In Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education (2013). Memphis, TN. http://www.moocshop.org

Roll Call: Taking a Census of MOOC Students

Betsy Williams[1],

[1] Stanford University, Graduate School of Education, 520 Galvez Mall, CERAS Building, 5th Floor, Stanford, CA, 94305, USA
{betsyw@stanford.edu}

Abstract. This paper argues for spending resources on taking a high quality census or representative survey of students on who enroll with all major MOOC platforms. Expanded knowledge of current students would be useful for business and planning, instruction, and research. Potential concerns of cost, privacy, stereotype threat, and maladaptive use of the information are discussed.

Keywords: MOOC population, education, data collection, survey, demography

1 Introduction

Quantitative education researchers are accustomed to piecing together complex analyses from the rather lifeless data available from administrative records and test scores. The fine-grained data collected by MOOCs—including detailed knowledge of students' attendance and attention patterns, response on formative and summative assessments, and discussions with instructors and fellow students—offer an opportunity for much greater understanding of teaching and learning.

Unfortunately, MOOCs are not making the most out of their big data because they are not collecting enough data on students' backgrounds. Borrowing Bayesian terms, platforms have few priors on students, even though these priors can have great predictive power if paired with existing knowledge, from fields like developmental psychology and higher education theory.

The major platforms optimize sign up to make becoming part of the platform as quick as possible, leaving students mostly mysterious. EdX requests a few valuable pieces of demographic data upon registration, asking for voluntary identification by gender, year of birth, level of education completed, and mailing address without a clear reason why.[1] Coursera's information gathering is more like social media or a dating site, encouraging students who visit the profile page to share their age, sex, and location. As part of its "About Me" prompt, Coursera suggests that among other things users might share "what you hope to get out of your classes," while EdX asks the question in an open-ended text box upon registration. While these questions yield some of the data that is valuable for improving courses, the platforms, and education

[1] No one reads terms of service [1].

research, I argue that the platforms should collect more key data, clearly identified as information that will not be sold or used for targeted marketing or for student evaluation.

The paper first describes the fields most useful for analysis based on priors, and then it explains the benefits to platform development, instructional quality, and research. Potential drawbacks are discussed, including cost, privacy concerns, the risk of invoking stereotype threat, and the potential for undesirable changes to arise from this information.

2 Prior Information about Students

Given infinite data storage and infinite indulgence on the part of MOOC students, knowing every scrap of data about students might allow for inspired analyses and eerily predictive machine learning exercises. However, a more humble conception of student data would ably fulfill our research needs.

Core demographic information includes year of birth, gender, and race/ethnicity.[2] Asking users for their current city or place of residence should generate more accurate location results than IP address tracing or the information provided to appear on a semi-public profile. Combined with place of origin and native language, these questions provide a sketch of a student's likely history and culture.

A MOOC-run survey would also provide the opportunity to ask questions less often available in administrative education data, although extremely useful for understanding who enrolls. Although sensitive, questions about socioeconomic status and living situation would be tremendously helpful; for instance, is a student living with family, and to which generation does that student belong?

Adult students' lives are increasingly complex, and questions about work and education history should do their best to capture this. If a student's highest degree is a high school diploma (or equivalent), then have they ever enrolled in higher education? If so, in how many institutions? How many years and months would they estimate this spanned? Were they primarily taking full time or part time loads? What was the name of their primary institution, and what was their most recent course of study pursued? Those who have earned bachelor's degrees or higher should face similar questions. For all students, questions about previous or concurrent MOOC use would be very valuable. Work history can get a similar treatment, identifying such things as area of employment, and full- and part-time scheduling.

Although students themselves may not be entirely clear on the point yet, questions about educational and career goals, along with goals for the course, would be extremely useful. This information is captured to some extent in existing questions or for particular research. However, this may be incomplete or collected only in a piecemeal fashion. For instance, a study on learner patterns surveys the students in

[2] Race and ethnicity are social constructs whose meaning greatly varies by national context. For instance, being white in Norway has a different social meaning than being white in South Africa. And Belgium is split by a key ethnic marker—Walloon versus Fleming—that does not matter in other countries. Thus, choices for race/ethnicity should be based on the selected country of origin and/or country of residence.

one course, asking for intentions in the course, current employment status, years of work history, and highest degree attained [2].

Valuable information from surveys need not all be based on recall or opinion. Meaningful priors about academic preparation in particular fields can be generated by computer adaptive test questions in key content domains, based on existing work in psychometrics. Behavioral economics shows that survey questions can measure levels of risk aversion (asking for preferences between a gamble for $X and receiving $Y with certainty) and time discounting on money (asking about preferences for receiving $X now or $Y at a certain point in the future).[3]

Finally, there's a useful realm of information about how students use the platform. Within a class, how much time do they plan to devote, how do they plan to interact with peers, and will they use external supports, such as tutors, websites, and textbooks? What modes of access to the course are available to them? In particular, what electronic devices are available to them, is their use of the devices limited, and what kind of Internet access is available?

