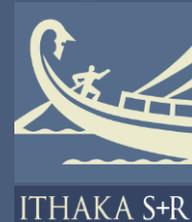REPORT

Student Data in the Digital Era
An Overview of Current Practices

September 6, 2016

Rayane Alamuddin
Jessie Brown
Martin Kurzweil

ITHAKA S+R

Ithaka S+R is a strategic consulting and research service provided by ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Ithaka S+R focuses on the transformation of scholarship and teaching in an online environment, with the goal of identifying the critical issues facing our community and acting as a catalyst for change. JSTOR, a research and learning platform, and Portico, a digital preservation service, are also part of ITHAKA.

Table of Contents

Introduction

Newly available data are making it possible to understand, improve, and represent student learning and other outcomes in profoundly different ways. With online learning platforms, technology-enabled educational tools, and other digital technologies, data about students and student learning in post-secondary settings have become unprecedentedly extensive and easy to access, interpret, and share. This growing ubiquity and granularity offer new opportunities for institutions, researchers, instructors, and other organizations to put student data to myriad uses: researchers can better understand student learning and behavior; institutions can identify institutional barriers to persistence and completion; advisors and instructors can proactively reach out to struggling students; and students can view their progress in real time and share representations of their accomplishments in new, more personalized ways.

Yet the potential of these new uses remains underdeveloped. Individual researchers, higher education institutions, and other organizations working in these areas are often hindered by challenges related to technical and analytical capacity and institutional culture, as well as sorting out what it means to collect and use data responsibly. Many have deferred or abandoned efforts in the face of these obstacles. Addressing these challenges, and achieving the potential benefits of the new student data will require a set of guiding principles, coordination within and across institutions, and enhanced technological infrastructure.[1]

To provide an overview of this landscape, we reviewed initiatives in three broad categories:

- **Research:** Student data are used to conduct empirical studies designed primarily to advance knowledge in the field, though with the potential to influence institutional practices and interventions.

- **Application:** Student data are used to inform changes in institutional practices, programs, or policies, in order to improve student learning and support.

- **Representation:** Student data are used to report on the educational experiences and achievements of students to internal and external audiences, in ways that are more extensive and nuanced than the traditional transcript.

---

[1] In June 2016, Ithaka S+R partnered with the Center for Advanced Research through Online Learning at Stanford University to host a convening of higher education scholars and leaders focused on these issues. The goals for the convening were to lay the groundwork for a shared understanding of the purposes, needs, and responsibilities of those working with student data in new ways. This report was drafted in advance of the convening, and incorporates feedback provided by participants.

Each of these three uses has its own goals and challenges, emerging norms, and a community of practitioners and commentators. Considering them together brings to the fore cross-cutting trends that can serve as a basis for building a national, multi-disciplinary dialogue about the use of student data in the digital era.

To that end, our research reveals several common themes:

- **Evidence of student learning is becoming increasingly visible and actionable, but there is wide variance in innovative use of student data.** Changes in the landscape of student information allow researchers, institutions, and other stakeholders to look beyond course grades, credit, and the degree to understand and improve learning. Activities such as time spent on task and discussion board interactions are at the forefront of research. With the assistance of technology, instructors are able to use learning objects and content mastery to better target interventions. And extended transcripts, e-portfolios, and badges make it possible for students to share more precise evidence of their learning within and outside of their home institution. Yet, while the leading edge of innovation continues to mark new boundaries of possibility, most researchers, instructors, and faculty are not working at these frontiers. Financial, technical, motivational, and cultural factors all pose obstacles to progress, leading to wide variance in student data practices across institutions.

- **Boundaries are eroding, but integration is hard.** Many of the new uses rely on the integration of data from multiple systems within institutions, each of which has a different business owner. Moreover, third-party providers are embedded throughout the process of generating, collecting, and analyzing student data. These efforts increasingly involve student data from multiple institutions, happen within the context of cross-institutional and cross-disciplinary collaboration, or demand consistent translation in diverse settings. The advantages of connecting these various sources and partners is significant, but without technical and relational standards, and a culture that encourages collaboration, reconciling them is extraordinarily difficult.

- **Ownership and governance are unclear, which carries multiple risks.** In part because of the boundary-defying nature of their work, researchers, institutions, and other organizations engaged in these areas often struggle to define who owns the data, who has authority to use them, in what ways, and for what purposes. The granularity and ubiquity of data collection and the large scale of analytical data sets complicate the question of what students must know about how data about them are being used, and what they must consent to. Furthermore, who gets to decide these questions is not clear: in some

circumstances, multiple existing governing organs claim jurisdiction; in other circumstances, none do. In an environment with unclear ownership and governance, the most prominent risk is overreach—that someone will take action that crosses an ethical line. But uncertainty can also lead to "underreach"— foregoing by inaction a real benefit to knowledge or to students.

- **There is a need for a more robust normative framework**. "Responsible use" is a term that we prefer to the more typical "privacy" or "human subjects protection." In our view, responsible use includes respect for students' privacy. But it also captures values such as transparency and student autonomy that are sometimes confusingly lumped under the heading of privacy, as well as concepts such as obligations to take action and reduce adverse affects that are not typically considered privacy concerns. Similarly, responsible use covers the Belmont Report considerations of respect for persons, beneficence, and justice,[2] but extends those values beyond the context of research to various other institutional uses. Furthermore, both existing normative frameworks are constraining; they limit potentially beneficial activity to protect other values (like keeping sensitive information to oneself). Responsible use constrains but also connotes a "responsibility to use" data in ways that improve student learning and other outcomes.

In the sections that follow, we outline the major categories of activity in research, application, and representation using new forms of student data, with illustrative examples. We discuss the practical challenges faced by individuals and institutions in those fields, as well as emerging efforts to address those challenges, focusing in particular on those related to technical infrastructure, capacity building, and coordination. Finally, in each section, we present questions related to the responsible use of large-scale student data.

The efforts we cover are based on collections of student data that were previously unavailable or not aggregated in the way they are now: primarily, they are more granular, collected in larger sets, longitudinal, or linked across systems and institutions. We are, furthermore, focused on data collected through the interaction between students

---

[2] United States, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (Bethesda, MD: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), http://ohsr.od.nih.gov/guidelines/belmont.html.

and higher education institutions or uses that inform practices and the policy context for institutions of higher education.[3]

Even within this scope, our report is by no means a comprehensive review of all initiatives currently underway. Rather, it aims to give a broad overview of the three major categories of uses we identified and some of the key issues with respect to each. We chose and reviewed the efforts we focus on through consultation with a set of advisors, extensive desk research, and conversations with key researchers and administrators. A detailed description of our research methodology and sources are presented in Appendix A.

Research

Research using large-scale learner data is progressing along a number of promising avenues. We briefly summarize four of them here, based on their overarching goals: advancing the science of learning; improving instructional design; predicting student success; and informing educational policy. Further development of these research areas is constrained by several technical challenges and infrastructure shortcomings, most notably those affecting access to data; data mining and analysis techniques; and standard data definitions, formats, and methodologies. A number of current cross-institutional initiatives are designed to address these logistical challenges. These new frontiers of research have also put pressure on the IRB system of review for responsible use; while some adaptation has begun, the transition has constrained researchers and strained reviewers.

Avenues of Research

Advancing the Science of Learning

Learning science researchers leverage educational technology, and the large and fine-grained data it offers about students' educational experiences and interactions, to better understand the underlying mechanisms of human learning. Research in this area is being undertaken mostly using educational data mining and learning analytics

---

[3] This report refers to "student" rather than "learner" data due to its explicit focus on institutions of higher education, which have unique obligations and responsibilities toward their students. Some of our examples, however, involve data collected from learners in other contexts that are relevant to practices at or the policy context for institutions of higher education.

techniques, to investigate a myriad of questions ultimately aimed at uncovering how students' educational experiences contribute to learning processes and learner success. [4]

One example is the empirical testing and development, over the past ten years, of sophisticated models of knowledge acquisition in the context of instruction. The most efficient model, Bayesian Knowledge Tracing, was first proposed in 1995. Analysis of data from Cognitive Tutor, an online tutoring program used in approximately six percent of U.S. high schools, allowed researchers to test and refine the model and develop practical applications of it.[5] Another important area of learning science is the study of engagement and disengagement from the learning process. For example, a group of researchers drew on MOOC data from over 800 students to analyze how different measures of student engagement with coursework relate to learning course performance.[6] The researchers studied various dimensions of engagement, including behavioral (e.g. submitting assignments), linguistic (e.g. polarity of language used discussion posts), temporal (e.g. time-period in which the student exhibited particular engagement), and structural (e.g. whether posts were made to the same discussion thread).

A main contribution of educational technology to the field of learning science is that it allows for a "positive feedback loop"[7] between research and practice that can substantially advance the science in unprecedented ways. Research discoveries about student learning mechanisms can lead to new hypotheses about learning, which are tested through subsequent discreet changes to course instruction that in turn yield new high-quality data for further analysis and subsequent theoretical refinement, and so on and so forth. For example, Bayesian Knowledge Tracing is the basis for Cognitive Tutor's use of the Cognitive Mastery approach to structuring curriculum, in which a student is advanced to the next skill only after they have reached mastery in a precursor skill.[8]

[4] Ryan Baker and George Siemens, "Educational Data Mining and Learning Analytics" (2014), http://www.columbia.edu/~rsb2162/BakerSiemensHandbook2013.pdf.

[5] Kenneth R. Koedinger and Albert Corbett, "Cognitive Tutors: Technology Bringing Learning Science to the Classroom," *The Cambridge Handbook of the Learning Sciences* (New York: Cambridge University Press, 2006) 61-77.

