Big Data and Assessment of Complex Skills

Piotr Mitros
edX

Historically, assessment in classrooms was limited to instructor grading, or problems that lend themselves well to relatively simple automation, such as multiple-choice questions. Progress in educational technology, combined with economies of scale, has allowed us to digitally measure student performance on authentic assessments such as engineering design problems and free-form text answers, radically increasing the depth and the accuracy of our measurements of what students learn, allowing us to tailor instruction to specific students needs and giving individualized feedback for an increasing range of issues. In addition, social interactions have increasingly moved on-line. We now have traces of a substantial portion of student-student interactions. By integrating these and other sources of data, we have data with which we can estimate complex skills, such as mathematical maturity, complex problem solving, and teamwork for large numbers of students. This paper looks at the potential information found in the data we now collect, some of the challenges with making sense of that data, and some early successes in analyzing that data. The data is complex. Actually extracting useful high-level metrics has proven difficult. The next grand challenge in big data in education will be finding ways to analyze complex data from heterogeneous sources to extract such measurements.

Keywords: educational datamining, assessment

Twenty years ago, most digital assessments consisted of multiple choice questions and most social interactions happened in person. Data was spread out over multiple systems with no practical means of integration. Over the past two decades, we have seen fundamental progress in educational technology, combined with broad-based adoption of such technology at scale[1]. Digital assessment has increasingly moved towards rich authentic assessment. Previously, widely available data for large numbers of students principally came from standardized exams or standardized research instruments such as the Force Concept Inventory. These assessments are limited to a short time window, and as a result, they either contain a large number of small problems (which are statistically significant, but generally fail to capture skills which require more than a minute or two to measure), or a small number of large problems (which, on a per-student basis lack statistical significance). Today, we are increasingly collecting data for students doing a large numbers of complex problems as part of their regular coursework. For example, the first edX/MITx course[2], 6.002x (Mitros et al., 2013) was implemented entirely with authentic assessment. Students completed circuit design problems (verified through simulation), and design and analysis problems (with answers as either equations or numbers). Since these types of questions have a near-infinite number of possible solutions, answers cannot be guessed. Students could attempt to submit an answer as many times as necessary in order to completely understand and solve a problem. The assessments were complex – most weeks of the course had just four assessments, but completing those four required 10-20 hours of work. We see similarly rich assessments in courses such as chemistry, biology, physics, digital electronics, and many others. Such complex assessments, taken together across many courses, give rich data about problem solving skills, creativity, and mathematical maturity.

Furthermore, we now collect microscopic data about individual student actions. We can see not only which problems students answered correctly, but how they got there. Extensive literature on expert-novice shows differences in problem solving strategy between novices and experts. For example, experts can chunk information (Schneider, Gruber, Gold, & Opwis, 1993) – an expert looking at an analog circuit will be able to remember that circuit, whereas a novice will not (Egan & Schwartz, 1979). In our data sets, we can see actions which reflect such differences. Continuing with the example of chunking, we record how many times a student flips between pages of a problem set, looks up equations in a textbook, and similar activities which are proxies for expertise.

Next, social interactions are increasingly moving on-line. As we introduce increased amounts of digital group work to

[1] We define at-scale learning environments as ones where thousands of students share common digital resources, and where we collect data about such use. This includes MOOCs, but also many educational technologies predating MOOCs, as well as formats such as SPOCs.

[2] Used both in a pure on-line format, as well as in a blended format in a number of schools

courses, we start to see traces of social activity in our logs. We can begin to look for students who under-perform or over-perform in group tasks, and directly measure students' contributions to groups. We have enough data to begin to look for specific actions and patterns that lead to good overall group performance, and hopefully we will be able to use such patterns to provide feedback to students. Natural language processing frameworks, such as the open-source edX EASE and Discern, are still used primarily for short-answer grading, but were designed to also apply to analysis of social activities, such as e-mails and forum posts, as well. We believe this will begin to give insights into soft skills, writing processes (Southavilay, Yacef, Reimann, & Calvo, 2013), communications styles, and group dynamics.

Finally, aside from just looking within individual courses, we can perform longitudinal analysis across a student's educational career. In most cases, a single group design project does not provide statistically significant information. However, all of the projects over the duration of a student's schooling are likely to be significant. Learning analytics systems are increasingly moving in the direction of aggregating information from multiple sources across multiple courses. Open analytics architectures (Siemens et al., 2011) such as edX Insights (Mitros, 2013) or Tin Can provide a common data repository for all of a student's digital learning activities.

However, going from data to measurement is a complex problem. In the next few sections of this paper, we will discuss some of the challenges, as well as early successes.

