Ethical considerations in using student data in an era of 'big data'

Paul Prinsloo and Michael Rowe

Abstract

Learning with technology enables the collection of data on students at a level unprecedented in face-to-face tuition and paper-based academic administration. Universities see the advantage in tracking students' engagement and progress, particularly when it comes to putting interventions in place for at-risk students. Our use of these data should be legal, ethical and seen as fair use by students. At no time should it cross the boundaries into the realm of 'creepy', a word used by Gartner analyst Frank Buytendijk in several of his presentations at the Gartner ITExpo in Cape Town in September 2014.

Recommendation

It is recommended that:

1. higher education engages critically with the potential and perils of big data;

2. the collection, analysis and use of student data be understood as a moral practice and duty;

3. students' agency and participation in the collection, analysis and use of their data are recognised and protected;

4. student identity and performance are accepted and analysed as temporal, dynamic constructs;

5. student success is interpreted as a complex and multidimensional phenomenon; and

6. higher education institutions commit to the transparent collection, analysis and use of student data.

These recommendations are explained in detail below.

Discussion and analysis

It is difficult, if not impossible, to assess the scope and permanency of the changes facing higher education (Altbach et al. 2009). While many factors contribute to the dynamic higher educational context, the impact of technological advances on the curriculum cannot (and should not) be underestimated. Amid these changes, higher education institutions (HEIs) are increasingly accountable to more stakeholders than ever before and must provide evidence of the institutional processes linked to student success, along with the need to optimise the collection, analysis and use of data that serve to inform decision making. It is important to remember that, fundamentally, the collection, analysis and use of student data are activities to improve students' learning and their chances for success (Booth 2012).

As higher education increasingly integrates digital systems into teaching and learning processes, the amount of data generated increases exponentially. As data increase in volume, velocity, variety, scope, resolution, flexibility, scalability and indexical properties (Kitchen 2014), so do the extent and complexity of the associated ethical issues and challenges (Slade and Prinsloo 2013). So, while big data in the educational context are imbued with almost magical properties by many academics – together with mainstream media, some authors are calling for a more skeptical and critical approach to its use. Siemens and Long (2011) describe the need to use data for (better) decision making as one of the most dramatic factors shaping the future of higher education. It is becoming increasingly clear that, while quantitative research and analysis provide much-needed information, it is primarily information about 'what is happening' and not necessarily 'why'. Qualitative data, on the other hand, explains *why*. They uncover the behaviour and motivation behind those actions.

In 2014, the Open University published its policy on the ethical use of learning analytics (Open University 2014), becoming one of the first universities to articulate this important perspective. There are still very few HEIs that have responded to the fast-changing data environment with its different legal and ethical issues (Prinsloo and Slade 2013), with current data protection regimes failing to address the complexities and different nuances of an ethical use of student data (Prinsloo and Slade 2015). For example, while most HEIs address the issue of the use of student data for research purposes, very few explicitly inform students that their behaviour may be monitored or provide students with the opportunity to opt out of these processes. In research contexts, students may opt out as objects of research, and institutional ethical clearance processes make informed consent and anonymity non-negotiable (unless the latter is waived by the participant himself or herself). (See, for example, a recent and fairly comprehensive overview of the ethical issues and research into the ethical issues in learning analytics by Niall Sclater (Sclater 2015).)

Stewart (2014) writes: 'Thanks to the proliferation of personal computers, smartphones and wearables, we generate 2.5 quintillion bytes of data a day. That means, every two days, human beings create more data than they did from the dawn of civilisation up until 2003’. Kitchen (2014) states that data have changed from being a scarce commodity to a situation where the ‘production of data is increasingly becoming a deluge; a wide, deep torrent of timely, varied, resolute and relational data that are relatively low in cost and, outside of business, increasingly open and accessible’ (Kitchen 2014:xv). Big data, as a phenomenon, are therefore characterised by increased volume, the possibility to examine and profile entire populations instead of samples, fine-tuned resolution, strong relationality, increased velocity, wide variety, and high flexibility and scalability (Kitchen 2014). As HEIs move increasingly online and into the digital space, the characteristics of big data will add levels of complexity never before seen in this context.