[6] Arti Ramesh et al, "Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic" (2013), http://linqs.cs.umd.edu/basilic/web/Publications/2013/ramesh:nipsws13/ramesh-nipsws13.pdf.

[7] Baker and Siemens, "Educational Data Mining and Learning Analytics," 11.

[8] Koedinger and Corbett, "Cognitive Tutors: Technology Bringing Learning Science to the Classroom."

Improving Instructional Design

Research on instructional design sometimes draws heavily on and contributes to learning science research. However, instructional design research aims to study and test instructional strategies in order to improve their design for optimal student learning, rather than to uncover the actual mechanisms of learning. Research on instructional design has focused on two areas of inquiry: how course content presentation influences learning, and which features of the learning environment promote the psychological states necessary for adequate learning.

Research conducted through the Open Learning Initiative (OLI) provides a good example of both lines of work. The OLI offers open online courses, developed by interdisciplinary teams of experts, with the express intention of offering the best evidence-based instructional practices and designs possible and engaging in ongoing learning research. Examples of inquiries using OLI data that aim to improve instructional design include studies of how the level of support provided by the course instructor, course format (e.g. fully online or blended), and course pacing and structure influence student learning outcomes. [9]

Researchers have also studied how structural components of the learning environment promote internal psychological states or behaviors in students that are conducive to learning. For example, a study using OLI data found that offering practice quizzes and feedback at specific time points during a course allows students to better self-regulate their learning.[10] Similarly, a study that drew on data from educational video games found that a particular type of incentive structure promoted a growth mindset in students playing the game, which is associated with increased student motivation and achievement.[11]

[9] Candace Thille, "Education Technology as a Transformational Innovation," White House Summit on Community Colleges: Conference Papers, 2010, http://www2.ed.gov/PDFDocs/college-completion/11-education-technology-as-a-transformational-innovation.pdf.

[10] Paul Steif and Anna Dollar, "Engagement in Interactive Web-based Courseware as Part of a Lecture Based Course and the Relation to Student Performance," paper presented at the 121st ASEE Annual Conference and Exposition, Indianapolis, IN, 2014.

[11] Eleanor O'Rourke et al, "Brain Points: A Deeper Look at a Growth Mindset Incentive Structure for an Educational Game," *Proceedings of the Third ACM Conference on Learning @ Scale*, 2016.

Predicting Student Success

The availability of large-scale computerized education activity and data has launched a prominent line of research that focuses on predicting student success in higher education settings. Researchers engage in educational data mining and learning analytics techniques to uncover models that predict specific student outcomes, such as on-time graduation or a passing grade in a course, based on student background characteristics (such as demographics) or observed behaviors (such as grades in previous coursework, choice of major, interactions with learning management systems). One important feature of this research is that it can be used to predict variables that are difficult to collect in real time, such as student engagement or progress toward degree completion. In addition, predictive learning analytics research is often completely non-intrusive, as it draws on pre-collected and -stored student data.[12]

Unlike research that aims to advance the science of learning, this research is focused on identifying meaningful patterns and developing robust predictive models that can inform student support. For example, the Open Academic Analytics Initiative (OAAI)[13] developed and empirically tested a predictive model that draws on students' demographic and background academic data, as well as course-specific data mined from the institution's learning management system (e.g. number of course sites visits, assignments submitted) to predict the likelihood of students' success in a course.[14] Although the model does not offer an understanding of how students learn, it allows institutions to provide targeted supports to students predicted to have low levels of course success.

Research on Educational Policy

The availability of digitized student records coupled with the technological ability to store, transfer, and combine datasets within and across institutions means that longitudinal data systems can be created and shared to answer key questions pertaining

---

[12] Baker and Siemens, "Educational Data Mining and Learning Analytics," 5.

[13] Led primarily by Marist College with the support of the EDUCAUSE Next Generation Learning Challenges program, the Open Academic Analytics Initiative (OAAI) leveraged academic analytics and data mining to create, test, and release predictive models of student course outcomes that can be used across institutions to generate early alerts, in an effort to improve the postsecondary outcomes of at-risk students. It developed the only open-source predictive analytics platform, and tested the portability of the model across institutional contexts and its success in informing interventions with students.

[14] Sandeep M. Jayaprakash et al, "Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative." *Journal of Learning Analytics*, 1,1 (2014): 6–47.

to college access, completion, cost, and outcomes.[15] As states and institutions are coming under increasing pressure to drastically improve completion efficiency and equity in their systems amidst tightening budgets, there is growing interest in using these data to improve and develop policies at the state and federal level.

For instance, some researchers have constructed longitudinal datasets connecting varied educational and income or employment data at the state or national level, allowing more sophisticated analysis of the consequences of policies than was ever possible in the past. One study, for example, draws on multiple datasets with information on student aid applications, loan transactions, and college completion, resulting in over 45 million individual observations on four million student loan borrowers. When they combined these data with Federal Student Aid data, the researchers were able to shed light on the role of federal government and institutional policies on student loan practices and default rates.[16]

Research Infrastructure

The infrastructure supporting the research advances described above is still in an early stage of development. Much of the research relies on the sometimes heroic efforts of individuals or teams of researchers cobbling together data from multiple sources, developing techniques iteratively, and exploiting loopholes or special relationships to gain access to data. For the field to expand—for more researchers to participate in both developing and replicating the science—a more robust infrastructure is needed. Three specific areas in need of improvement are access to data; support for mining and analyzing data; and the standardization of research formats and practices.

Improving Access to Data

In light of the different populations served by different institutions of higher education, the most powerful analyses typically rely on data from multiple institutions. Yet access to such massive and varied datasets is not a given. For example, it is frequently difficult for researchers to gain access to datasets collected by other researchers or institutions for other purposes. Data sharing over open-source platforms can create ambiguous rules

---

[15] Jamey Rorison et al, "Employing Postsecondary Data for Effective State Finance Policymaking," Lumina Foundation (2016).

[16] Adam Looney and Constantine Yannelis, "A Crisis in Student Loans? How Changes in the Characteristics of Borrowers and in the Institutions They Attended Contributed to Rising Loan Defaults," Brookings Papers on Economic Activity, Fall 2015 Conference.

about data ownership and publication authorship, or raise concerns about data misuse by others, thus discouraging liberal sharing of data. In other cases, institutional silos pose a challenge as learner data may reside in different departmental systems that do not collaborate or whose data structures are incompatible, making it difficult to gather and aggregate needed data. IRB requirements may also restrict access to data. Student consent forms, for instance, often prohibit the sharing and alternative use of de-identified data. At the state and national level, Congress' 2008 ban on the creation of a federal student unit record system and the exclusion of large swaths of students from existing federally-mandated datasets greatly limits policy-oriented research.

There are several important initiatives designed to address these data access challenges, for individual researchers as well as institutions and states. LearnSphere, a cross-institutional community infrastructure project, aims to develop a large-scale open repository of rich education data by integrating data from its four components.[17] For instance, DataShop stores data from student interactions with online course materials, intelligent tutoring systems, virtual labs, and simulations, and DataStage stores data derived from online courses offered by Stanford University. Click-stream data stored in these repositories include thousands and even millions of data points per student, much of which is made publicly available to registered users who meet data privacy assurance criteria. On the other hand, MOOCdb and DiscourseDB, also components of LearnSphere, offer platforms for the extraction and representation of student MOOC data and textual data, respectively, surrounding student online learning interactions that are otherwise difficult to access or are highly fragmented. By integrating data held or processed through these different components, LearnSphere will create a large set of interconnected data that reflects most of a student's experience in online learning.

Given the importance of access to high quality student-level institutional data to inform educational policy, a number of voluntary data initiatives are improving access to learner data at the national level by creating and collecting new and more robust student data, and aggregating them across their numerous participating institutions.[18] For example the Multistate Longitudinal Data Exchange (MLDE) project, spearheaded by the Western Interstate Commission for Higher Education, has aggregated student data across four states and linked all public high school and postsecondary data for specific cohorts of

[17] LearnSphere is a joint project of Carnegie Mellon University, Stanford University, the Massachusetts Institute of Technology, and the University of Memphis, that aims to transform learning science by creating a large, distributed data and community software infrastructure and developing the capacity for course developers, instructors, and learning engineers to make use of it.

[18] See J. Rorison et al, "Employing Postsecondary Data for Effective State Finance Policymaking," Appendix A, for a list of voluntary data initiatives.

students with their employment data. By filling gaps in existing state datasets, MLDE has allowed researchers to capture student mobility and better explain employment patterns and outcomes relevant to the states' workforce and educational attainment policies and goals.[19]

The Predictive Analytics Reporting (PAR) Framework also aggregates student data from various institutions across multiple states, creating large cross-institutional datasets that focus on postsecondary student data, including student demographics, high school information, and a range of student course-taking behaviors and outcomes.[20] These data have been used to construct and validate predictive models of student success. For example, datasets from two PAR members, the University of Maryland University College and the University of Hawaii system, were used to assess how community college completion metrics predict students' future success in four-year institutions.[21] In addition to aggregating student-level data in the absence of a federal unit record system, organizations are putting forth policy recommendations and solutions to improve access to data. The Institute for Higher Education Policy's (IHEP) Postsecondary Data Collaborative, for example, recently published 11 policy papers in a significant and concerted collective effort to improve national postsecondary data systems.[22]

Facilitating Mining and Analysis of Student Data

At the moment, the technical capacity to analyze data in the ways described above is limited to relatively few researchers and institutions. This is in part because the methodologies for extracting, organizing, and mining the data are still in development and are not widely known. Even when they are known, most education researchers do not have the computer science expertise required to implement them. It is also the case that significant computing power, not accessible to every researcher, is needed to process massive datasets.