Challenges – Pedagogical Design

There is substantial friction between the design for different educational purposes, of which, measurement is just one. Assignments and assessments in courses have several objectives:

- **Initial and formative assessment as an ongoing means of monitoring what students know**. This allows instructors and students to tailor teaching and learning to problematic areas (Sadler, 1989).
- **The principal means by which student learn new information**. In many subjects, most student learning happens through assignments where they manipulate, derive, or construct knowledge (Chi, 2011) – not lectures, videos, or readings.
- **A key components of grading**. Grading itself has multiple goals, from certifying student accomplishment to providing motivation for desired student behaviors.
- **Summative assessment of both students and courses**. Summative assessment has many goals, such as student certification and school accreditation.

Historically, different research communities emphasized different objectives and gave very different principles around how good assessments ought to be constructed. For example, the psychometrics community principally relies on metrics such as validity and reliability. These suggest a high level of standardization in assessments. In contrast, the physics education research community emphasizes concepts such as the trade-off between authentic assessment and deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993), as well as principles such as rapid feedback, active learning, and constructive learning. Educational psychology (Bloom, 1984) and gamification emphasize mastery learning (where students eventually get all questions right).

Numerical techniques which presume that assessments are developed designed based on principles which optimize for measurement often fail when applied to the much broader set of classroom assessments. There is an inherent friction between:

- Having a sufficient number of problems for statistical significance vs. long-form assessments which allow students to exercise complex problem solving and mathematical maturity.
- Measuring individual students vs. group work[3].
- Standardized assessments vs. diversity in education. The US economy benefits from a diverse workforce, and the educational system, especially at a tertiary level, is designed to create one. There are over ten thousand distinct university-level courses.
- Aiming for 50% of questions correct (maximize measurement) vs. 100% of concepts mastered (mastery learning)

To give an example of how friction comes into play, the MIT RELATE group applied item response theory (Embretson & Reise, 2000), a traditional psychometric technique, to calibrate the difficulty of problems in 6.002x, the first MITx/edX course. However, IRT presumes that problem correctness is a measure of problem difficulty. 6.002x is based on mastery learning, and students can continue trying until they answer a question correctly – any sufficiently dedicated student could answer all questions correctly. To apply IRT in this context, RELATE had to substantially adapt the technique (Champaign et al., 2014).

Challenges – Diversity and Sample Bias

Many traditional psychometric techniques rely on a relatively uniform dataset generated with relatively unbiased sampling. For example, to measure learning gains, we would

[3]At this point, we have overwhelming evidence that well-structured groupwork leads to improved student outcomes.

typically run a pre-test and a post-test on the same set of students. In most at-scale learning settings, students drop out of classes, take different sets of classes, and indeed, the set of classes taken often correlates with student experience in previous classes. We see tremendous sampling bias. For example, a poor educational resource may cause more students to drop out, or to take a more basic class in the future. This shifts demographics in a future assessments to a stronger students taking weaker courses, giving a perceived gain on post-assessment if such effects were not controlled for.

Likewise, integrating different forms of data – from peer grading, to mastery-based assessments, to ungraded formative assessments, to participation in social forums – gives an unprecedented level of diversity to the data. This suggests a moves from traditional statistics increasingly into machine learning, and calls for very different techniques from those developed in traditional psychometrics.

Challenges – Data Size and Researcher Skillset

Traditionally, big data educational research was conducted by statisticians in schools of education with tools such as spreadsheets, and numerical packages such as R. This worked well when data sets were reasonably small. A typical data set from a MOOC is several gigabytes. The data at a MOOC provider is currently several terabytes. While this is not big data in a classic sense, the skills and tools required for managing this data go far beyond those found at many schools of education. With continuing moves towards technologies such as teleconferencing, we expect datasets to grow manyfold.

As a result, most data science in MOOCs has been conducted in schools of computer science by researchers generally unfamiliar with literature in educational research. This shortcoming is reflected in the quality of published results – for example, in many cases, papers unknowingly replicating well-established decades-old results from classical educational research.

Meaningful research requires skillsets from both backgrounds. There are few researchers with such skillsets, and collaborations are sometimes challenging due to substantial cultural differences between schools of education and schools of computer science.

Early Successes

An early set of high-profile successes in this sort of data integration came from systems which analyzed data across multiple courses in order to predict student success in future courses. This includes systems such as Purdue Course Signals (Arnold & Pistilli, 2012), Marist Open Academic Analytics Initiative (Lauría, Moody, Jayaprakash, Jonnalagadda, & Baron, 2013), and Desire2Learn Student Success System (Essa & Ayad, 2012).

There have been early successes with system which look at different types of data as well. For example, the first prototype of the edX Open-ended Response Assessment (ORA1) system integrated:

- **Self-assessment** – students rate their own answers on a rubric.
- **Peer assessment** – students provide grading and feedback for assignments submitted by other students.
- **Instructor assessment** – the traditional form of assessment.
- **AI assessment** – a computer grades essays by attempting to apply criteria learned from a set of human-graded answers.

