(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

Why Theory Matters More than Ever in the Age of Big Data

Alyssa Friend Wise
Simon Fraser University, Canada
alyssa.wise@sfu.ca

David Williamson Shaffer
University of Wisconsin, Madison, USA

ABSTRACT: It is an exhilarating and important time for conducting research on learning, with unprecedented quantities of data available. There is a danger, however, in thinking that with enough data, the numbers speak for themselves. In fact, with larger amounts of data, theory plays an ever-more critical role in analysis. In this introduction to the special section on learning analytics and learning theory, we describe some critical problems in the analysis of large-scale data that occur when theory is not involved. These questions revolve around what variables a researcher should attend to and how to interpret a multitude of micro-results and make them actionable. We conclude our comments with a discussion of how the collection of empirical papers included in the special section, and the commentaries that were invited on them, speak to these challenges, and in doing so represent important steps towards theory-informed and theory-contributing learning analytics work. Our ultimate goal is to provoke a critical dialogue in the field about the ways in which learning analytics research draws on and contributes to theory.

Keywords: Learning analytics, learning theory, learning design, research methodologies, statistics, large-scale data

1 INTRODUCTION

The quantities of learning-related data available today are truly unprecedented. Whether the size comes from the number of individuals involved, such as thousands of learners taking a MOOC, or the fine-grained nature of the capture process, such as second-by-second changes in a learner's gaze, it provides exciting new opportunities to probe the patterns and processes of how people learn. It is an exhilarating and important time for conducting research on learning. However, there is a danger in falling into the trap of thinking that with sufficient data, the numbers speak for themselves. In fact, the opposite is true: with larger amounts of data, theory plays an ever-more critical role in analysis.

2 WHERE TO CAST OUR FISHING NETS

There is an important cascade of problems in data analysis and interpretation that scale rapidly when theory is not involved. The first is somewhat obvious but bears repeating: if we collect tens or hundreds of variables from millions of individuals, in the absence of theory, how does a researcher decide which

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

ones to include in an analysis? Each variable could be tested in isolation, or a backward-stepwise approach could eliminate variables that contribute little to the explanatory power of a model, but any approach that relies solely on statistical techniques raises a critical conceptual problem:

*What counts as a meaningful finding when the number of data points is so large that **something** will always be significant?*

The conceptual and mathematical machinery of statistical sampling was developed for datasets of a particular size and in a particular context: large enough that random effects are normally distributed in the data, but small enough to be obtained using traditional methods. Thus inferential statistics were designed to help us tell whether a big idea can be warranted from a small sample. With relatively small samples, statistically significant effects are also those with larger effect sizes, and thus a practical significance as well. Increasing sample sizes stresses these techniques. Statistically significant results are now plentiful, but appear even for very small — perhaps tiny, effects. There have been numerous studies in recent years showing that one or more variables in a large data set is associated with student success of one form or another. But a result derived from a test of 2 million data points that is significant with $p = 0.01$ has an effect size (Cohen's d) on the order of 0.004. To put that in perspective, this effect is over 100 times smaller than the impact of a student's overall motivation on their learning outcomes (Hattie, 2009). In other words the mathematics of statistical analysis means that macro-data will consistently produce micro-results.

There may, of course, be multiple variables each with that effect size. But unfortunately, effects do not typically add linearly. One hundred and fifty variables with a very small effect do not simply add up to a moderate effect overall. The impact of any two variables may reflect the common influence of some other underlying latent factor, or there may be interaction effects between variables. Without a theoretical framework, in other words, it is hard to know what variables to include in a model, how they might interact, which micro-results to pay attention to, or how to select a useful model from the immense array of combinatorial possibilities. This exacerbates the more general problem, pointed out many years ago by Hill (1965), that the "glitter" of statistics can be a hypnotizing distraction from inadequacies in the original data and the critical influence of the many decisions researchers must make in cleaning, structuring, and modelling it (Leek & Peng, 2015). To safeguard against the danger that analytic outcomes are a result of arbitrarily taken decisions (Simmons, Nelson, & Simonsohn, 2011), theory is a critical tool to limit researchers' degrees for freedom by providing a coherent and reasoned framework from which to make decisions. In sum, when working with big data, theory is actually more important, not less, in interpreting results and identifying meaningful, actionable results. For this reason we have offered Data Geology (Shaffer, 2011; Arastoopour et al., 2014) and Data Archeology (Wise, 2014) as more appropriate metaphors than Data Mining for thinking about how we sift through the new masses of data while attending to underlying conceptual relationships and the situational context.

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

3 BIRDS OF FEATHER MODEL TOGETHER

The challenges of atheoretical analysis of large-scale data are not just concerned with having large numbers of features to consider. Large sample sizes in and of themselves create problems in the absence of a theoretical framework. In a sample of 2 million learners, there will almost certainly be some pertinent subgroups (and the number of these may scale with the size of the dataset). It is easy to see how unidentified subgroups can mask results or lead to faulty conclusions about the population as a whole. Consider a variable, such as external pressure to succeed, which has a moderate positive effect on male students and a large negative effect on female students. In a course setting where the genders are equally represented, the variable might show no impact overall. In undergraduate engineering, it might show an overall positive impact, even though the effect on female students was negative, because female engineering students are a small but critical minority in the field. The same problem can occur when we seek to combine data from different course offerings without a clear theoretical rationale for why we expect key variables or relationships to be similar across them. For example, a recent study by Gašević, Dawson, Rogers, and Gašević (2015) showed that predictive modelling across multiple courses consistently misidentified the predictors most relevant for specific ones. Techniques such as structural equation modelling or data partitioning can account for subgroups (and even nested subgroups) in a data set, but this requires the researcher to specify the relevant groups in advance based on some theory of relevant differences. Without such a theory, we run the risk of our analyses both drawing inappropriate conclusions for the population as a whole and failing to detect more nuanced findings for relevant subgroups within it. Both create serious concerns (and potential ethical issues) for using the resultant analytics to make diagnoses appropriate for improving learning processes and outcomes for all learners.