---

[19] Brian T. Prescott, "Beyond Borders: Understanding the Development and Mobility of Human Capital in an Age of Data-Driven Accountability," Western Interstate Commission for Higher Education (2014).

[20] The Predictive Analytics Reporting (PAR) Framework, recently acquired by Hobsons, is a national, non-profit provider of learning analytics and data mining collaborative that focuses on improving student retention in US higher education. PAR aggregates data from two-year, four-year, public, proprietary, traditional and progressive colleges and universities in a single data resource, and applies systematic exploratory, inferential, and descriptive techniques to identify patterns of risk for non-success among students. Additionally, PAR offers tools to make an inventory for and measure the effects of interventions on student outcomes.

[21] See http://www.parframework.org/2015/05/par-framework-releases-results-of-groundbreaking-research-predicting-transfer-student-success-with-community-college-data/.

[22] See http://www.ihep.org/research/initiatives/postsecondary-data-collaborative-postsecdata.

A number of emerging tools and structures are intended to increase the ease, efficiency, and expediency of data mining and analysis. A goal of the LearnSphere project is to support researchers in their use of the student data it provides. To that end, its four initiatives are developing systems that allow researchers to upload their data or run data through the platform software as easily as possible, without needing assistance from their institutions' IT departments to access the platform functionalities. These functionalities include tools that help researchers extract, mine, visualize, interpret, and analyze their data in new efficient ways.

DataShop, for example, offers push-button tools through its web-based interface such as the "performance profiler," which allows researchers to select a list of data variables within a particular domain and view a range of performance metrics for those variables to facilitate analysis (e.g. view the error rate for problem-solving questions about the area of a rectangle from a math cognitive tutor course). The "learning curve" tool creates graphs for visualizing student performance on select variables, with the option of clicking on different points on the graph to display descriptive data for a specific interaction. Among other tools, MOOCdb organizes data in new ways that make it amenable to inquiry and analysis, and offers scripts that researchers can run on their data to create new higher-order variables for analysis. DiscourseDB, which allows researchers to represent and analyze textual student data from multiple sources, facilitates data mining by simplifying the data and supports analysis by offering data contextualization and annotation tools.

OAAI and PAR, on the other hand, developed portable predictive analytics models, which have been validated across institutional contexts. These predictive models can be applied to datasets from different institutions, without the need to repeatedly mine data and develop models completely anew. This expedites the research process and consequently the time and resources needed to develop and implement consequent interventions. The OAAI model is open-sourced and freely available to researchers and institutions.

Standardizing Practices in the Field

Standards of practice common to more mature fields are still developing for the emerging forms of student data research. Definitions and format of data differ from source to source, institution to institution, and researcher to researcher. Expectations for data quality, analytical rigor, and reporting are still being iterated. Standardizing these practices would accelerate progress by facilitating broader access to data and analytical capacity, as well as greater collaboration among researchers.

Several of the data sharing projects described above are developing these kinds of standards. To create its cross-institutional predictive models, the PAR Framework established standard data definitions for the transcript and student-level data its member institutions contribute to the platform. These definitions have been shared with the broader higher education research community through publications under a Creative Commons license in an effort to encourage their adoption.

DataShop developed an XML logging format that allows for data from various sources to be collected into the repository uniformly and efficiently, in part by specifying which specific learner transactions get logged and how each is defined. Similarly, DataStage, MOOCdb, and DiscourseDB developed software schemas that allow researchers to standardize their data according to their particular format. Once data is in one of these platforms' standard format, researchers using that platform can efficiently organize their data based on common definitions and variables, run analyses using existing tools, and share comparable data and findings with the research community. Furthermore, researchers can create and share additional analysis tools, which can be adopted by other researchers in the field thanks to the use of standard formats and definitions. These initiatives make documentation of their process and definitions available to the public through their websites, to encourage widespread standardization.

Responsible Use

Researchers collect, contribute, and analyze student data under the auspices of IRB review of human subjects research. This system of review, while sophisticated and familiar, developed its processes and policies for a context different from new research areas and data uses. For instance, the automatic collection of students' data through interactions with educational technologies as a part of their established and expected learning experiences raises new questions about the timing and content of student consent that were not relevant when such data collection required special procedures that extended beyond students' regular educational experiences of students. Additionally, IRBs differ greatly in their policies and practices, and their capacity for keeping up with technological advancements in the field and their implications for responsible use. Researchers and IRBs have handled this mismatch in different ways; creatively and effectively in many instances, but contributing nevertheless to a confusing mash of policies and procedures and a general sense of uncertainty.

The cross-institutional nature of much of the research has complicated matters further. In some instances, cross-institutional research ventures have trusted that each collaborating researcher secured necessary IRB approval and is following privacy and ethical use requirements determined by their home institution. In other cases, lead

researchers have required collaborators to satisfy their home institution's IRB requirements. Some institutions, especially community colleges, do not have their own IRB and therefore rely on approval by their president or defer to collaborators instead.

To avoid uncertainty and data privacy challenges, DataShop and DataStage both developed a strong working relationship with their institutions' IRBs. DataShop, for instance, worked closely with the IRB at the Carnegie Mellon University to clarify the technical details and implications of its work. In turn, the IRB helped DataShop develop guidelines and forms to share with researchers who are considering contributing their data to the project, and to develop an efficient system for checking that contributed data meet privacy rules and can thus be used on the platform. Other researchers working on cross-institutional ventures had to unexpectedly work with different IRBs on a case-by-case basis to clarify the nature of the technologies they are using and data they are extracting, and their implications for student privacy.

Among the researchers we interviewed, protecting student data from identification or re-identification was a salient concern. This is not surprising considering the primacy of the protection of personally identifiable information under FERPA. To that end, researchers put in place procedures for de-identifying data and conducting manual or automated data checks of fields that may unexpectedly contain identifying information (such as free-response fields). Some use double-encryption to secure files that link student names with identifiers, and others rely only instead on datasets that have already been made public. In some cases researchers avoided collecting data that are particularly sensitive or pose higher risks of unintentional re-identification, such as financial aid or demographic data. Additionally, some of the initiatives we reviewed include features that allow researchers to specify different privacy settings for their data based on the particular circumstances. Some are looking to emerging privacy technologies, including encryption schemes, to secure their data.

As learner data become more interconnected and comprehensive, however, new privacy and responsible use concerns arise. For instance the risk of re-identification increases by virtue of having more data points on students from multiple contexts. Additionally, new possibilities for data collection and analysis emerge as technologies advance and their use in education becomes more affordable and widespread. For example, fine-grained education data could be merged with genetic, biomarker, market behavior, and portable electronics data in ways that are unforeseen by students at the time of data collection, raising new questions about student consent. The advance of fingerprint, eye-tracking, face and voice recognition, and GPS technologies can also move education research into new uncharted territory. Furthermore, as student data are increasingly automatically generated and the contributions of research to improving students' experiences and outcomes increase, researchers and institutions may become faced with questions about

their obligations to analyze data that can advance knowledge in the field and subsequently improve educational practices and policies at scale. This will not only necessitate a continuous review of responsible use policies and practices, but also an infrastructure for close collaboration between experts from a growing number of disciplines who are increasingly involved in educational data collection but historically isolated from each other.

Application

Modern higher education institutions have unprecedentedly large and detailed collections of data about their students, and are growing increasingly sophisticated in their ability to merge datasets from diverse sources. As a result, institutions have great opportunities to analyze and intervene on student performance and student learning. While there are many potential applications of student data analysis in the institutional context, we focus here on four approaches that cover a broad range of the most common activities: data-based enrollment management, admissions, and financial aid decisions; analytics to inform broad-based program or policy changes related to retention; early-alert systems focused on successful degree completion; and adaptive courseware.

Despite the promise of new applications of student data, some institutions have struggled to take advantage of them, stymied by technical challenges or the inability to scale innovations among stakeholders. These applications also generally fall outside of the IRB structure, and few institutions have alternative policies and procedures governing appropriate activity. Thus, in many situations, it is not only unclear what it means for an instructor or administrator to use student data responsibly, it is also unclear who decides what constitutes responsible use.

Areas of Application

Data-Driven Enrollment Management

College admissions and enrollment management have always been data-driven practices. However, as colleges and universities gain access to more data about students, and as they augment their capacity to analyze these data in sophisticated ways, they can more precisely and efficiently predict which students will attend and succeed at their institutions, and are using predictive algorithms to inform recruitment campaigns, admissions decisions, and financial aid offers.

In addition to building improved algorithms to predict yield and success, institutions are expanding the types of data they use to make admissions decisions. For example, Ithaca College uses applicant activity on IC PEERS, a social media website on which prospective students can connect with one another and Ithaca faculty, to gauge student interest in the college and predict how likely they are to enroll.[23] Although rarely as sophisticated as Ithaca's efforts, such review of social media information is quite common: a 2015 Kaplan Test Prep Survey of 397 admissions officers found that 40 percent of admissions officers visit applicants' social media profiles, often to verify information presented on their applications.[24]

Institutions also use student data to inform policies related to diversity and financial aid. For example, Franklin & Marshall College, which increased the share of Pell Grant-eligible students in its entering class from five percent in 2008-09 to 21 percent in 2014-15, partnered with Third Coast Analytics to determine how to design admissions and financial aid policies that would maximize yield and retention of those students in a cost-effective way.[25] Similarly, University of Richmond used historical data on admissions, financial aid, and yield to build a predictive model that informed its "Richmond Promise to Virginia" program, which was a key strategy in increasing its share of Pell-eligible students from nine percent in 2008-09 to 16 percent in 2012-13.[26] Institutions have also relied on careful data analysis to target financial aid awards, to maximize yield or to craft a diverse class.