3 Value for Planning and Strategy

The background information on users discussed above provides extremely valuable data for the operations of the course platforms. Let us stipulate that there are limitations on the data being used for targeted marketing purposes. Even so, having aggregate background information on who is using which MOOCs is a huge advance.

In a traditional business mindset, the primary questions would be who is willing and who is able to pay. However, more advanced uses could help a course recommendation engine distinguish who is taking the course as a consumption good versus as investment in their future; the follow-up courses the students are interested in may be vastly different.

The survey may also suggest a greater than anticipated demand for classes taught at a certain level or on a certain topic. Students' locations, educations, and work histories might help the platform identify other institutions that may be good partners, either because they are very well-represented or under-represented.

4 Value for Instructional Design

A strong finding in educational research is that there is not a single correct way to teach or structure a course. Instead, learners matter, and knowledge about the students and their characteristics is important for teaching well [3]. Knowing more about the students also allows instructors to effectively call on their existing knowledge and address likely misconceptions; this is part of Pedagogical Content Knowledge [3] and a prominent contribution of Piagetian constructivism [4].

[3] For the most accurate answers, survey takers would actually receive the payout they say they prefer, subject to a gamble or delay as the case may be.

For example, knowing the age distribution and native languages of students can improve instruction. Instructors may choose allusions, words, and examples better.

An inherent challenge within the online classroom is that some feedback that is obvious in a physical context is not available. One student falling asleep in a lecture hall is far more obvious and effective of a signal than a thousand who never rewind the recorded lectures. While learning analytics is tackling this paucity of data in clever ways, we would also benefit greatly from leaning on priors. Imagine two students who do not watch the second week lecture by the beginning of week three: one has a doctoral degree in the field, while the other is a high school graduate who has attended several different institutions of higher education and intends to take a course for professional development. Applying theory to this prior knowledge, we might think the former finds the course matter unnecessary to review, while the latter may be struggling to stay motivated in the class.

In short, better prior knowledge can be paired with data collected in courses to better identify how students are learning the course content and improve the course.

5 Value for Education Research

MOOC populations are so wildly self-selected, and the field so new, that external validity is extremely questionable. At best, we might extend findings in a class to perhaps the same class the next time it is taught or use the results to develop hypotheses and learning theory.

While there is great value in using research to improve a single course, ideally the lessons could be transferred more broadly, so that the effort of analysis pays greater dividends. However, results cannot generalize until the population of the study is understood; once more is known about incoming characteristics of MOOC students who were studied, researchers can seek other classes that resemble them in salient details.

More concretely, MOOCs offer radical levels of access to education, and so they include many non-traditional and out-of-school learners. These nontraditional learners can be elusive research subjects, and there is also great diversity among their numbers. Having additional background data allows us to tag them and better understand their behavior. If a course platform is successful with a particular college level course and is contemplating recommending it to a partner community college, it would be wise to understand how students of different backgrounds performed. The inference is not direct, but it is far more useful than a recommendation based on coarser data.

The MOOC is also a fantastic platform for learning about how everyone learns, not just how self-selected MOOC users learn. The large number of students and the computerized means of instruction mean that MOOCs are very amenable to experimentation and careful observation. In addition, the very design of MOOCs strips down the traditional classroom; greater insight about learning and traditional instruction can come from adding back in some of these elements that are taken for granted in other classrooms.

Yet again, the great advantages of MOOCs as a place for learning research have the caveat that results are hard to generalize. However, if researchers control for the observable background data of the students who opt into MOOCs, their results will be far more plausibly applicable to a wide array of classes.

A key challenge within the online classroom is that feedback that may be obvious in a physical context, such as real-time indications of student engagement or confusion, is usually not available. One student falling asleep in a lecture hall is far more obvious and effective of a signal than a thousand who never rewind the recorded lectures. While learning analytics is tackling this paucity of data in clever ways, we would also benefit greatly from leaning on priors. Imagine two students who do not watch the second week lecture by the beginning of week three: one has a doctoral degree in the field, while the other is a high school graduate who has attended several different institutions of higher education and intends to take a course for professional development. Applying theory to this prior knowledge, we might think the former finds the course matter unnecessary to review, while the latter may be struggling to stay motivated in the class.

In short, better prior knowledge can be paired with data collected in courses to better identify how students are learning the course content.

6 Concerns and Limitations

There are genuine concerns with collecting this much data. Here, I discuss cost, privacy, stereotype threat, and maladaptive use. I present these cursorily not to dismiss these points, but to begin what must be a larger discussion.

