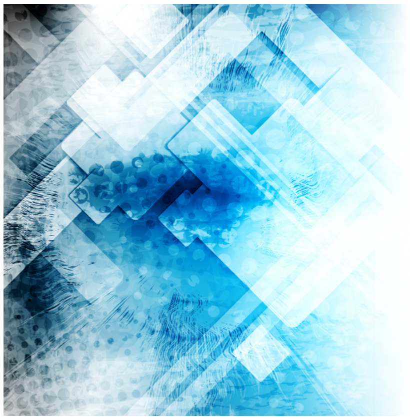


Abstract

The article addresses the question of how the assessment process with large–scale data derived from online learning environments will be different from the assessment process without it. Following an explanation of big data and how it is different from previously available learner data, we describe three notable features that characterize assessment with big data and provide three case studies that exemplify the potential of these features. The three case studies are set in different kinds of online learning environments: an online environment with interactive exercises and intelligent tutoring, an online programming practice environment, and a massive open online course (MOOC). Every interaction in online environments can be recorded and, thereby, offer an unprecedented amount of data about the processes of learning. We argue that big data enriches the assessment process by enabling the continuous diagnosis of learners' knowledge and related states, and by promoting learning through targeted feedback.

AUTHORS
Candace Thille, Ed.D.

Emily Schneider, B.A.

René F. Kizilcec, B.A.

Christopher Piech, M.S.

Sherif A. Halawa, M.Sc.

Daniel K. Greene, B.A.
Stanford University

The Future of Data–Enriched Assessment

A fundamental goal of education is to equip people with the knowledge and skills that enable them to think critically and solve complex problems. The process of quantifying the degree to which people have acquired such knowledge and skills is at the heart of assessment. Over the last decades, large–scale assessment of knowledge has become increasingly standardized, primarily to provide policy and other decision makers with clearer signals on the effectiveness of educational institutions and practices (Shavelson, 2007). Yet the merits of effective assessment extend far beyond informing policy decisions: instructors can gain valuable insights into the effectiveness of their instructional methods and learners receive feedback on their learning approach and overall progress. In providing an opportunity to apply the acquired knowledge and skills with subsequent feedback, assessment can promote learning if designed appropriately (Black & Williams, 1998; Gikandia, Morrowa, & Davisa, 2011; Roediger & Karpicke, 2006).

Education is becoming ever more augmented by technology to create new ways of interacting with educational content and communicating with instructors and peers. A number of promising technologies fall under the broad category of online learning environments, which rely on digital, networked systems but vary substantially in the features they provide to instructors and learners. Some such environments attempt to provide a holistic learning experience by integrating instruction, assessment, and social interaction. Other environments serve as a complementary resource to augment an in–person learning experience. In this paper, we present three case studies, which are set in different kinds of online learning environments: an online environment with interactive exercises and intelligent tutoring, an online programming practice environment, and a massive open online course (MOOC). The latter is an online learning environment in which thousands

CORRESPONDENCE

Email
cthille@stanford.edu

of people worldwide can learn about a given topic from lecture videos, quiz questions, longer assignments, and discussions with peers on a forum, to name but a few of the many forms of interaction that can occur in these environments (Kizilcec, Piech, & Schneider, 2013). Similar to non–educational online content providers, every interaction in these environments can be recorded and, thereby, offer an unprecedented amount of data about the processes of learning.

Online learning environments hold the potential to better support learning and to create opportunities for novel forms of assessment. The question we address in this article is: how will the assessment process with large–scale data derived from online learning environments be different from the assessment process without it? To address this question, we first explain our definition of big data, and how we believe it is different from previously available learner data. We then present three notable features that characterize assessment with big data and provide three case studies that exemplify the potential of these features. We argue that big data enriches the assessment process by enabling the continuous diagnosis of learners' knowledge and related states, and by promoting learning through targeted feedback.

Big Data

How will the assessment process with large–scale data derived from online learning environments be different from the assessment process without it?

Big data, in the context of assessment, is learner data that is deep as well as broad.[1] Large amounts of data can occur not only across many learners (broad between–learner data), but also within individual learners (deep within–learner data). Moreover, the depth of data is determined not only by the raw amount of data on a given learner, but also by the availability of contextual information that adds semantic meaning to within–learner data. Clickstream data is a good example of big data that tends to fall short of providing meaningful information in the context of assessing learning (cf. Case Study 1), although it may be sufficiently deep for assessing persistence (cf. Case Study 3). Therefore, the dimensionality of big data depends fundamentally on the object of assessment. More importantly, the converse is also true: new forms of data–enriched assessment require collecting deeper and broader data in order to gain insight into the new object of assessment.