It is therefore crucial that we take note of the warning that we should not accept the simplistic premise that ‘more data = better data’ at face value. Data are not neutral, raw or exempt from being shaped and framed by technical, ethical, economic, philosophical and political interests. It is impossible to separate data from the epistemologies, contexts, assumptions and power relations that were used to select, process and analyse them (Gitelman 2013; Kitchen 2014). There is increasing consensus that the way data are ‘ontologically defined and delimited is not a neutral, technical process, but a normative, political and *ethical* one that is often contested and has consequences for subsequent analysis, interpretation and action’ (Kitchen 2014:19; emphasis added).

Most authors agree that we are, indeed, facing a data revolution, one that higher education cannot and should not ignore. HEIs increasingly have access to disparate data sources, inside and outside the conventional boundaries of student information and learning management systems (LMSs) (Prinsloo et al. 2015). This raises ethical concerns and questions such as the fact that, firstly, students have not provided consent to harvest information outside the scope of the learning contract with the institution and, secondly, the danger of context-collapse. Higher education may therefore make decisions regarding students’ applications for access based on digital assemblages without students knowing the reasons or rationale (Solove 2004).

Recommendation 1: *Higher education engages critically with the potential and perils of big data*. There is no question that higher education should collect, analyse and use students’ digital data (Slade and Prinsloo 2013). It would be irresponsible and unethical for higher education not to do so, as HEIs are accountable for the effectiveness and efficiencies of the programmes and resources used (Prinsloo and Slade 2014). However, we must remain critical and even skeptical regarding our ontologies, epistemologies and assumptions regarding data. ‘We need to be cognisant of the impact and unintended consequences of our assumptions underpinning the algorithmic turn in higher education’ (Prinsloo et al. 2015:294) (also see Danaher 2014; Napoli 2013), as ‘bad use of data can be worse than no data at all’ (Kakaes 2015.) It is also crucial that we shy away from the belief that our students' digital profiles or footprints present a complete picture of their potential and the challenges they face. Several authors (Cloggy 2011; Duval nd) have cautioned that learning analytics can very easily serve to bureaucratise students' learning even further, or serve a panoptical purpose and culture of increasing surveillance rather than empowering students and their institution to facilitate more appropriate choices.

Recommendation 2: *The collection, analysisand use of student data be understood as a moral practice and duty*. Amid the many constraints and challenges facing higher education, it is crucial that we are reminded, especially in the South African context, of education as a *moral* practice. While higher education cannot ignore the need for accountability and increased effectiveness and efficiency, we should, in equal measure, be concerned about the *appropriateness* of our curricula, assessment methodologies and pedagogies. As Biesta (2007) has warned, something may be very effective and efficient without being moral. In the context of the collection, use and analysis of student data, it is therefore important to realise, firstly, that increased knowledge about our students increases our responsibility to them and, secondly, to use our increased knowledge and understanding to serve *learning* (Gašević and Siemens 2015). We should not assume that knowing more about our students may, necessarily, result in more just decisions. There is ample evidence of how increased surveillance and gathering of personal information can actually result in unjust and unfair decisions and the marginalisation of those who are already vulnerable (Henman 2004).

Recommendation 3: *Students' agency and participation in the collection, analysis and use of their data are recognised and protected.* Students are not passive recipients of services, but active agents in a reciprocal relationship with the institution. Students' agency with regard to the collection, analysis and use of their data is much wider and more nuanced than simply needing to provide consent. Learning analytics should be student-centric in that it should not conceive students as data objects. Rather, our policies and frameworks must allow students to have access to, and permission to edit, additional contextual information associated with the 'hard' data that will allow HEIs to have a more holistic picture of students and their learning. Students' learning journeys are more than the number of clicks, logins or time spent looking at online content. What possibilities emerge when students can edit the data that institutions gather about them? Should students be able to set the permissions on which data sets the institutions can use and for what purposes? (Kruse and Pongsajapan 2012; Prinsloo and Slade 2014). These, and other questions, should serve to guide decision making about students' role in the use of their data by HEIs.