In the theoretical formulation (Mitros & Paruchuri, 2013), each of the four grading systems contributes a different type and amount of information. The system routes problems to the most appropriate set of grading techniques. An algorithm combines responses from graders to individual rubric items into feedback and a final score. A simplified form of this algorithm was experimentally validated.

Conclusion

While many of the goals of an educational experience cannot be easily measured, it is much easier to improve, control, and understand those that can. The breadth and depth of data now available has the potential to fundamentally transform education.

Students and instructors are incentivized to optimize teaching and learning to measured skills, often at the expense of more difficult-to-measure skills. While we have seen tremendous progress in education with the spread of measurement, limited or inaccurate assessments can actually cause harm if relied on too much. Measurement in traditional education is tremendously resource-constrained which severely restricts what can be measured. Standardized high-stakes tests are typically 3-4 hours long, and must be graded for millions of students in bulk. In most cases, such high-stakes exams can only accurately measure some skills and use those as proxies for more complex to measure skills. Many completely fail to capture skills such as mathematical maturity, critical thinking, complex problem solving, teamwork, leadership, organization, time management, and similar skills. While time constraints in traditional classroom settings are somewhat more relaxed than in high-stakes exams, instructors still often rely on proxies. For example, when measuring communication skill, a common proxy is an essay – a medium relatively rare in outside of the classroom. Instructors cannot effectively critique longer formats of communications, such as e-mail threads, meetings, and similar without extreme student:faculty ratios – computers can.

Digital assessments have long been effective means to liberate instructor time, particularly in blended learning settings, as well as for providing immediate formative feedback (VanLehn, 2011) (National Research Council, 2000) (Patterson, Gavrin, & Christian, 1999). Building on this work, we are increasingly seeing a move to authentic assessment, approaches where humans and machines work in concert to quickly and accurately assess and provide feedback to student problems (Basu, Jacobs, & Vanderwende, 2013), where data is integrate from very diverse sources, and where data is collected longitudinally.

With this shift, for the first time, we have data about virtually all aspects of students skills – including complex ones that are, ultimately, more important than simple factual knowledge (Sternberg, 2013). We have the potential to provide new means to assess students in ways which can improve the depth, frequency, and response time, potentially dramatically expanding the scope with which students and instructors can monitor learning, including assessment of higher-level skills, and proving personalized feedback based on those assessments. However, the tool for understanding this data (edX ORA, Insights, EASE, and Discern, in our system, and their counterparts in others) are still in their infancy. The grand challenge in data-intensive research in education will be finding means to extract such knowledge from the extremely rich data sets being generated today.

References

Arnold, K. E. & Pistilli, M. D. (2012). Course signals at purdue: using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402.

Bloom, B. (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.

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D., & Pritchard, D. E. (2014). Correlating skill and improvement in 2 moocs with a student's time on tasks. In *Proceedings of the first acm conference on learning@ scale conference* (pp. 11–20). ACM.

Chi, M. T. H. (2011). Differentiating four levels of engagement with learning materials: the icap hypothesis. *International Conference on Computers in Education*.

Egan, D. E. & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–158.

Embretson, S. & Reise, S. (2000). *Item response theory*. Psychology Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Essa, A. & Ayad, H. (2012). Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 158–161). ACM.

Lauría, E. J., Moody, E. W., Jayaprakash, S. M., Jonnalagadda, N., & Baron, J. D. (2013). Open academic analytics initiative: initial research findings. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 150–154). ACM.

Mitros, P. (2013). Open platforms for pedagogical innovation. In *Learning analytics summer institute*. LAK.

Mitros, P., Affidi, K., Sussman, G., Terman, C., White, J., Fischer, L., & Agarwal, A. (2013). Teaching electronic circuits online: lessons from MITx's 6.002x on edX. In *Iscas* (pp. 2763–2766). IEEE.

Mitros, P. & Paruchuri, V. (2013). An integrated framework for the grading of freeform responses. *The Sixth Conference of MIT's Learning International Networks Consortium*.

National Research Council. (2000). How people learn. (pp. 67–68, 97–98). National Academy Press.

Patterson, N. G., Gavrin, A., & Christian, W. (1999). Just-in-time teaching: blending active learning with web technology. Upper Saddle River NJ,: Prentice Hall.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119–144.

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess expertise and memory for chess positions in children and adults. *Journal of Experimental Child Psychology*, *56*(3), 328–349.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., … Baker, R. (2011). *Open learning analytics: an integrated & modularized platform* (Doctoral dissertation, Open University Press).

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 38–47). ACM.

Sternberg, R. (2013, June 17). Giving employers what they don't really want. *Chronicle of Higher Education*.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.