Large–scale standardized tests, for instance, are broad but not deep; they yield large amounts of data consisting of test scores for thousands of learners with the primary focus of providing comparisons across learners, but which provide relatively little information about each individual learner. In contrast, a virtual reality learning experience (e.g., a mathematics lesson in a virtual classroom) can track learners' body positions to generate a substantial amount of behavioral and other information, but only for a small number of learners. Data–enriched assessment in appropriately instrumented online learning environments can, for a large number of learners, provide insights into each individual learner's problem–solving processes, strategic learning choices, misconceptions, and other idiosyncratic aspects of performance. In practice, this typically implies that information about learner performance is plentiful enough to gain new insights by applying modern data mining and machine learning methods (Romero, Ventura, Pechenizkiy, & Baker, 2011), such as hidden Markov modeling (cf. Case Study 1), probabilistic graphical modeling (cf. Case Study 2), or natural language processing methods (cf. Baker & Corbett, 2014).

Previously available data in assessment have been large in one of the two dimensions, but rarely before have education researchers been in a position to collect large amounts of data on both dimensions at once. The promise of big data in online learning environments is that capturing semantically meaningful information both across and within learners provides a powerful basis for assessing and supporting learners.

Elements of Data–Enriched Assessment

Deep and broad learner data in an interactive online learning environment can enable assessment tasks that are continuous, feedback–oriented, and multifaceted.

Continuous. In an online learning environment, an individual's learning process can be continually observed: the steps in solving a math problem, the chemicals combined on a virtual lab bench, and the learner's contributions to a discussion forum are all captured by the system. Interactions with learning resources, with peers, or with the instructor each contain

[1] A technical definition of big data focuses on the technological constrains that occur during computation, and which tend to require distributed processing and approximations instead of exact computations.

evidence about the concepts and skills over which the learner currently has command. There is no need to distinguish between learning activities and moments of assessment. Instead, a model of the learner's knowledge state is continually assessed and updated – as are models of other facets of the learner, as described below. This enables learning to be modeled as an ongoing process rather than as a set of discrete snapshots over time.

Feedback–oriented. Feedback is central to an assessment process that is designed to support learning. Well–designed feedback presents the learners' current state, along with enough information to make a choice about the appropriate next action. Feedback can be provided directly to the learner, to an instructor, or to the system (e.g., an adaptive test or an intelligent tutor). Providing learners with the choice of when to receive feedback and an opportunity to reflect on feedback may have additional benefits for developing metacognitive competencies. Drawing on prior work on the relative benefits of different types of feedback for learners with particular characteristics, online learning environments can also provide personalized feedback. For instance, based on a design principle proposed by Shute (2008) in a review of the feedback literature, the system could offer direct hints to low–achieving learners and reflection prompts to higher–achieving learners.

The effective presentation of feedback in online learning environments poses an interesting design challenge. Graphs, maps, and other information visualization techniques can be used to represent learner progress through the multiple concepts and competencies that learners are attempting to master. The information visualization community has developed an increasingly sophisticated visual language for representing complex datasets (e.g., Ware, 2013), and the efficacy of particular visualization strategies for supporting learners and instructors is a fruitful area for future research.

Multifaceted. Learners' abilities to learn from resources or interactions with others is influenced by factors beyond their current knowledge state. There are many reasons that a learner may start a task, struggle with it, or complete it successfully. Detecting these factors can contextualize observations about cognitive competencies, which provides the system or an instructor with additional information to target feedback or an intervention. The learner's life context is an important facet for developing deeper understanding of the learner's experience (cf. Case Study 3). Affective state – the learner's mood or emotions – can also have an impact on the learning processes (cf. Baker & Corbett, 2014), as can interpersonal competencies, such as the ability to communicate and collaborate effectively with others (De Laat & Prinsen, 2014).

Big data, in the context of assessment, is learner data that is deep as well as broad. Large amounts of data can occur not only across many learners (broad between–learner data), but also within individual learners (deep within–learner data).

Other critical facets of the learner include self–regulation – a learner's awareness and effective application of study strategies (Zimmerman, 1990); goal orientation – a learner's purpose in engaging with the learning activity (Pintrich, 2003); and mindset – a learner's beliefs about whether intelligence is fixed or malleable (Dweck, 2006). In addition, a rich history of research in social and educational psychology highlights the impact of learners' attributions of social cues in their environment (Cohen & Sherman, 2014; Steele, 1997), for example, whether a learner experiences a sense of social belonging in an environment (Walton & Cohen, 2011). Each of these intrapersonal, affective, contextual, and interpersonal states can be included in a model as latent states of the learner or directly reported features. Complex multifaceted models are enabled by big data and can advance research on the impact of each of these factors on learning.