Recommendation 4: *Student identity and performance are accepted and analysed as temporal, dynamic constructs*. Learning analytics can provide snapshots of individual students at a particular moment in time, but often with no context. 'Students should be allowed to evolve, and adjust and learn from past experiences without those experiences becoming blemishes on their development history' (Slade and Prinsloo 2013:1520). Students' digital records should not be permanent 'tattoos' that follow them for the rest of their lives (Mayer-Schönberger 2009:14) and data collected should therefore have an agreed-upon life span and expiry date, perhaps even determined in collaboration with the student.

Recommendation 5: *Student success is interpreted as a complex and multidimensional phenomenon.* Student success is the result of 'mostlynon-linear, multidimensional, interdependent interactions at different phases in the nexus between student, institution and broader societal factors' (Prinsloo 2012). In the context of big data and learning analytics, we shall therefore have to move beyond claims of causality and rather attempt to understand relationships between different variables at different points in a student's journey. One of the big promises of big data is the increasing use of algorithmic decision making. What happens when algorithms – supposedly neutrally coded – make choices that reflect social contexts that are inherently biased? For example, if an algorithm is tested or trained using baseline data that reflect an existing bias towards a minority, the algorithms will tend towards that same bias. What are the implications for this in education? What are the challenges that emerge when our 'neutral' code makes unfair decisions about students? This might, for example, have an impact on admissions processes, as the systems for student admissions become increasingly dependent on computer ranking and other methods for sorting through applications (Danaher 2014; Henman 2004).

Recommendation 6: *Higher education institutions commit to transparent collection, analysis and use of student data.* It is becoming clear that different data sources are harvested and combined, often without regard to the original context (Kitchen 2014). This 'context-collapse' that takes place amid increasing concerns regarding pervasive surveillance and privacy (Prinsloo 2014) highlights the issue of increasingly asymmetrical power relationship between students and institutions (Davis and Jurgenson 2014; Vitak 2012). Therefore, HEIs should be transparent about the type of data collected, for what purposes, by whom, and the measures that will be taken to protect individuals' identities. It is also increasingly apparent that students should be informed about the implications of their sharing of personal information in other online contexts – perhaps those that are not even related to their academic lives – and how this information may be used by higher education to make decisions around access, curricula and support within the academic context. It is therefore clear that the notion of informed consent is more nuanced than thinking in terms of the binary of opting in *or* out (Prinsloo and Slade 2015). There is also increasing concern about the re-identification of de-personalised data (Tene and Polonetsky 2012).

It is also crucial that HEIs in South Africa engage with the Protection of Personal Information Act, Act 4 of 2013 (Government of South Africa 2013) to ensure compliance.

Current trends internationally

Although there is an increase in theoretical and conceptual research regarding the ethical collection, analysis and use of student data in an era of big data, there are very few current examples of how institutions respond to the ethical challenges and issues (Slade and Prinsloo 2013; Prinsloo and Slade 2015). The ground-breaking work of the Open University in this regard may point the way to how to approach the ethical collection and use of student data (Open University 2014), taking into account the recommendations presented above.

Affordances

Despite and amid a sobering reflection on the realities facing the realisation of the potential of learning analytics in higher education, it is clear that the appropriate and intelligent collection, combination, analysis and use of student data can inform and increase the effectiveness and efficiencies of higher education (Siemens and Long 2011; Prinsloo and Slade 2014).

Costs: licensing, infrastructure, personnel

The costs, infrastructure and personnel involved in ensuring the ethical collection and use of student data relate mostly to policy and staff development, and possibly operational structures and processes to oversee the implementation of policy and regulatory initiatives. If a commercial data analytics program is used, licensing will become a factor. Most HEIs use some form of business intelligence program that might or might not be built into the general IT provision. Blackboard users have the Retention Centre for basic analytics or can acquire Blackboard Analytics for Learn. Online software comes with licensing cost implications. Possibly the greatest gap in the system is the availability of data analysts and data scientists, people who can interpret data and make actionable recommendations.