Analytics to Inform Program and Policy Change

Institutions are increasingly basing decisions about their business processes and programs on analysis of student outcome data. Georgia State University (GSU), which

---

[23] See Emmanuel Felton, "Colleges Shift to Using 'Big Data'—Including from Social Media—in Admissions Decisions," *The Hechinger Report* (August 21, 2015), http://hechingerreport.org/colleges-shift-to-using-big-data-including-from-social-media-in-admissions-decisions/.

[24] "Kaplan Test Prep Survey: Percentage of College Admissions Officers Who Check Out Applicants' Social Media Profiles Hits New High; Triggers Include Special Talents, Competitive Sabotage," Kaplan Test Prep (January 13, 2016), http://press.kaptest.com/press-releases/kaplan-test-prep-survey-percentage-of-college-admissions-officers-who-check-out-applicants-social-media-profiles-hits-new-high-triggers-include-special-talents-competitive-sabotage.

[25] "Posse Program: STEM Cohort, Franklin & Marshall College," Building Blocks to 2020: First in the World Competition (January 2014), http://www.naicu.edu/special_initiatives/2020/detail/posse-program-stem-cohort-2. Franklin & Marshall Financial Aid, http://www.fandm.edu/financial-aid/financial-aid-resources; some data provided by Franklin & Marshall.

[26] Data from U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics Integrated Postsecondary Education Data System. The Richmond Promise offers a grant equal to tuition, room, and board (without loans) to students whose families make less than $60,000. See Nicole Cohen, "Cutting Costs," *Richmond Magazine* (September 11, 2013), http://richmondmagazine.com/life-style/cutting-costs-09-11-2013/.

---

increased graduation rates from 32 percent in 2003 to 54 percent in 2014, achieved this improvement by using data to systematically identify and address barriers to retention and completion through analysis of its data.[27] For example, advisors and administrators used an analysis of courses in which students consistently performed poorly to target supplemental instruction, and to inform a redesign of introductory math courses. GSU has also developed a program targeting small, conditional grants to students who are highly likely to complete a degree but would be forced to drop out for a semester because of a small amount due to the bursar.

The University of Texas, Austin (UT Austin) has employed a similar approach, creating intensive support programs for students who, according to an algorithm based on the institution's historical data, have a low probability of graduating in four years. For example, UT Austin's University Leadership Network (ULN) provides academic and financial support to the 500 students in each entering cohort whose unmet financial need is greatest and whose predicted probability of graduation is lowest.[28]

Early-Alert Systems for Advisors, Instructors, and Students

In addition to using student data analysis to inform institution- or program-level change, institutions are increasingly putting predictive analytics into the hands of instructors, advisors, or students themselves. For example, early-alert systems aggregate and analyze large datasets from multiple sources (such as gradebooks, LMS log-files, student information systems), and automate the process of identifying student behavior that is associated with hindered progress to successful course or degree completion. Such systems reduce the time advisors or faculty must spend monitoring student performance, and help them better target support. When the information is used to prompt students directly, it may motivate them to change their behavior. All of these features have the potential to make student support more effective and more efficient.[29]

---

[27] Martin Kurzweil and Derek Wu, "Building a Path to Student Success at Georgia State University," Ithaka S+R (April 23, 2015), http://www.sr.ithaka.org/wp-content/uploads/2015/08/SR_Case_Study_Building_Pathway_Student_Success_042315_0.pdf.

[28] Paul Tough, "Who Gets to Graduate," *The New York Times Magazine* (May 15, 2014), http://www.nytimes.com/2014/05/18/magazine/who-gets-to-graduate.html. Other institutions that have used aggregate analyses of student demographics, academic history, or registration behavior to design broad-based interventions to improve retention and completion include Virginia Commonwealth University, University of Maryland University College, and Florida State University. See Joseph Yeado, Kati Haycock, Rob Johnstone, and Priyadarshni Caplot, "Learning from High-Performing and Fast-Gaining Institutions," Education Trust (October 2013), http://edtrust.org/wp-content/uploads/2013/10/PracticeGuide1.pdf.

[29] These applications also create a risk of algorithmic bias. See the "Responsible Use" section below, or for a more sustained discussion, see "The Predictive Analytics Revolution: Leveraging Learning Data for Student Success," ECAR Working Group Paper (October 7, 2015), http://net.educause.edu/ir/library/pdf/ewg1510.pdf.

Early-alert systems vary in the type of data they use, the interventions they inform, and their technical complexity. GSU's advisor-facing Graduate Progression System (GPS), developed by EAB, merges course grade data with student information, academic history, and registration data. GPS determines characteristics and behaviors associated with retention and completion based on ten years of GSU student data. If a student engages in behavior that decreases their chances of graduating, their advisor receives one of 700 different system alerts based on the student's behavior, risk level, and context. Potential triggers include failure to achieve a minimum grade in a required course, failure to complete a course by a particular point in one's academic career, and registration for a course that is not part of a student's program of study. Advisors also have access to extensive dashboards, which are updated daily, that give them easily digestible information about their students so that they can determine how to best provide support.[30]

Other advisor-facing systems, like Arizona State University's e-Advisor, use information about student activity in learning management systems in addition to registration data and student background information. When students get off track, their advisors are notified and asked to intervene. The tool also uses students' academic performance data to make registration suggestions to students and advisors. Advisors can also manually flag students in e-Advisor for sustained tracking and intervention.[31]

While GPS and E-Advisor target advisors as the primary audience for their alerts, and progress to degree as the primary unit of intervention, other institutions focus their predictive analytics on successful course completion. These institutions use systems that offer instructors or students alerts when a student is at risk of failing. Purdue University's Course Signals tool and Rio Salado's RioPace tool are two well-known examples. These tools merge student demographic information and academic history with learning management system log-file data to predict student likelihood of success within a given course.[32] Commonly observed behaviors include student LMS log-ins, assignments submitted, assessment and discussion board activity, and content pages viewed. All of this information is used to determine a student's risk level, which is then communicated to instructors in an easily interpretable format. For example, Purdue's system, Course Signals, presents student risk levels in three categories: red (highest

---

[30] GPS Advising at Georgia State University," Georgia State University, http://oie.gsu.edu/files/2014/04/Advisement-GPS.pdf.

[31] ASU only uses LMS data for fully online students. See https://eadvisor.asu.edu/.

[32] For RioPace, see "RioPace," Rio Salado College, http://www.riosalado.edu/riolearn/Pages/RioPACE.aspx. For information on Course Signals, see Matthew D. Pistilli and Kimberly Arnold, "Signals: Using Academic Analytics to Promote Student Success," *Educause Review* (July 17, 2012), http://er.educause.edu/articles/2012/7/signals-using-academic-analytics-to-promote-student-success.

risk), yellow, and green (not-at-risk). Instructors can run analyses on demand and reach out to students who are identified as at-risk.

Finally, a number of early alert systems provide information directly to students (sometimes in addition to advisors or instructors). Students at Purdue can see their Course Signals risk level and compare it to an aggregate and anonymized view of the rest of the class. ASU's eAdvisor also has a student-facing dashboard. E2Coach, a tool used in University of Michigan introductory STEM courses, automatically sends students personalized messages based on student preferences and a continually updated algorithm. Students are also provided with graphics that allow them to compare their progress to that of the rest of the class.[33] This approach invests academic responsibility in the student—rather than in the instructor or advisor—and rests implicitly on the assumption that data can most effectively change student behavior when students are given ownership over its insights and results.

Adaptive Courseware

A final, increasingly popular application of student data analytics is adaptive courseware. Adaptive courseware collects information on student learning activity on the platform— such as time spent on task, performance on tasks and assessments, and platform engagement—to create "personalized learning paths" for students based on their performance. The algorithms that determine these paths are often based on the learning sciences research discussed in the previous section. Adaptive platforms offer dashboards and analytics tools that allow instructors to both see where the class as a whole is struggling and drill-down on individual student performance. Many solutions also provide students with a dashboard so that they can better understand their progress and roadblocks. While adaptive platform dashboards for instructors and students have parallels to some of the early-alert systems discussed above, they typically focus on student progress towards mastery in learning outcomes, rather than successful course completion.[34]

Though few institutions have built their own adaptive learning platforms, the marketplace of third-party adaptive learning providers is growing in size and usage,

---

[33] "E2Coach: Tailoring Student Support for Students in Introductory STEM Courses," *Educause Review* (December 6, 2013), http://er.educause.edu/articles/2013/12/e2coach-tailoring-support-for-students-in-introductory-stem-courses.

[34] For an articulation of this distinction, see Keith Hampson, "Analytics in Online Higher Education: Three Categories," Acrobatiq (April 25, 2014), http://acrobatiq.com/analytics-in-online-higher-education-three-categories/.

especially for introductory and remedial coursework.[35] These providers often have in-house research, data-science, and design teams that develop the algorithms and machine learning processes to structure students' personalized learning paths. Some offer their own curriculum, while others work with faculty at their client institutions to develop a customized solution.

Technical and Cultural Barriers to Application

While the application of student data analysis to improve instruction and support is becoming more common, a majority of institutions are not systematically engaged in such efforts.[36] Technical challenges are a significant impediment to further development, with many institutions confronting poorly integrated systems and a lack of capacity to use them. But, even at institutions that have overcome these technical challenges, innovations frequently remain at the margins, limited by poor planning for scale and a culture and incentives that oppose their adoption.