The multiple facets of a learner translate into key competencies for individuals to be productive and resilient in future educational and professional settings. Explicitly assessing these competencies as desired outcomes of learning can inform the design of learning environments to support their development and thereby better serve learners for the long term.

Case Studies

In the following case studies, we draw on our work in three online learning environments to describe multiple approaches to data–enriched assessment. In each case study, learner data is deep because the learner is observed continuously, and broad as a result of the number of learners who engage with the online learning environment. Additional data dimensionality is added by specifying the relationship of learner activities to the concepts requisite for successful task engagement (Case Study 1) and to the appropriate next steps in a problem–solving process (Case Study 2). This specification, or "expert labeling," can occur in advance of developing an initial model or in the process of refining a learner model. Regardless of variations in the object of assessment or the timing of expert labeling, each case study uses machine learning techniques to develop or refine a learner state model.

In Case Study 1, the Open Learning Initiative, the assessment tasks are designed and embedded within the learning process. Data collected on learner performance on assessment tasks are used to diagnose the knowledge state of the learner and give feedback in real time and to refine underlying models. In Case Study 2, learners engage in open–ended software programming tasks, and assessment is focused on the processes of problem solving. Moreover, patterns in these processes are used to automatically generate suggestions for future learners who are struggling with the task. Case Study 3, focused on MOOCs, addresses the challenge of assigning meaning to learner activities that are outside of problem solving, such as forum interactions and video watching habits.

Case study 1: The open learning initiative (OLI). *Open Learning Initiative (OLI)* at Stanford University and Carnegie Mellon University is a grant funded open educational resources initiative. Data have been collected from over 100,000 learners that have participated in an OLI course for credit at academic institutions of all Carnegie Classifications and from over 1,000,000 learners that have engaged in one of the free and open versions of an OLI course.

An intelligent tutor is a computer program whose design is based on cognitive principles and whose interaction with learners is based on that of a good human tutor, making comments when the learner errs, answering questions about what to do next, and maintaining a low profile when the learner is performing well.

OLI courses comprise sequences of expository material such as text, demonstration videos and worked examples interspersed with interactive activities such as simulations, multiple choice and short answer questions, and virtual laboratories that encourage flexible and authentic exploration. Perhaps the most salient feature of OLI course design is found in the intelligent tutors embedded within the learning activities throughout the courses. An intelligent tutor is a computer program whose design is based on cognitive principles and whose interaction with learners is based on that of a good human tutor, making comments when the learner errs, answering questions about what to do next, and maintaining a low profile when the learner is performing well. The tutors in OLI courses provide the learner tailored feedback to individual responses, and they produce data.

OLI learning environments and data systems have been designed to yield data that inform explanatory models of a student's learning that support course improvement, instructor insight, learner feedback, and the basic science of learning. Modern online learning environments can collect massive amounts of learner interaction data; however, the insights into learning that can be gleaned from that data are limited by the type of interaction that is observable and by the semantic tagging (or lack of tagging) of the data generated by the interaction. Many MOOC platforms and traditional learning management systems collect clickstream data that can measure frequency and timing of learner log–ins, correctness (or incorrectness) of learner responses, learner use of resources, and learner participation in forums. While such clickstream data may be used to predict which learners are likely to complete the course, they do not explain if or how learning is occurring.

In OLI, the learning data are organized by learning objective. Learning objectives identify what a learner should be able to do or demonstrate they know by the end of the learning experience. Each learning objective comprises one or more skills. Skills break down the learning objective into more specific cognitive processes.

The course design process starts with the articulation of the learning objectives and skills. During the design of the course, the opportunities for learner action (e.g., answering a question, taking a step in a multi–step task, acting on a simulation) in an interactive activity are associated with the learning objectives and skills. The relationships among learning objectives, skills and learning activities are fully many–to–many: each learning objective may have one or more component skills, each skill may contribute to one or more learning objectives, each skill may be assessed by one or more steps in a task, each task step may assess one or more skills. Typical OLI courses comprise about 30 to 50 learning objectives and 100 to 1,000 skills.