6.1 Cost

Course platforms are in a unique position It can be extremely costly to ask survey questions. User attention is limited and a choice to ask an additional question may implicitly limit their engagement later during the session, or even drive them away from the service at the extreme. Higher quality survey data can be generated by using internal resources to follow up with non-responders; higher response rates can also be generated by incentives, such as monetary payment, entry in a lottery, or access to a premium site feature. In addition, comprehensive surveys offered by a platform itself can be more easily embedded in the site, making the survey more available and more salient.

Administering a vast survey at the site level also better captures students who might be over-sampled if asked class-by-class. Cross-course analyses can be conducted more easily if the relatively permanent, detailed background information is available at the platform level, rather than asked for in individual courses.

Stratified sampling methods could be used to reduce the burden on students and the cost burden on the platform. For instance, core questions could be asked of the main sample of students, while additional long forms of the survey ask different questions of different students. The aggregate picture can be pieced together with a smaller

burden on most students and a lower cost to the platform. While this is less than ideal, it may be a necessary tradeoff in some cases.

6.2 Privacy

Privacy concerns are important and complex, and researchers are used to the question of balancing privacy concerns against the benefit of the research. The more background data a platform collects, the more risk that personally identifiable information about subjects is available through composite reports or if the data are intercepted. Access security and care in reporting results are thus crucial and should be considered ahead of time.

Because of these concerns or others, some students may wish not to provide information, which could systematically bias the survey sample, making our inferences worse. Some students who wish to opt out may be reassured if the reasons for the research and the protection of the data are made clear. Others may be more comfortable with anonymized options for responding or techniques designed for collecting sensitive data. [5]

6.3 Stereotype Threat and Maladaptive Use of Information

Arguably, the Internet provides one of the few places in society where people are not forced to reveal information about their social position, which may be of value in itself.[4] A powerful strand of research in social psychology suggests that invoking identities that are attached to negative stereotypes can hinder educational performance; people are especially vulnerable to this "stereotype threat" if they feel there is a power imbalance and that they are being defined by others' judgments [8]. This threat could both change answers provided and potentially harm the student. However, a sustained harm seems unlikely to result from the trigger of a few questions on a survey; rather, the underlying negative social context or vulnerability might be in play. It would be unfortunate if a detailed survey triggered stereotype threat, even temporarily, but making sure the questions are seen as low-stakes could help.

There may also be a risk that instructors change their courses in unintended ways if they find out more about the students. An instructor might make a college-level course less rigorous if he finds out high school students are enrolled, for instance. While this raises concerns, it is ultimately up to policy and instructors' judgment.

[4] Perhaps the Internet is the place where students "will not be judged by the color of their skin, but by the content of their character." [6] Less seriously, "On the Internet, nobody knows you're a dog." [7]

7 Conclusion

Platform operations, instructional design, and educational research would all benefit from collecting more systematic background data about students. Better knowledge about who takes MOOCs is crucial at this stage in their lifetime. I propose not only a census of MOOC users on each platform, capturing a snapshot of users today, but an ongoing effort to capture these detailed demographic snapshots at least every three years.

Acknowledgments. Many thanks to Eric Bettinger, Susanna Loeb, Tom Dee, Roy Pea, Mitchell Stevens, and participants in the Lytics Lab for valuable discussions that have influenced this paper.

References

1. Gindin, S.E.: Nobody Reads Your Privacy Policy or Online Contract? Lessons Learned and Questions Raised by the FTC's Action Against Sears. 8 Nw. J. Tech & Intell. Prop. 1, 1--38 (2009)
2. Kizilcec, R.F, Piech, C., Schneider, E.: Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In: 3rd Conference on Learning Analytics and Knowledge. Leuven, Belgium (2013)
3. Shulman, L.S.: Knowledge and Teaching: Foundations of the New Reform. Harvard Educational Rev. 57, 1--22 (1987)
4. Ackermann, E.K.: Constructing Knowledge and Transforming the World. In: Tokoro, M., Steels, L. (eds.) A Learning Zone of One's Own: Sharing Representations and Flow in Collaborative Learning Environments. pp. 15--37. IOS Press, Amsterdam, Berlin, Oxford, Tokyo, Washington, DC (2004)
5. Du, W., Zhan, Z.: Using Randomized Response Techniques for Privacy-Preserving Data Mining. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 505-510. ACM, New York (2003)
6. King, M.L.: I Have a Dream. In: Carson, C., Shepard, K. (eds.) A Call to Conscience: The Landmark Speeches of Dr. Martin Luther King, Jr. IPM/Warner Books, New York (2001)
7. Steiner, P.: On the Internet, Nobody Knows You're a Dog. The New Yorker LXIX, 20, p. 61 (1993)
8. Walton, G.M., Paunesku, D., Dweck, C.S.: Expandable Selves. In: Leary, M.R., Tangney, J.P. (eds.) The Handbook of Self and Identity, Second Edition, pp. 141--154. Taylor and Francis, New York (2012)