One of the biggest technical challenges that institutions face is aggregating data from multiple systems. The data needed for sophisticated analytics are usually dispersed and differentially formatted in student information systems, registrar's data systems, LMS log-files, and other systems. Some institutions have the technical and human resource capacity to merge these data into a common database for mining and analysis, but smaller and less-resourced institutions often do not. Several emerging efforts at interoperability standards, such as the Tin-Can API and the Caliper Analytics framework, aim to simplify data integration across systems.[37]

---

[35] See Jessie Brown, "Personalizing Post-Secondary Education: An Overview of Adaptive Learning Solutions for Higher Education," Ithaka S+R (March 18, 2015), http://www.sr.ithaka.org/wp-content/uploads/2015/08/SR_Report_Personalizing_Post_Secondary_Education_31815_0.pdf; "Learning to Adapt: Understanding the Adaptive Learning Supplier Landscape," Tyton Partners (April 2013), http://tytonpartners.com/tyton-wp/wpcontent/uploads/2015/01/Learning-to-Adapt_Case-for-Accelerating-AL-in-Higher-Ed.pdf, and Michael Feldstein, "What Faculty Should Know About Adaptive Learning," e-literate (December 17, 2013), http://mfeldstein.com/faculty-knowadaptive-learning/.

[36] "Embracing innovation: 2015-2016 Higher Education Industry Outlook Survey," KPMG (2015), http://www.kpmg-institutes.com/content/dam/kpmg/governmentinstitute/pdf/2015/he-outlook-2016.pdf.

[37] The Advanced Distributed Learning Initiative (ADL), has developed the Experience API (xAPI), also known as the Tin Can API, which standardizes learning activity tracked in disparate systems (for example LMS log-ins, reading an article watching a training video, having a conversation with a mentor), and stores them together in a Learning Record Store. While the Learning Record Store is a novel representation of student learning (discussed in a later section), it also serves a similar purpose to an institutional data warehouse, and aggregates data in such a way so that it can be analyzed and acted upon. Similarly, the IMS Global Learning Consortium is in the process of developing the Caliper Analytics Standard, which will establish a means for consistently tracing and presenting measures of learning activity—tracked in different systems—in a standardized way. Like the xAPI, the Caliper framework will enable interoperability amongst systems. In addition, it will establish guidelines for formatting, labeling, and presenting learning activity data. For information on the

---

A similar challenge relates to the tools that institutions use to analyze data and present them to stakeholders. In-house analytical capability is not a given, but it is perhaps even more rare to find staff who can visually represent data to stakeholders in meaningful and actionable ways.[38] While some institutions, such as Purdue and Rio Salado, have the resources to build their own solutions, many more are turning to the growing market of third-party analytics providers. For institution- and program-level analytics, popular providers include Civitas's Illume; IBM Analytics; Starfish's Enterprise Success Platform; and Blackboard Analytics. Many of these companies, along with EAB, which created GSU's GPS, also provide solutions for early alert systems.

These third-party platforms offer customization options, but the core algorithms they use tend to be proprietary and are not shared with clients. This secrecy can make it hard for institutions to gauge the integrity and flexibility of the algorithms; it also raises questions about the ethics of making decisions about students' instructional pathways based on a black box that administrators, instructors, and students do not understand.[39] Several prominent scholars in the field, including Candace Thille and George Siemens, have stressed the importance of academic institutions taking the lead in developing learning analytics solutions, for just this reason. Cross-institutional collaborations focused on the development of shared or open analytics models and resources, such as the Open Academic Analytics Initiative and the PAR Framework (discussed in the Research section) offer some potential for augmenting institutional capacity for this kind of innovation.

Institutions also face challenges to accessing external data on students, which can be helpful in developing a more holistic or long-term understanding of student learning and other outcomes. Increasingly, institutions are entering partnerships with local or regional entities that also interact with their students to share data. For example, Long Beach City College has entered a data-sharing agreement with the Long Beach Unified School District to gain access to entering students' data, and has redesigned the way students are placed in developmental courses based on a new understanding of which high school outcomes correlate with success. Similarly, Florida's Valencia College, the University of Central Florida, and several Orlando-area public school districts are

---

xAPI, see "What is the Tin Can API," Tin Can API, https://tincanapi.com/overview/. For information on the Caliper framework, see "Caliper Analytics Background," IMS Global Learning Consortium, https://www.imsglobal.org/activity/caliperram

[38] See, for example, JISC, "Learning Analytics in Higher Education," and Randall J. Stiles, "Understanding and Managing the Risks of Analytics in Higher Education," *Educause* (June 2012), http://net.educause.edu/ir/library/pdf/EPUB1201.pdf

[39] For a discussion of these concerns, see Goldie Blumenstyk, "As Big Data Companies Come to Teaching, a Pioneer Issues a Warning," *The Chronicle of Higher Education* (February 23, 2016), http://chronicle.com/article/As-Big-Data-Companies-Come-to/235400.

creating a federated data system, and plan to use information on entering students to strengthen its support system for these students.

Finally, in order for data-driven interventions to be wide-spread, institutions must sustain a culture that embraces the use of data, and create incentives for data-driven activities amongst administrators, instructors and student support staff. Large-scale, data-driven policy changes are implemented with minimal friction and maximal buy-in when leaders demonstrate a commitment to data-informed decision-making, and create multiple opportunities for stakeholders to make sense of and contribute to the direction of the change. Users not only need to be trained on the proper ways to use these tools and communicate with students, they also require meaningful incentives to take on the potentially steep learning curve.[40]

Responsible Use

New applications of student data raise new questions regarding institutional obligation: Who should have access to student data? Who should intervene on insights? What should students know about how their data are being used? Because the work is not considered "research"—as it is not intended to be published or to promote general understanding of the issues—most Institutional Review Boards claim no jurisdiction, and few institutions have formal policies to guide their applications of student data. Despite this lack of official documentation, all of the institutional stakeholders with whom we spoke faced a similar set of questions regarding responsible use. In addition, each had devised similar—yet institutionally specific—ways of answering these questions to protect student privacy, autonomy, and integrity.

At the most basic level are questions of who has access to a student's information and how much information they can access. Typically, the answer differs depending on an individual's relationship to students and the purpose for which they use the information. For example, Rio Salado's RioPace offers identifiable information to instructors about students in their courses so that instructors can reach out to students on an individualized basis. On the other hand, the programmers charged with monitoring and updating RioPace's predictive model have access to all student data in an anonymized form. Once decisions about access have been made, institutions must reinforce them with appropriate security protocols, such as anonymization, de-identification, and encryption.

---

[40] See "The Predictive Learning Analytics Revolution," ECAR Working Paper.

Another concern is student awareness of and consent to their data being used in a particular way. As in the case of research involving new forms of student data, the conventional conceptions of informed consent and opt-out are not a clean fit for new applications of student data. For example, predictive models may be compromised if students are non-randomly excluded. In other scenarios, student awareness of predicted outcomes, the algorithms underlying them, or the interventions meant to address them may undermine the effectiveness of the interventions. Moreover, if administrators or faculty know that having certain information about a student or undertaking a certain intervention increases the likelihood that a student will succeed, they may feel an obligation to use those tools to support students, regardless whether the student consents.[41] Currently, most institutions do not provide students the opportunity to opt out, and many of those we interviewed felt that their institutions could do a better job of communicating clearly with students about how their data was being used.

Another major ethical concern is that predictive systems may reproduce bias or stereotypes built into algorithms or underlying datasets. A key question in this regard is whether predictions are based on mutable or immutable characteristics. Purdue and Rio Salado, for example, base their predictive models primarily on student activity logged in a learning management system, because those data represent behavioral patterns that can be changed, rather than demographic information that might lead to stereotyping in algorithm design or intervention. By contrast, UT Austin relies more heavily on unchanging variables like student demographics, family characteristics, and pre-college academic history in order to sort students into immersive programs of support from their first term at the institution. For UT Austin, the model's potential to guide impactful interventions outweighs its potential for bias.

Regardless of which information is used in the algorithm, the predictions that these models yield can be used in ways that impact students' experiences and mindsets in different ways. If the results of an algorithm automatically trigger certain pathways or interventions, any biases present in the algorithm will be reproduced in reality. Monitoring for such results and allowing discretionary judgment to override the algorithm can mitigate such consequences, but could also reinforce the idiosyncratic biases of the individual making the judgment or, if it happens frequently, undermine the general efficacy of the system.

There is also a risk that students, advisors, or instructors internalize probabilistic predicted outcomes and turn them into self-fulfilling prophecies. Advisors and

---

[41] For a more comprehensive framing of these questions, see James E. Willis III, "Ethics, Big Data, and Analytics: A Model for Application," Instructional Development Center for Publications (May 6, 2013), http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1001&context=idcpubs.

instructors who use early alert systems need to be reminded that the information they receive from these systems does not constitute a full picture of student performance and behavior, and must communicate with students accordingly. Researchers at Purdue and University of Michigan have even done extensive research on the type of feedback that has the biggest impact on student success, and structure automated and manual communications to students in ways that maximize student agency and growth mindset while minimizing the potential for discouragement. Similarly, UT Austin is careful to give their support programs names and inform students of their selection in ways that emphasize students' leadership and scholarship.