Teams of faculty domain experts, learning scientists, human–computer interaction experts, assessment experts, and software engineers work collaboratively to develop the OLI courses and a parameterized model that predicts learner mastery of component skills. Skills are ranked as easy, moderate, or difficult based on perceived complexity. Initially, the labels are based on an analysis of the domain and on the expert's prior teaching experience. The rankings are used to adjust baseline parameters and, during the initial design of the course, the adjustments are heuristic, not empirical. The model associates learner practice with individual skills rather than with larger topics in a domain or activity in the course in general. The underlying theory is that learning is skill specific and it is practice on the specific skill that matters rather than practice in the course in general.

The skill model that the development team has created is considered naïve until it has been validated by data. Machine learning algorithms support learning researchers to improve upon the initial human–generated model by searching for models of learning that produce a better fit to the learner–generated data. The algorithms split and combine existing skills and suggest new skills where appropriate but, to date, a human must supply labels for the changes suggested by the algorithm. The researchers use the data to evaluate the fit of the model and to tune the parameters for the model. The course design team also uses the data to refine the learning activities and the response–level feedback.

While such clickstream data may be used to predict which learners are likely to complete the course, they do not explain if or how learning is occurring.

The skill model serves a number of purposes, including assisting in the iterative course improvement process; measuring, validating and improving the model of learning that underlies each course; and offering information necessary to support learning scientists in making use of OLI datasets for continued research. In the original versions of OLI courses, learning is modeled using a Bayesian hierarchical statistical model with the latent variables of interest, learners' knowledge state, becoming more accurate as more data is accrued about performance on a given skill. Skills are modeled using a multi–state hidden Markov model. The Markov model is hidden because the knowledge states cannot be observed directly; inferences need to be made about which state a learner is in based on the learner's answers to questions. In the original models, individual skills are treated as mathematically independent variables and it is assumed that learning a skill is a one–way process: once a skill is learned, it is not unlearned.

One of the most important uses of the skill model is to support learning analytics for instructors and learners. The OLI system analyzes the learner activity in real time against the skill model. When a learner responds to a question or engages in an OLI activity, the system uses the skill model mapping to identify the skills related to that question or activity. The learning estimates are computed per skill per learner and use simple algorithms with low computational overhead to allow real time updates. Data are aggregated across skills for a given learning objective and reported to instructors and students at that level. It is this real time feedback to instructors and students about mastery of learning objectives that helps guide the instructional and learning process throughout the course.

Case study 2: Code webs. The Code Webs Project is a Stanford machine learning research collaboration to analyze logs of learners completing open ended programming assignments with the intention to (a) uncover new perspectives into individual learner abilities, (b) paint a picture of how learners in general approach problems, and (c) understand how to help learners navigate complex programming assignments.

The project studies logs of learners solving assignments in three courses: The Code.org Hour of Code (Code.org), The Coursera Machine Learning class (ML) and Stanford's Introduction to Computer Science course (CS1). The Code.org and ML courses are both open access online courses, whereas the CS1 is a traditional in–person college course. The data are wide and deep. In each course learners complete a set of challenging programming tasks and each time a learner saves or runs an intermediate solution to a task, an entire snapshot of their current work is recorded. When the learner submits a final answer, or stops working on an assignment, all of the learner's partial solutions are composed into a trajectory. From the three courses, the Code Webs project has compiled trajectories from over 1,000,000 learners.

One of the most generally applicable results of this research has been to demonstrate the tremendous potential towards better assessment that comes from digital logs of how learners work through assignments, as opposed to just the learner's final submission. In future educational settings, the data on how learners develop their homework solutions from start to finish will become more ubiquitous and machine learning techniques applied to this format of data will generate important insights.

The first nugget that can be discovered from learner trajectories is a depiction of how learners, both as a cohort and individually, solve open ended work. In CS1, the Code Webs team instrumented the programming environment that learners used to generate their homework solutions. Using the data gathered, the research team modeled how groups of learners proceed through the assignment, using a Hidden Markov model that involved:

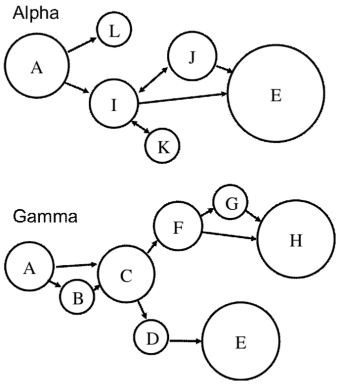
a. Inferring the finite set of high–level states that a partial solution could be in.

b. The transition of probabilities of a learner moving from one state to another.

c. The probability of seeing a specific partial solution given that a learner is in a state.