Application in different contexts in South Africa

All public higher education institutions collect data for reporting to the Department of Higher Education and Training (HEMIS data). These data are normally highly aggregated so individual student information or formative interventions to improve pass rates, for instance, are not the focus. Of course, the underlying demographic and module success data exist and could be extracted, integrated and manipulated using business intelligence tools. In many ways, though, HEMIS data are summative and not that useful for student success in a formative phase, although historical trends might be useful in predictive algorithms, with all the inherent dangers discussed above.

Many universities enable their students to apply online and interact online for all their academic administration. Student interactions in this environment can be tracked. Most universities also provide an online LMS for their students and data are produced each time the student enters the system. Some universities or individual modules might use Facebook, Twitter, Google or any of a number of web-based technologies, and everything leaves a trace. To what extent can data generated by social media be ethically integrated with other data to profile students?

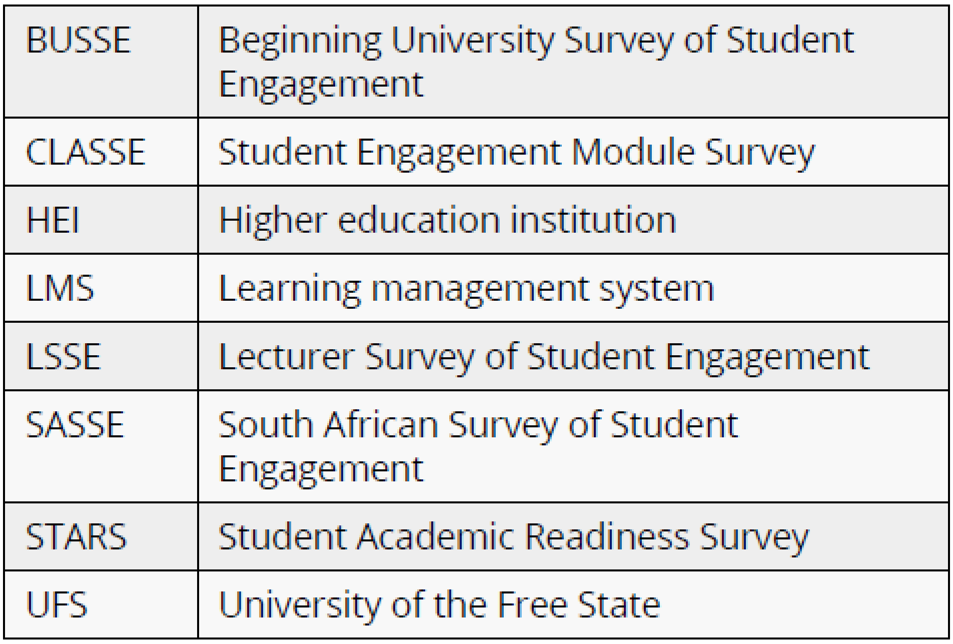
Many South African institutions are already experimenting with home-grown or commercial online student tracking systems that will identify student engagement while they are learning in order to put timely interventions in place.

One example is the use of Blackboard's Analytics for Learn at the University of Pretoria. The most effective use of these data would be to alert students themselves to the fact that they are not engaging sufficiently or achieving well enough relative to fellow students. The alert could be accompanied by recommendations to interventions that are in place, such as tutorial groups. A South African Survey of Student Engagement (SASSE) was developed at the University of the Free State (UFS) and piloted a couple of years ago with the support of the Council on Higher Education. It is based on the original survey in the USA that has also been contextualised for countries such as Australia, Hong Kong and Ireland. There is a companion lecturer survey (LSSE). The SASSE was relaunched in 2014 and administered online. The UFS is busy engaging with institutions that participated on the results. Each institution receives its own raw data and care should be taken in its responsible and ethical use. Student surveys are popular instruments for gathering data on students. For instance, in the SASSE suite, there is a survey for students just starting (BUSSE), as well as a module survey (CLASSE). The University of Pretoria has its own Student Academic Readiness Survey (STARS), implemented during the registration period. Students selfidentify aspects such as study skills, time management, family support, etc. As a result of the survey, students might be allocated a mentor or be referred to faculty student advisors for support. Each student receives his or her own profile online as well.