Running through all of these issues is the question of who gets to decide. The current modal situation seems to be that, with the IRB and other decision-making bodies declining jurisdiction, the administrators or faculty members working on these projects make decisions on their own, perhaps in consultation with the president or provost, or the institution's counsel, in an ad hoc manner. Some institutions have begun to develop governance structures around these non-research applications of data. For instance, the chief privacy officer of UCLA, one of the growing set of institutions to have such a role, has developed a data governance plan that includes a cross-disciplinary and cross-functional data governance committee and a set of guidelines for adjudicating the various internal uses of the institution's student data. Educause has taken a leadership role in coordinating efforts like these and in defining the role of chief privacy officers. Since 2014, the organization has published an Information Security Guide for Chief Privacy Officers and other information security professionals that provides guidelines on FERPA and HIPPA compliance, de-identification, information security governance, and risk management frameworks. These guidelines are useful for privacy professionals involved in both research and applications with student data.[42]

While frameworks like these, as well as FERPA and HIPPA regulations, provide some direction for protecting student privacy, few institutions have developed comprehensive guidelines for addressing the other risks described here. Notable exceptions include Open University and JISC, both in the United Kingdom, which have developed ethical policies to guide learning analytics.[43] Additionally, while no official framework for responsible use exists, numerous researchers have outlined broader sets of principles

---

[42] For the 2014 Information Security Guide, see https://spaces.internet2.edu/display/2014infosecurityguide/Home. Much has been written on ethical and responsible use of student data in learning analytics.

[43] For Open University's policy, see "Policy on Ethical Use of Student Data for Learning Analytics," http://www.open.ac.uk/students/charter/sites/www.open.ac.uk.students.charter/files/files/ecms/web-content/ethical-use-of-student-data-policy.pdf. For JISC's policy, see "Code of Practice for Learning Analytics," JISC, https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics.

that might guide responsible applications of student data across institutional contexts.[44] One of the most widely cited of these frameworks was authored by Sharon Slade and Paul Prinsloo and serves as the basis for Open University's Policy on Ethical use of Student Data for Learning Analytics. The document outlines six principles, or ways of thinking about the use of student data, that should guide learning analytics. These are: learning analytics as a moral practice; students as agents in the construction of their records; student identity and performance as temporal, dynamic constructs; student success as a complex and multidimensional phenomenon; transparency regarding data practices; and an institutional obligation to use data.[45] As the use of data for interventions on student success and learning becomes more widespread, principles like these, which extend beyond privacy protection to maintain student autonomy, integrity, and agency, should be generated and iterated upon both within and amongst institutions.

Representation

New sources and ways of organizing student data have fueled a proliferation of more-detailed representations of student learning and achievement. As more students, and especially non-traditional students, accumulate credit and experiences across multiple institutions and platforms, institutions and other providers need a way to benchmark and recognize these attainments in broadly recognizable ways. Furthermore, there is growing recognition that course grades represented on traditional transcripts do little to communicate a student's actual skill set, and employers are increasingly relying on other documents to assess students' skillset and fit.[46] This growing demand for more nuanced information is being met by the increasing activity of students on platforms that capture and archive the work they produce, which open up new possibilities for representing learning more holistically than a traditional transcript does.

---

[44] See, for example, Abelardo Pardo and George Siemens, "Ethical and Privacy Principles for Learning Analytics," *British Journal of Educational Technology,* 45:3 (2014), pp. 438-450; Sharon Slade and Paul Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," *American Behavioral Scientist,* 57: 1510 (March 2013), http://abs.sagepub.com/content/57/10/1510; Jenni Swenson, "Establishing an Ethical Literacy for Learning Analytics," in Abelardo Pardo and Stephanie Teasley, ed. "Proceedings of the Fourth International Conference on Learning Analytics and Knowledge," New York: ACM (2014), pp 264-250. These papers outline issues and propose frameworks.

[45] Slade and Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas."

[46] In AAC&U's 2015 public opinion survey of employers, 80 percent of employers said they would find an ePortfolio very or fairly useful in helping to evaluate job applicants' potential to succeed at their company. Just 45 percent held the same view about traditional college transcripts. See "Falling Short? College Learning and Career Success: Selected Findings from Online Surveys of Employers and College Students Conducted on Behalf of the Association of American Colleges and Universities, AAC&C (January 20, 2015), https://www.aacu.org/leap/public-opinion-research/2015-slides.

Services such as LinkedIn and Degreed have seized the initiative in innovative forms of representation, and students (and alumni) are increasingly relying on them as records of their accomplishments to present to external stakeholders. These services challenge the institution as the primary owner, purveyor, and sharer of student learning, but a number of institutions are engaging with this shift, seeking to reinvigorate their role in verifying learning by participating in the creation of a more student-centric and experientially diverse student record.

Approaches to Representation

The New Student Transcript

Several institutions are developing new records intended to supplement, rather than replace, the traditional transcript. In particular, two initiatives, are coordinating such projects. The Comprehensive Student Record Project, funded by the Lumina Foundation and led by the American Association of Collegiate Registrars and Admissions Officers and NASPA: Student Affairs Professionals in Higher Education, has coordinated twelve diverse institutions to rethink how they represent student learning and how they share the student record.[47] At the same time, the IMS Global Learning Council is working with five institutions participating in Lumina's Competency-Based Education Network to create a digital "new transcript for the 21st-century." This record will be student-centered, competency-based, secure, and can integrate information from learning management systems, student information systems, and other disparate sources.[48]

University of Maryland University College (UMUC), which focuses on serving adult learners, is a member of both initiatives. As it shifts its curriculum to a competency framework, UMUC is developing a digital transcript that shows progress on designated competencies, as well as the courses or experiences through which the student learned the competency, performance on assessments, and associated artifacts. UMUC's new transcript will also allow students to add sources from which they accumulate

---

[47] See "Comprehensive Student Record Project," American Association of Collegiate Registrars and Admissions Officers, http://www.aacrao.org/resources/record. Participating institutions are: Elon University, Indiana University – Purdue University Indianapolis, Quinsigamond Community College, Stanford University, University of Houston-Downtown, University of Maryland University College, University of South Carolina, and University of Wisconsin Colleges and University of Wisconsin – Extension. For more information on the project, see also, "Summary of Comprehensive Student Records Project Convening," ACCRAO & NASPA (October 28-29, 2015), http://www.aacrao.org/docs/default-source/Lumina/aacrao-lumina_executive-summary_v4.pdf?sfvrsn=0. For a broader overview, see "Tomorrow's Transcript: Effort to Reform sSudent Records Puts Learning at the Core," Lumina Foundation Focus Magazine (Fall 2015), http://focus.luminafoundation.org/archives/fall2015/.

[48] "Digital Credentialing," IMS Global Learning Consortium, https://www.imsglobal.org/activity/digital-credentialing.

competencies, such as previous educational or professional experiences, and would indicate whether a competency is constructed from registrar verified sources, student sources, or both.[49] Stanford University, which serves a very different demographic than UMUC and has a more traditional curricular model, is also in the process of designing a verifiable, digital transcript called the "certified electronic verified certificate," that would organize coursework and represent student learning by learning outcomes in a standardized format.[50]

Other institutions are redesigning their transcripts around experiences, rather than competencies. Elon University, for example, has since 1994 been capturing and validating students' co-curricular experiences such as leadership, service, research, and study abroad on a supplemental "Experiences Transcript." While the institution has historically represented these experiences in a list of activities with dates and hours spent, it is developing a prototype that represents categories of experience as infographics and on a timeline. The digital-only document is an official document of the university, so students cannot enter information into the transcript on their own. Rather, students submit evidence of their participation in a program to the university, which validates the experience and maintains information about it in a centralized system.[51]

A final category of the extended transcript is the contextualized transcript, such as the one the University of North Carolina, Chapel Hill (UNC) has developed. UNC's transcript compares a student's course grades to average grades and percentile rank for students in the class, as well as the grades of students taking similar courses.[52]

---

[49] "eT, CSR, CBE and the Role of the Registrar: Providing Evidence of the Student's Learning Journey," presented by and provided the University of Maryland University College Registrar's Office. See also, Carl Straumsheim, "Transcript of Tomorrow," *Inside Higher Ed* (February 29, 2016), https://www.insidehighered.com/news/2016/02/29/u-maryland-university-colleges-extended-transcript-new-type-student-record.

[50] See "A Student Record a Student Can Love," AACRAO (April 19, 2016), http://www.aacrao.org/resources/resources-detail-view/a-student-record-a-student-can-love.

[51] "Elon Experiences Transcript," Elon University, http://www.elon.edu/e-web/students/elon_experiences/transcript.xhtml. For the new visual transcript, see "Visual Experiential Transcript," http://www.elon.edu/e-web/students/elon_experiences/VisualEXP.xhtml. See also "Growing Student Records Beyond the Traditional Transcript," AACRAO (March 8, 2016), http://www.aacrao.org/resources/resources-detail-view/growing-student-records-beyond-the-traditional-transcript.

[52] "Providing Context for the Contextualized Transcript: A Case Study," AACRAO (June 12, 2015), http://www.aacrao.org/conferences/conferences-detail-view/providing-context-for-the-contextualized-transcript--a-case-study

E-Portfolios

E-portfolios provide a means for students to organize, archive, and display digital artifacts of their learning and experiences, which can be shared with educators and potential employers. In 2010, more than half of private universities and nearly 40 percent of public universities reported that they used e-portfolio services for their students.[53]

E-portfolios have both internal and external uses, and like some versions of the extended transcript, give students more agency over the representation of their learning, provide more detail on how students gained skills and knowledge, and can represent student achievement across multiple contexts, rather than solely in institution-offered courses.[54] Some institutions, like LaGuardia Community College and Boston College, use e-portfolios as pedagogical tools, prompting students to reflect on their learning across contexts and semesters.[55] Another common use of e-portfolios is internal learning outcomes assessment; instructors at institutions like CUNY's Guttman Community College and Alverno College use rubrics to assess artifacts in students' e-portfolios and discuss their findings in a process to refine curriculum and pedagogy. Alverno College also encourages students to follow their own learning progress, review assessment feedback, and look for patterns in their academic work that might help them improve.[56]

Finally, like extended transcripts, e-portfolios can be used to supplement resumes and traditional transcripts for employment. Portfolium, an e-portfolio service used by more

---

[53] "The 2010 National Survey of Information Technology in U.S. Higher Education," The Campus Computing Project (October 2010,) http://www.campuscomputing.net/sites/www.campuscomputing.net/files/Green-CampusComputing2010.pdf. See also George Lorenzo and John Ittelson, "An Overview of E-Portfolios," Educause Learning Initiative (July 2005), https://net.educause.edu/ir/library/pdf/ELI3001.pdf; "ECAR Study of Undergraduate Students and Information Technology," Educause (August 2015), https://library.educause.edu/resources/2015/8/~/media/24ddc1aa35a5490389baf28b6ddb3693.ashx.