Once transition patterns for each learner had been fit, we then clustered the transition patterns to produce different prototypical ways that learners approach programming assignments.

In the CS1 dataset we discovered two notable prototypical patterns: A "Gamma" group whose progress is defined by steady work towards the objective and an "Alpha" group in which learners tend to get stuck in states where they would spend a lot of time before moving back to a previous state and then manage to make a large jump to a solution. Figure 1 demonstrates the pattern for a particular assignment in CS1.



State	*Description*
A	Start state.
B	Places a single beeper.
C	Places a perfect row in all worlds.
D	Places two perfect rows of the comb soln.
E	The comb soln. that works on all worlds.
F	The zigzag soln. that works except one col worlds.
G	The zigzagoln. that crashes on odd col worlds.
H	The zigzag soln. that works on all worlds.
I	Places one line. Misses a beeper on odd col worlds.
J	Two rows of the zigzag soln. Second row shifted.
K	Repeats first two rows infinitely.
L	Hardcoded to place first row on the 8x8 world.

Figure 1. Visualization of the two prototypical patterns for solving an open ended assignment in CS1. While most learners submitted a correct final solution, how they arrived at their answer was more predictive of miderm score. Only the most popular states and transitions are visualized.

One of the most generally applicable results of this research has been to demonstrate the tremendous potential towards better assessment that comes from digital logs of how learners work through assignments, as opposed to just the learner's final submission.

In CS1, almost all learners turn in working solutions to the class assignments; however on the class midterms and finals, some learners are unable to solve basic tasks. A promising result of this work was that the learners' problem solving patterns on the first assignment were more predictive of midterm grades than were their final scores on the first assignment.

Data logs on learners' solving problems can give insights into how learners are approaching problems and to what extent they understand the material. In addition to finding prototypical patterns, the autonomous process also computes to what extent each learner's progress matches the common patterns, and the overall distribution of the class.

Trajectories can also be used to autonomously learn what learners should do when working on open ended problems. For example, if we observe thousands of past learners who got stuck on the same problem, it seems plausible that we could use the subsequent actions that they took to predict the ideal solution to that problem. To explore this avenue, the Code Webs project team looked at the trajectories from half a million predominantly middle school learners solving the same programming assignments in Code.org's Hour of Code. We devised an experiment where experts generated a strategy of what partial solution a learner should transition to next given their current state and, using trajectory data, learn an algorithm that could recreate the expert strategy.

Surprisingly, many reasonable statistics on learner trajectories are not particularly useful for predicting what expert teachers say is the correct way forward. The partial solution that learners most often transition to after encountering a problem does not tend to correspond with what experts think learners should do. The wisdom of the crowd of learners, as seen from this angle, is not especially wise. However, there are other signals from a large corpus of trajectory data that shed light onto what a learner should do from a current partial solution. One algorithm generates a data–driven approximation of a complete journey from any current state to a solution that it expects would be most common if students were evenly distributed amongst the problem solving space. The first step in the generated journey overwhelmingly agrees with expert labels of how learners should proceed. This algorithm can be applied to logs of learners working on problems for which there are no expert labels, and will produce an intelligent strategy for what learners ought to do.

By modeling how learners progress through an assignment we open up the possibility for data driven feedback on problem solving strategies. By learning a predictor for how experts think a learner should proceed through a project, the process for generating a hint is simplified, both because we know what part of an open ended problem a stuck learner should work on next and we know what change they should make. Since the feedback can be autonomously generated it could be continuously and immediately provided to learners.

In future educational settings, the data on how learners develop their homework solutions from start to finish will become more ubiquitous and machine learning techniques applied to this format of data will generate important insights.

Trajectories seem like a promising medium through which we can leverage large amounts of data to generate better and more scalable assessment for learners that do their work in an instrumented environment. Though this case study was about computer programming, the algorithms used would apply to any trajectories of learner data, given an appropriate representation of partial solutions. While the Code Webs project has made progress towards its goal, this is still an active line of research, and better techniques will help uncover the deeper educational gems hidden in how learners work through assignments.

Case study 3: MOOCs and multifaceted dropout factors. Big data inspires us to ask questions that we could not ask with previous types of educational data. Among these questions is whether we can predict learners' persistence in a course and understand the challenges they encounter, given data from their interactions with the system. In earlier learning environments, it was much easier to acquire data about a learner's skill through assessment tasks than it was to learn about the learner's motivation, volition, or other latent factors that affect persistence similarly. Newer online platforms record new types of interactions that make assessment of such latent factors more feasible. For instance, passive forum participation is a potential signal of motivation for learners who did not participate actively in the forum. Total time of a learner on the course site might be a signal of time availability.

