining,* 117-126.
WINNE, P.H. 2001. Self-regulated learning viewed from models of information processing. In B.J. ZIMMERMAN and D.H. SCHUNK (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd Ed.). Lawrence Erlbaum Associates, Mahwah, 153-189.
WINNE, P.H. 2006. How software technologies can improve research on learning and bolster school reform. *Educational Psychologist* 41, 5–17.
WINNE, P.H. 2011. A cognitive and metacognitive analysis of self-regulated learning. In B.J. ZIMMERMAN and D.H. SCHUNK (Eds.), *Handbook of Self-Regulation of Learning and Performance*, New York, Routledge, 15-32.
WINNE, P.H. and HADWIN, A.F. 1998. Studying as self-regulated learning. In D.J. HACKER, J.E. DUNLOSKY, and A.C. GRAESSER (Eds.), *Metacognition in Educational Theory and Practice*, Lawrence Erlbaum Associates, Mahwah, 277-304.
WINNE, P.H. and NESBIT, J.C. 2010. The psychology of academic achievement. *Annual Review of Psychology* 61, 653-578.