[54] For a useful framework through which to understand e-portfolio's many uses and their institutionalization, see the Catalyst for Learning's ePortfolio Resources and Research, available at http://c2l.mcnrc.org/. The website also includes a list of institutions with robust e-portfolio initiatives.

[55] For a deeper discussion of uses of ePortfolio, see "LaGuardia and the ePortfolio Field," http://eportfolio.lagcc.cuny.edu/about/field.htm.

[56] For more information on Guttman's outcomes assessment program, see Jessie Brown and Martin Kurzweil, "Student Success by Design: CUNY's Guttman Community College," Ithaka S+R (February 4, 2016), http://www.sr.ithaka.org/publications/student-success-by-design/. For more on Alverno College's Diagnostic Digital Portfolio, which is perhaps the most longstanding example of the use of ePortfolios within an assessment-as-learning" context, see "Diagnostic Digital Portfolio," Alverno College, http://www.alverno.edu/ddp/.

than 2,000 postsecondary institutions, offers an interface through which students can use their e-portfolio to apply for jobs that match their represented skills.[57]

Micro-credentials and Badges

Badges are validated, visual representations of student accomplishment that can be shared easily and securely through social media, e-portfolios, extended transcripts, or other digital platforms. Most badging platforms also contain metadata about how a student earned the badge. Badges allow learners to represent learning not included on a traditional transcript, and are particularly aligned to competencies or experiences other than formal academic experiences.[58] Like e-portfolios, badges can also play a formative role in student learning; some research indicates that badges can increase student motivation by offering short-term, visible recognitions of their work.[59]

Though digital badges are most often associated with online or extra-institutional environments, a growing number of institutions are also awarding digital badges to represent and acknowledge student accomplishments, often on a smaller scale than the degree. For example, Penn State University offers a number of badges in areas like Liberal Arts-Digital Citizenship, Information Literacy, Emerging Leadership in Online Learning, and Media Commons. The Liberal Arts-Digital Citizenship badge recognizes experiences like studying abroad, participating in international student group events, or engaging in other travel opportunities.[60] The University of Notre Dame offers badges for a similar set of co-curricular experiences, and uses an integration between Credly (its badging platform) and Digication (its e-portfolio platform) to allow students to display badges in their e-portfolios.[61]

---

[57] See the Portfolium website, https://portfolium.com/educators.

[58] For an overview of digital badging, see Carla Casilli and Erin Knight, "7 Things You Should Know About Badges," Educause (Monday, June 2012), https://library.educause.edu/~/media/files/library/2012/6/eli7085-pdf.pdf; "What is a Digital Badge," HASTAC, https://www.hastac.org/initiatives/digital-badges.

[59] Samuel Abramovich, Christian Schunn, Ross Mitsuo, "Are Badges Useful in Education?: It Depends Upon the Type of Badge and Expertise of the Learner," *Education Teach Research Development* (March 17, 2013), http://www.lrdc.pitt.edu/schunn/research/papers/Abramovich-Schunn-Higashi.pdf.

[60] See Digital Badges at Penn State at http://badges.psu.edu/.

[61] "E2B@ Badge Directory," ePortfolio@ND, http://eportfolio.nd.edu/directory/badge-directory/. See also, "Digital Badging a Conference on Digital Portfolios and Badges," Research and Assessment for Learning Design Lab News (July 13, 2016), https://blogs.nd.edu/real/.

Representation Infrastructure

Though some institutions have been using e-portfolios or extended transcripts for decades, efforts to rethink the student record are dynamic and multi-directional. There are a number of concurrent efforts to ground the work in a consistent set of guiding principles, conceptual frameworks, and technological platforms across institutional contexts.

As discussed above, many of the most sophisticated extended transcript initiatives are part of an Lumina-funded and AACRAO and NASPA-led initiative that aims to accelerate the creation of extended transcripts, coordinate and document efforts, and develop a framework for institutions involved in this work. AACRAO's framework, still in draft form, calls for institutions to think carefully about the experiences to include in the transcript; to ensure that transcripts are available in a digital format; and to supplement—rather than replace—the traditional transcript.[62] The IMS Global Learning Consortium, a newer player to this sort of coordination work than ACCRAO, has outlined a more detailed set of technical specifications for competency-based transcripts, and requires that these records must be digitally verifiable and secure so that they are only viewable by authorized recipients. The IMS specifications also imagine an easily sharable public component of a competency-based credential, much like a badge.[63]

A number of vendors provide the technical infrastructure to support these kinds of representation innovations. Parchment, a credential technology company, supports Elon's Experiences Transcript, and is actively engaged in rethinking credentialing and how to support it for other institutions. Parchment has established its own framework for innovations in credentialing and transcripts, and has identified five ways in which institutions can alter transcripts to contribute to "learner-empowerment."[64]

E-portfolios are more established components of institutional practice than extended transcripts, and therefore there are a number of organizations and research efforts that

---

[62] "Summary of Comprehensive Student Records Project Convening," ACCRAO & NASPA (October 28-29, 2015), http://www.aacrao.org/docs/default-source/Lumina/aacrao-lumina_executive-summary_v4.pdf?sfvrsn=0.

[63] "Enabling Better Digital Credentialing," IMS Global Learning Consortium, https://www.imsglobal.org/initiative/enabling-better-digital-credentialing.

[64] Briefly, these are: go digital, do what paper can't, create new pathways (e.g. ensure portability), communicate more content, and make it actionable or easily embeddable into social media sites. Each of these approaches resonates with the examples that we have discussed. See "Extending the Credential, Empowering the Learner," Parchment (2015), http://info.parchment.com/rs/parchment/images/ExtendingCredentialsWhitePaper.pdf.

provide best practices and "connective" ligaments for institutions that use e-portfolios.[65] Additionally, there is a marketplace for e-portfolio platforms, some of which are more conducive to some purposes than to others. Portfolium is a newer player in the e-portfolio market and has been primarily adopted by career services and alumni associations. Other leading e-portfolio vendors include those with strong assessment management features such as LiveText, Taskstream, FolioTek and Chalk & Wire in addition to professional portfolio features. Digication is primarily e-portfolio-focused and has broad adoption across the U.S. in institutions such as CUNY, the Rhode Island School of Design, Stanford, SUNY, and the University of Alaska, Anchorage. PebblePad, out of the UK, has a personal learning environment emphasis and has been adopted by Duke and Portland State University as well as institutions in the UK and Australia. Finally, Seelio and Pathbrite have a career-oriented focus and a strong visual interface. Open source tools include Mahara out of New Zealand and Karuta which is the latest iteration of the Open Source Portfolio (Sakai).[66]

These vendors offer institutions guidelines and tools to manage challenges such as the integration of information from student information systems and learning management systems, privacy and intellectual property protection, assessment, and long-term maintenance. However, these guidelines have largely emerged iteratively from practice and observations, rather than thoughtful consideration of the short-term and long-term implications of potential decisions and policies. Therefore, the need for a normative framework is every bit as salient in uses of e-portfolio and other innovative forms of representation as it is for other areas of student data use.

The key practical challenge for badging is cross-contextual integration, so that students can embed badges in portfolios, social media platforms, or other learner-centric digital records. The Humanities, Arts, Science, and Technology Alliance and Collaboratory (HASTAC) has documented badge design principles that emerged from the Macarthur Foundation's Digital Media and Learning Program (which funded a number of badging initiatives).[67] Included in these principles are recommendations that badges map learning trajectories (e.g. fit in with larger goals or experiences); align to standards or

---

[65] These include The Association for Authentic, Experiential & Evidence Based-Learning (http://www.aaeebl.org/); ePortfolio Action & Communication (http://epac.pbwiki.com); Inter/National Coalition for ePortfolio Research (http://ncepr.org/); and International Journal of ePortfolio Research (http://www.theijep.com).

[66] We are indebted to Helen L. Chen, Senior Researcher in the Designing Education Lab and Director of ePortfolio Initiatives at Stanford University, for helping us build this list.

[67] Andi Rehak and Daniel Hickey, "Digital Badge Design Principles for Recognizing Learning," HASTAC (May 20, 2013), https://www.hastac.org/blogs/andirehak/2013/05/20/digital-badge-design-principles-recognizing-learning. For more on the MacArthur Foundation's Digital Media and Learning Program, see https://www.macfound.org/programs/learning/.

learning outcomes; are issued by experts; are externally endorsed; recognize diverse learning; can be used to externally communicate knowledge and skills; and can be linked to formal academic credit when appropriate.