This case study describes our attempt to leverage the richer types and scale of data to predict who is going to drop out from a MOOC, and whether they are going to drop out due to difficulty, lack of motivation, or lack of time. To predict who will drop out, we developed an algorithm that uses features extracted from learners' interactions with the videos, assignments, and forums in multiple MOOCs (Halawa, Greene, & Mitchell, 2014). Our model uses four features we found highly correlated with dropout: the amount of time taken to complete the first week's videos, assignment scores, and the fraction of videos and assignments skipped. The model predicted dropouts with a recall of 93% and false positive rate of 25%.

We developed an instrument and collected data to predict the reason(s) that learners drop out. We emailed a diagnostic survey to 9,435 learners who were red–flagged by our dropout predictor in a course. The survey was sent out via email in the middle of the third week of the course, and 808 recipients responded to the survey (a typical survey response rate in a MOOC). Constructing our diagnostic models based on the optional survey introduced a selection bias, whose consequences on the suitability of the designed interventions to non–respondents are the subject of future research. In the survey, learners were asked to report on various persistence factors, including their commitment level (the extent to which learners believed they committed a sufficient portion of their free time to the achievement of their course goals), and perceived difficulty (how difficult they found the course materials, including assessment tasks). Learners were also asked to report on the average amount of weekly free time they had. We used each learner's responses to compute three binary variables indicative of potential interventions:

1. Dropped out due to procrastination (which results from a lack of volition)
2. Dropped out due to difficulty
3. Dropped out due to lack of time

Next, we used learner interaction data to compute scores for various activity features describing the learner's pace, learning session times, and interactions with the lecture videos, assignments, and forums as shown in Table 1. We selected the features that we believe would correlate with particular reasons for dropout (or lack thereof). For instance, joining a study group may be predictive of the learner's intention to persist in the course for a long period. Giving up on problems after a first incorrect attempt might indicate a lack of motivation or grit.

Surprisingly, many reasonable statistics on learner trajectories are not particularly useful for predicting what expert teachers say is the correct way forward.

Table 1
Candidate Features Used to Predict Reasons for Dropout

Video interactions	● Fraction of visited video duration seeked back ● Fraction of visited video duration seeked forward ● Number of times the learner reviewed a previously visited video ● Fraction of videos skipped ● Fraction of videos viewed until the end
Assignment interactions	● Fraction of course quizzes attempted ● Fraction of problems answered incorrectly on first attempt that were reattempted ● Time between attempts
Forum interactions	● Number of forum posts ● Number of comments to posts by other learners ● Number of threads read ● Did the learner post a self-introduction to the forum? ● Did the learner create or join a study group?
Pace	● What fraction of released videos had the learner visited by various time points in the course?
Learning sessions	● How many times per week did the learner visit the course? ● How long is the learner's average session?

We trained three logistic regression models, one for predicting each of the three dropout factors, which meant that a learner could be red–flagged for multiple dropout reasons. Accuracy was measured for each risk factor individually via recall – the fraction of learners who self–reported the risk factor that was red–flagged by the prediction model – and false positive rate (fpr) – the fraction of learners who were self–reportedly unaffected by the risk factor but red–flagged by the predictor (see Figure 2).