Glossary

The primary concept to understand is 'big data' and how it differs from the data that we might have gathered on students in the past. It is the wealth of data generated by the use of technology for administrative or learning purposes. The ethical collection and use of student data in an era of big data are inherently multi- and intra/interdisciplinary. Therefore, it is almost impossible (and possibly counterproductive) to attempt to define the different concepts and terms used in the disparate discourses. The Open University (2014) briefly describes learning analytics, defines an intervention, as well as data, and specifically sensitive data, and the notion of informed consent. The list of references and readings below provides a rich source of information. Interested individuals and institutions are invited to consult the reading list.

Acronyms and abbreviations



BUSSE	Beginning University Survey of Student Engagement
CLASSE	Student Engagement Module Survey
HEI	Higher education institution
LMS	Learning management system
LSSE	Lecturer Survey of Student Engagement
SASSE	South African Survey of Student Engagement
STARS	Student Academic Readiness Survey
UFS	University of the Free State

References and resources for further reading

Altbach, P.G., Reisberg, L., & Rumbley, L.E. (2009). Trends in global higher education: tracking an academic revolution. A report prepared for the UNESCO 2009 World Conference on Higher Education. http://atepie.cep.edu.rs/public/Altbach,_Reisberg,_Rumbley_Tracking_an_Academic_Revolution,_UNESCO_2009.pdf (accessed 28 April 2015).

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics. An issue brief. Washington, DC: U.S. Department of Education, Centre for Technology in Learning. http://www.nku.edu/content/dam/StrategicPlanning/docs/implementationteams/technologysupport/library/learning-analytics-ed.pdf (accessed 28 April 2015).

Biesta, G. (2007). Why 'what works' won't work: Evidencebased practice and the democratic deficit in educational research. *Educational Theory*, 57(1), 1–22. http://www.vbsinternational.eu/files/media/ research_article/Evidencebased_education_Biesta1.pdf (accessed 28 April 2015).

Biesta, G. (2010). Why 'what works' still won't work: from evidence-based education to value-based education. *Studies in Philosophy of Education*, 29, 491–503. doi:10.1007/s11217-010-9191-x.

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, 18(10). http://firstmonday.org/ojs/index.php/fm/article/view/4869 (accessed 28 April 2015).

Booth, M. (2012). Learning analytics: The new black. *Educause Review*, 47(4), 52–53.

Boyd, D., & Crawford, K. (2013). Six provocations for big data. http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=1926431 (accessed 28 April 2015).

Cloggy, W. (2011). The value of analytics in an educational and learning context. Blog post: http://ritakop.blogspot.com/search?updated-min=2011-01-01T00:00:00-08:00&updated-max=2012-01-01T00:00:00-08:00&max-results=7 (accessed 13 August 2015).

Crawford, K. (2013). The hidden biases in big data [Web log post]. *Harvard Business Review*. http://blogs.hbr. org/cs/2013/04/the_hidden_biases_in_big_data.html (accessed 28 April 2015).

Danaher, J. (2014). Rule by algorithm? Big data and the threat of algocracy. http://ieet.org/index.php/IEET/ more/danaher20140107 (accessed 28 April 2015).

Davis, J.L., & Jurgenson, N. (2014). Context collapse: theorizing context collusions and collisions. *Information, Communication and Society*, 17(4), 476–485.

Duval, E. (n.d.). Blog: Learning and knowledge analytics. http://www.learninganalytics.net/?page_id=54(accessed 13 August 2015).

Gašević, D., & Siemens, G. (2015). Let's not forget: learning analytics are about learning, TechTrends. http://link. springer.com/article/10.1007/s11528-014-0822-x (accessed 28 April 2015).

Gitelman, L. (Ed.). (2013). 'Raw data' is an oxymoron. London: MIT Press. Government of South Africa (2013). Protection of Personal Information Act, No. 4. http://www.justice.gov.za/legislation/acts/2013-004.pdf (accessed 28 April 2015).

Henman, P. (2004). Targeted!: Population segmentation, electronic surveillance and governing the unemployed in Australia. *International Sociology*, 19(2004), 173–191. doi: 10.1177/0268580904042899.