On the more technical end, Mozilla Foundation has developed an open badging technical specification that standardizes the way that badging information is packaged, made portable, and verified. The specification is currently being refined by the Badge Alliance, a collection of working groups dedicated, more broadly, to refining badging infrastructure and ecosystem issues such as standards, data, research, and endorsement.[68] There are a number of third-party providers, such as Credly and plug-ins for Blackboard, Moodle and Canvas, which build on Mozilla's open badging specification to allow educators and institutions to build and verify badges.[69]

Comparing the different forms of representation to one another also presents a challenge. Since the late nineteenth century, institutions have used the Carnegie Unit as a discrete, standard and transferable unit of learning. While financial aid and other regulations will necessitate that most institutions continue to structure and recognize learning through this framework, the growing scope of what is recognized as learning, and the expanding landscape of providers, also requires that institutions look to other measures to integrate learning.[70] To that end, the Lumina Foundation, building on efforts to establish learning outcomes as interchangeable units of learning, has outlined a beta framework for "connecting credentials" across multiple institutional contexts, and funds initiatives that translate various credentials into a common set of competencies.[71] A group of community colleges engaged in the "Right Signals" initiative is implementing Lumina's credentials framework to recognize and verify the multiple credentials from various issuers that their many non-traditional students have accumulated.[72]

---

[68] For Mozilla Open Badges, see http://openbadges.org/ and Erin Knight and Carla Casilli, "Case Study 6: Mozilla Open Badges," Educause (May 2, 2012), https://library.educause.edu/resources/2012/5/case-study-6-mozilla-open-badges. For the Badge Alliance, see http://www.badgealliance.org/.

[69] For a list of platforms that use the Mozilla specification, see Bernard Bull, "Want to Issue Open Badge? Here are some options," *Etale-Digital Age Learning* (July 27, 2014), http://etale.org/main/2014/07/27/want-to-issue-open-badges-here-are-some-options/.

[70] For an in-depth discussion of this development, see Ethan Hutt, "A Brief History of the Student Record: A Paper for the Asilomar Conference on Student and Records in the Digital Era (Draft)," (May 31, 2016), included in Asilomar pre-reading material.

[71] "Connecting Credentials: A Beta Credentials Framework," Lumina Foundation (June 11, 2015), https://www.luminafoundation.org/resources/connecting-credentials.

[72] "The Right Signals Initiative," American Association of Community Colleges, http://www.aacc.nche.edu/Resources/Pages/right_signals.aspx. For more information on the project guidelines, see "Request for Proposals: The Right Signals Initiative," American Association of Community Colleges, http://www.aacc.nche.edu/Resources/Documents/Right_Signal_RFP2.pdf.

Responsible Use

Expanding what is represented, and who controls what is represented, requires tradeoffs among quality, accuracy, authenticity, and student agency and privacy. The external-facing nature of the student record requires some coordination and standardization in how these tradeoffs are addressed. While FERPA established students' rights to inspect, review, amend, and exercise some control over the disclosure of information from their educational records, the boundaries of what constitutes students and educational records have blurred with the growing use of online courses and platforms. Additionally, as is the case in our discussions of research and application of student data, the "responsible use" of student data expands beyond concerns regarding privacy. Institutions must also consider how to minimize adverse effects, promote equitable outcomes, and maintain student autonomy and integrity as they circulate student information. The proper role of government, institutions, and other organizations in establishing these principles is not yet well-defined.[73]

This new landscape raises a number of questions. One set of questions relates to what should be included in the student record, and how to weigh the potential benefits of including more information with the potential violations of student privacy. For example, it is not inconceivable that new student records could contain the sort of micro-level data that are used in some of the research and application initiatives discussed in this paper, giving students, instructors, or employers access to information about student learning styles, mindset, and motivation. Should this sort of information be part of the student record? If so, with whom should it be shared? To what extent should students decide what is included and what is made public? Because the boundaries of what should be included on the student record are still being defined (and may vary depending on contexts), the answers to these questions remain malleable, and the status of this information remains unclear under FERPA.[74]

Related questions have to do with how information from new student records can be used. Can data contained in student records and shared with outside organizations be used for predictive analytics? For research? How much control should the issuer or

---

[73] For a sustained inquiry into ethical issues regarding badging, see James E. Willis III, Joshua Quick, and Daniel T. Hickey, "Digital Badges and Ethics: The Uses of Individual Data in Social Contexts," in *Proceedings of Open Badges in Education (OBIE 2015) Workshop,* Poughkeepsie, New York (March 16, 2015), http://er.educause.edu/articles/2016/4/microcredentials-and-educational-technology-a-proposed-ethical-taxonomy.

[74] See Department of Education, "Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices," (February 2014), https://tech.ed.gov/wp-content/uploads/2014/09/Student-Privacy-and-Online-Educational-Services-February-2014.pdf.

student have over these uses? Mozilla, for example, shares and uses information collected in badges to improve its products, but de-identifies this information for research purposes. When it shares information with employers, contractors, and service providers, it does so only if those parties agree to handle the data in accordance with Mozilla's privacy policy.[75] As institutions partner with third-party platforms for new transcripts, ePortfolios, and badging, establishing normative guidelines about how data can be used will be crucial.

A final set of questions relates to authentication, verification, and ownership. Should institutions be the sole verifiers of student experiences? As discussed, UMUC allows students to add in their own sources of learning, and indicates on the transcript which sources are student added and which are institutionally verified. Elon, on the other hand, requires that all experiences represented on a transcript be verified by the institution. Maintaining the institution as the sole verifier of learning experiences adds credibility to what is represented, but it also excludes valuable experiences that only the student, or some other party, can verify.

Conclusion

As datasets about students and student learning grow larger and more granular, and as the technical capacity to merge data from disparate systems grows, there is a mounting expectation that institutions, researchers, and other stakeholders will use these data to improve student outcomes and student learning. Researchers are leveraging increasingly fine-grained data to gain insight into student learning and instructional effectiveness; administrators, advisors, and instructors can proactively address barriers to completion; and students and institutions have the opportunity to represent student learning in increasingly expansive and personalized ways.

Yet there are cultural, technical, and coordination challenges that currently limit the realization of these opportunities. Interest in a scientific approach to learning is only emerging, and few institutions provide incentives for researchers, instructors, and other staff to incorporate this approach into their work. Technical capacity for the new forms of research, application, and representation remains uneven. Although a growing number of third-party providers offer institutions tools to leverage their data, these organizations operate under different rules and norms than institutions of higher education, and most institutions are still working through the details of relationships

---

[75] "Privacy & Mozilla Badge Backpack," Mozilla Backpack, https://backpack.openbadges.org/privacy.html#sharing.

with these providers. Moreover, while the use of student data has necessitated some erosion of inter-institutional and intra-institutional silos, there is a growing need for standards and collaboration in the field, and unresolved questions about which bodies should be responsible for coordination.

Finally, this unprecedented capacity to collect, analyze, merge, and act on student information poses a number of cross-cutting questions related to responsible use, many of which are not clearly addressed by existing frameworks like those governing privacy or human subjects research. Our review surfaced several—sometimes convergent—efforts to define responsible use, but few widely shared or articulated principles through which these practices can be defined, refined, or reviewed.

The coordinating efforts we do discuss in this paper—from the multiple initiatives that make research with large and diverse datasets possible to the AACRAO's coordinating role in efforts to redefine the student record—build on a long tradition of cross-institutional coordination and capacity building, led by non-profit organizations, government agencies, and institutions themselves. Some of the enduring—though perhaps outdated—infrastructure we discuss in this report, such as IRB regulations and the Carnegie unit, have emerged from these coordinating efforts. As the use of student data for research, application, and representation grows, each field can leverage the relationships, processes, and standards already in existence. However, the field must also chart out new territory for innovation and integration across practices, while maintaining enough flexibility to evolve in a rapidly shifting landscape of institutional capacity, student needs, and technical opportunities.

Appendix A

This report was created to support a June 2016 convening on Student Data and Records in the Digital Era, co-hosted by Ithaka S+R and Stanford's Center for Advanced Research through Online Learning. We developed the scope and gathered contextual information and guidance on selecting examples with guidance from our co-planners and several advisors working on the convening, including Tom Black, Helen L. Chen, Mitchell Stevens, and Candace Thille of Stanford, Ken Koedinger of Carnegie Mellon University, Tim McKay of the University of Michigan, and Sharon Slade of Open University UK.

Through this consultation and our own desk research, we identified more than a dozen well-developed research, application, and representation initiatives for a deeper dive. In each case, these initiatives fit our criteria of using large-scale, technology-enabled sets of granular data on learners and learning, either generated through the interaction of students and higher education institutions or with relevance to practice or the policy environment for higher education institutions. We purposely selected initiatives based in or focused on a variety of institution types (public, private, four-year, two-year, etc.) and included both intra- and inter-institutional efforts.

After conducting extensive additional desk research on selected initiatives, we gathered information from participants in those initiatives through email communications and phone interviews, during which we focused on the major features of the initiative, technical and practical challenges, and concerns and practices surrounding responsible use. We also asked them to reflect on these issues for their field more broadly. We corresponded with or interviewed the following individuals:

- Alex Kindel, Stanford University (DataStage)

- Beth Davis, Predictive Analytics Reporting Framework (PAR)

- Carolyn Connerat, University of Texas at Austin

- Celeste Schwartz, Montgomery County Community College

- Clint McElroy, Central Piedmont Community College (Online Student Profile)

- Josh Baron, Lumen Learning, formerly Marist College (Open Academic Analytics Initiative)

- Ken Koedinger and Gail Kusbit, Carnegie Mellon University (DataShop, LearnSphere)

- Leah Lommel, Arizona State University (eAdvisor)

- Matthew Pistilli, Indiana University – Purdue University Indianapolis (Purdue University's Purdue Course Signals)

- Oliver Ferschke, Carnegie Mellon University (DiscourseDB)

- Shannon McCarty, Rio Salado University (RioPACE)

- Sharon Slade, Open University UK

- Una-May O'Reilly, Massachusetts Institute of Technology (MOOCdb)

In addition to this initiative-specific research, we reviewed the literature on each category of activity (research, application, and data), and on privacy, human subjects research, and ethics in the context of student data. Throughout the report, we cite published sources in footnotes; we do not cite our interviews or correspondences.