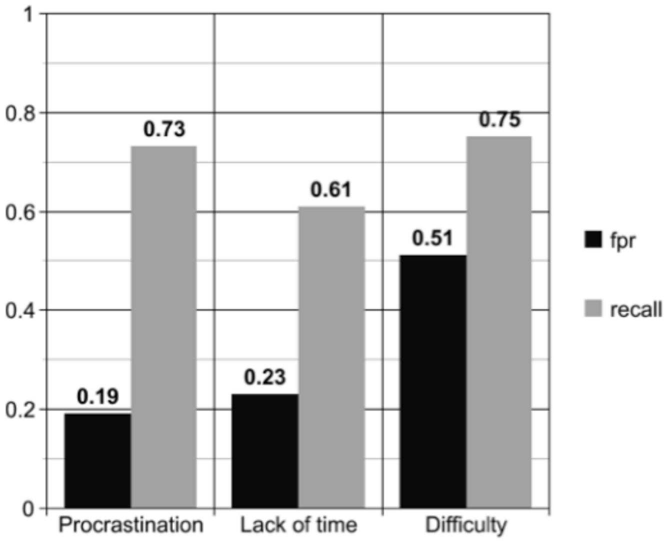


Figure 2. Prediction accuracy for our dropout diagnostic models

The procrastination detection model was able to predict procrastination with a false positive rate (fpr) of 0.19 at a recall of 0.73. The key contributing features of the model were interactions with the forum and assignments. We generally observed that learners with lower motivation or volition spent all of their time on the course in activities that yield direct personal rewards, such as viewing videos, taking assignments, reading the forum, or posting questions to the forum. Activities such as joining a study group, socializing on the forum, and commenting on other people's posts originated mainly from learners who self–reported higher levels of volition. We were also able to predict learners who reported time constraints with almost the same fpr but a lower recall. Contributing features included the patterns of spending time on the course, and it was observed that learners who report less free time tend to have shorter learning sessions. Predicting reports of perceived difficulty was less accurate due to the weakness of correlation between reported difficulty and our features including assignment scores. Improving this prediction is a subject of our future research.

This case study exemplifies two facets of data–enriched assessment, namely its multifaceted and feedback–oriented nature. In this study, we focused on specific facets of learners' contexts that are critical for their success in the learning environment: procrastination behavior, time constraints, and perceived difficulty. Moreover, this work will be extended to provide targeted feedback about these non–cognitive factors to at–risk learners. Potentially, such modeling capability allows us to assess these persistence factors and design more effective interventions that address the restraining and promoting forces relevant to each individual learner.

General Discussion

The preceding case studies illustrate how big data can enrich assessment by directly supporting learning as it assesses multiple facets of learning such as competencies and persistence. We argue that this is for three reasons. First, the next generation of online learning environments allows us to collect data continuously and at large scale. In turn, large–scale data collection allows researchers to more effectively use modern statistical and machine learning tools to identify and refine complex patterns of performance. For example, the work on programming trajectories described above illustrates that massive amounts of time–series data on learner programming problems can be used to predict later success and potentially to provide just–in–time hints.

Online learning environments also allow educators to record multifaceted measurements of skills and tendencies that normally evade traditional assessment tasks. The work on identifying dropout factors in MOOCs illustrates this point. Halawa and colleagues (2014) initially measured motivational variables using surveys, which are a familiar assessment instrument for academic motivation researchers. But they were then able to predict survey responses using data on forum engagement, pace, and other aspects of course interaction. In a traditional educational setting, these or analogous behavioral variables would be largely unmeasured. In addition, the continued development of educational games, complex simulations, and VR environments makes us confident that future educators will have a much more multifaceted set of data than ever before (Bailenson et al., 2008; Schwartz & Arena, 2013).

Big data inspires us to ask questions that we could not ask with previous types of educational data.

Third, and perhaps most crucially for learning, online learning environments are capable of delivering personalized feedback at the right moment. The Open Learning Initiative demonstrates this advantage by harnessing decades of research into cognitive skill development in order to model learner knowledge and provide more appropriate instruction in real time. Meta–analyses of what works in improving learning have placed appropriate feedback at or near the top of the list (Hattie, 2013), and researchers have argued that effective feedback is also the primary source of the oft–quoted "two–sigma" positive effects of tutoring (Bloom, 1984). Big data allows educators to build and refine model–driven feedback systems that can match and surpass human tutors (Corbett, 2001).

Finally, all of the examples in this article illustrate that big data can benefit multiple stakeholders in the learning ecosystem. As a more formative enterprise, data–enhanced assessment can benefit learners themselves, but it can also provide feedback to instructors to guide their attention and teaching strategies. The benefits of data–enriched assessment are available not only to instructors teaching in purely online environments but also to instructors teaching in hybrid (a blend of online and face to face instruction) or traditional classrooms. In hybrid environments, the data collected from the students in a class provide information to the instructor to make immediate adjustments to classroom teaching. Even instructors who are teaching in traditional classrooms without any technology will benefit from the insights about how students learn a subject that are developed from the big data collected in online learning environments. Big data have also clearly informed researchers to develop better learner models and experiment with just–in–time interventions. And Macfadyen, Dawson, Pardo, and Gaševic (2014) show that big data can inform questions about equitable and effective learning at a policy level.

Conclusion

We have been quite positive about the promise of data–enriched assessment, and so it seems reasonable to end with a note of caution. There is a difference between how we use assessment tasks and what they are intended to measure, and the history of psychometrics is littered with incorrectly interpreted test results. How will big data affect the interpretation and validity judgments of the next generation of assessment tasks? It may be helpful to look to the misapplication of current generation assessment tasks for lessons. Assessment experts generally agree that since the start of No Child Left Behind, data from high–stakes tests in K–12 settings have been used to make inaccurate judgments about the performance of teachers, schools, districts, and states in an attempt to establish benchmarks for accountability and quality improvement (Baker et al., 2010). According to a recent review, ten years of test–based accountability policies has shown little to no effects on student performance (National Research Council, 2011).