Kakaes, K. (2015). The big dangers of big data [Web log post]. http://edition.cnn.com/2015/02/02/opinion/kakaes-big-data/index.html (accessed 28 April 2015).

Kitchen, R. (2014). *The data revolution. Big data, open data,data infrastructures and their consequences*. London: Sage.

Kruse, A., & Pongsajapan, R. (2012). Student-centered learning analytics. *CNDLS Thought Papers*. https://cndls.georgetown.edu/m/documents/thoughtpaperkrusepongsajapan.pdf (accessed 28 April 2015).

Lanier, J. (2013). How should we think about privacy? Making sense of one of the thorniest issues of the digital age. *Scientific American*, November, 64–71.

Lupton, D. (2014). Self-tracking modes: Reflexive selfmonitoring and data practices. Available at SSRN 2483549.

Mayer-Schönberger, V. (2009). *Delete: The virtue of forgetting in the digital age*. Princeton, NJ: Princeton University Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data. Arevolution that will transform how we live, work, and think*. New York, NY: Houghton Miffling Harcourt Publishing Company.

Morozov, E. (2013). The real privacy problem. *MIT Technology Review*. http://www.technologyreview.com/featuredstory/520426/the-real-privacy-problem/ (accessed 28 April 2015).

Napoli, P. (2013). The algorithm as institution: Toward a theoretical framework for automated media production and consumption. *Media in Transition Conference* (pp. 1–36). doi: 10.2139/ssrn.2260923.

Open University (2014). Policy on the ethical use of student data for learning analytics. http://www.open.ac.uk/ students/charter/essential-documents/ethical-usestudent-data-learning-analytics-policy (accessed 28 April 2015).

Prinsloo, P. (2009). Modelling throughput at Unisa. The key to the successful implementation of ODL. http://uir.unisa.ac.za/xmlui/bitstream/handle/10500/6035/ Success%20%20Throughput%20Model%20DD%20corrected%20.pdf?sequence=1 (accessed 28 April 2015).

Prinsloo, P. (2012). ODL research at Unisa. Presentation at the School of Management Sciences, University of South Africa, Pretoria.

Prinsloo, P. (2014). A brave new world: student surveillance in higher education. Presentation at SAAIR, 16–18 October, Pretoria. http://www.slideshare.net/ prinsp/a-brave-new-world-student-surveillance-inhigher-education (accessed 28 April 2015).

Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., & Van Zyl, D. (2015). Big(ger) data as better data. *International Review of Open and Distributed Learning*, 16(1), 284–306. http://www.irrodl.org/index.php/irrodl/article/view/1948 (accessed 28 April 2015).

Prinsloo, P., & Slade, S. (2013). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp.240–244). ACM.

Prinsloo, P., & Slade, S. (2014). Educational triage in open distance learning: Walking a moral tightrope. *The International Review of Research in Open and Distributed Learning*, 15(4).

Prinsloo, P., & Slade, S. (2014). Student data privacy and institutional accountability in an age of surveillance. *Using data to improve higher education* (pp. 197–214). SensePublishers.

Prinsloo, P., & Slade, S. (2015). Student privacy selfmanagement: implications for learning analytics. Paper presented at LAK15, Poughkeepsie, NY, 16–20 March. http://dl.acm.org/citation.cfm?id=2723585 (accessed 28 April 2015).

Prinsloo, P., Slade, S., & Galpin, F. (2012). Learning analytics: challenges, paradoxes and opportunities for mega open distance learning institutions. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 130–133). ACM. http://dl.acm.org/citation.cfm?id=2330635 (accessed 28 April 2015).

Richards, N. M., & King, J.H. (2013). Three paradoxes of big data. http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2325537 (accessed 28 April 2015).

Sclater, N. (2015). Code of practice for learning analytics. A literature review of the ethical and legal issues. http://repository.jisc.ac.uk/5661/1/Learning_Analytics_A-_Literature_Review.pdf (accessed 28 April 2015).

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in learning and education. *Educause review*, 46(5), 30. http://www.Educause.edu/ero/article/ penetrating-fog-analytics-learning-and-education (accessed 28 April 2015).