In earlier learning environments, it was much easier to acquire data about a learner's skill through assessment tasks than it was to learn about the learner's motivation, volition, or other latent factors that affect persistence similarly. Newer online platforms record new types of interactions that make assessment of such latent factors more feasible.

Exploring the network of causes for the misuse of standardized test data is beyond the scope of this paper, but there are two substantial causes worth noting that are deeply related to the tests themselves. The first is simply that our ambitions to capture learning have often outpaced our abilities to design effective assessment tasks – learning is a multifaceted construct that is difficult to measure. The second reason is that it is also difficult to appropriately aggregate, report, and act upon test data (National Research Council, 2011).

We have argued that a data–enriched assessment process can potentially measure multiple facets of learning, as well as learning processes, more effectively than previous assessment approaches. However, our case studies also show that these assessment tasks depend on broad and deep learner data that may not always be available. The hype around online assessment, and the excitement over measuring novel motivational and other non–cognitive competencies, may continue to fuel ambitions that outstrip our capabilities. Moreover, data–enriched assessment methods can be far more complex and opaque than traditional methods, and their results can be difficult to interpret without expert assistance (Siemens & Long, 2011).

The availability of big data allows assessment methods to continually measure and support a broader range of learning outcomes while simultaneously providing feedback throughout the learning process. This is creating more of a need to provide thoughtful and actionable explanations of assessment results for all of the stakeholders involved, including teachers and learners.

AUTHOR'S NOTE

This work is a collaborative work by the researchers in the Stanford University Lytics Lab (http://lytics.stanford.edu). Each listed author is an equal contributor to the work.

References

Bailenson, J. N., Yee, N., Blascovich, J., Beall, A. C., Lundblad, N., & Jin, M. (2008). The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The Journal of the Learning Sciences, 17*(1), 102–141.

Baker, E. L., Barton, P. E., Darling–Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. EPI briefing paper# 278. *Economic Policy Institute.* Retrieved from http://eric.ed.gov/?id=ED516803

Baker, R. S., & Corbett, A. T. (2014). Assessment of robust learning with educational data mining. *Research & Practice in Assessment, 9*(2), 38-50.

Black, P., & Williams, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one–to–one tutoring. *Educational Researcher, 13*(6), 4–16.

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self–affirmation and social psychological intervention. *Annual Review of Psychology, 65*, 333–371.

Corbett, A. (2001). Cognitive computer tutors: Solving the two–sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva, (Eds.), *User modeling* (pp. 137–147). Heidelberg, Germany: Springer.

De Laat, M., & Prinsen, F. R. (2014). Social learning analytics for higher education. *Research & Practice in Assessment, 9*(2) 51-60.

Denley, T. (2014). How predictive analytics and choice architecture can improve student success. *Research & Practice in Assessment, 9*(2), 61-69.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Ballantine Books.

Gikandia, J. W., Morrowa, D., & Davisa, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333–2351.

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the European MOOC Summit*. Lausanne, Switzerland.

Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta–analyses relating to achievement*. New York, NY: Routledge.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the Third International Conference on Learning Analytics and Knowledge, ACM* (pp. 170–179).

Macfadyen, L. P., Dawson, S., Pardo, A., & Gaševic, D. (2014). The learning analytics imperative and the sociotechnical challenge: Policy for complex systems. *Research & Practice in Assessment, 9*(2), 17–28.

National Research Council. (2011). *Incentives and test–based accountability in public education*. Committee on Incentives and Test–Based Accountability in Public Education, M. Hout & S. W. Elliott, (Eds.). Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667–686.

Roediger, H. L., & Karpicke, J. D. (2006). Test–enhanced learning: Taking memory tests improves long–term retention. *Psychological Science, 17*(3), 249–255.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice–based assessments for the digital age.* Cambridge, MA: The MIT Press.

Shavelson, R. J. (2007). *A brief history of student learning assessment: How we got to where we are and where to go next*. Washington, DC: Association of American Colleges and Universities.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30–32.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629.

Walton, G. M., & Cohen, G. L. (2011). A brief social–belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447–1451.

Ware, C. (2013). *Information visualization: Perception for design* (3rd ed.). Waltham, MA: Elsevier.

Zimmerman, B. J. (1990). Self–regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 1–25.