This relates to the critical issue of generalization in learning analytics. It is straightforward to take a training set, develop a model, and then test it on a validation set from another corpus collected under the same circumstances. But to extend this to another situation, a researcher needs to have an explanation of what features of the data are salient to make the model (or findings from it) applicable in another context. For example, our own recent work suggests that similar discourse patterns within a discipline can support the transfer of MOOC discussion forum models despite differences in topic-specific vocabulary (Cui & Wise, 2015).

4 BEYOND "WHO" AND "WHAT" TO "WHY" AND "WHAT NOW"?

The problems multiply when we want to move beyond simple descriptive and predictive findings to make claims about causality and provide a basis for action. Educational research using big data frequently relies on post-hoc analysis, and a correlation between a student's actions and some outcome does not imply causality. As in all non-experimental designs, there is the possibility of reverse causation or an underlying third variable. In such cases, we can still present evidence to support causal claims if we can document both a logical (theoretical) explanation of the observed relationship and eliminate

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

plausible rival explanations (Johnson, 2001). There are statistical techniques that can help address the issue, for example by controlling for possible confounds, but the researcher first needs to identify the important variables that should be controlled for. Given the infinite number of possibilities, theory is needed to direct the attention and efforts of researchers. A good example is provided in the Miyamoto et al. paper (this issue), which identifies and controls for the possibility of individual student differences as driving both the spacing of study sessions and the ultimate certification rates by making within-subject comparisons. They also introduce additional variables to try to account for the more complicated possibility that these variables are a function of a student–course interaction factor: the degree of struggle a student experiences.

Beyond improving explanatory potential, there is also the critical issue of how to make learning analytics intelligence actionable (Clow, 2012). At a basic level, we need to ensure that some of the variables studied can be changed to influence the learning process. In this vein, Holland (1986) argues that, while useful for prediction, attributes of a person (e.g., a student's gender, socio-economic status, or prior achievement) can never be considered true causal variables since they provide neither an explanation of a mechanism for why the associated outcome occurred nor any recourse for remedying the situation. This is an important critique given the many attribute variables used for prediction in learning analytics work. One way to avoid such problems is to invert our research logic; that is, instead of starting with an outcome and retroactively searching back for what might have caused it, we begin with identified input (or process) variables and look for what effects they may have.

It is also important to keep in mind that there is frequently a gap between knowing that a variable matters and knowing what to do about it. For example, a recent study of MOOC learner behaviour suggested an association between the increased use of certain resources and the likelihood that a student would drop out of the class in the following week (Breslow et al., 2013). What do we do with such information? Learning is not likely to be enhanced by advising students to study these resources less. Similarly, van der Maas and Wagenmakers (2005) show that chess expertise can be predicted by how rapidly a player makes their moves, but telling novices to move more quickly will not help them improve their game. There are an infinite number of other possible actions we could take in response to such information, but we cannot test them all. Thus there has to be some basis — some underlying theory — for whatever choice or choices we make. Unless we want to return to a new age of dustbowl empiricism,[1] theory plays a crucial role in developing models, interpreting them, and converting those interpretations to meaningful — and scientifically justified — actions.

[1] A term referring to an approach to research that focuses on the haphazard accumulation of empirical observations and relationships between variables without attention to logic or meaning.

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

5 NOT ONLY A "ONE-WAY BRIDGE"

The role of theory in the analysis of large-scale data thus has several important functions:

- Theory gives a researcher guidance about which variables to include in a model
- Theory gives a researcher guidance about what potential confounds, subgroups, or covariates in the data to account for
- Theory gives a researcher guidance as to which results to attend to
- Theory gives a researcher a framework for interpreting results
- Theory gives a researcher guidance about how to make results actionable
- Theory helps a researcher generalize results to other contexts and populations

In saying this, we want to be clear that exploratory data analysis is a good thing. One of the most exciting aspects of large-scale data and learning analytics is the ability to discover patterns and associations across modalities (e.g., coordinating gaze and talk), over time (e.g., in the revisiting of previously studied material), or at a micro-genetic level (e.g., how a teacher uses analytics to monitor and support student learning activity). However, if there is not some theory to which such studies later contribute, it will be hard to develop any systematic understanding of learning. A useful analogy is to the antiquarian movement of the late 18th and early 19th centuries, where amateur collectors gathered bones, shells, and other natural objects from around the world in "curio cabinets" in a relatively haphazard fashion. These were (literally) curiosities that raised interest in natural science. It was a stunning collection of data, but its primary scientific value was that it led to later, more systematic investigations of the natural world. Similarly, at the beginnings of discovering a new tool there is a phase of exploration, of seeing what the tool can do, and marvelling at what it can show us. But science only starts when we begin to synthesize findings and ask how the use of a new tool can help us to move forward as a field. In other words, exploratory investigations of big data can be done without an explicit theory to guide them, but they must lead to testable hypotheses, and eventually to explanations and appropriate generalizations of important phenomena in learning. There is only so long that one can celebrate individual findings that certain data is useful to predict some outcome measure of students' learning, such as eventual completion or grade in a course; eventually these need to become part of a systematic scientific framework — that is, a theory.

6 GOAL AND STRUCTURE OF THE SPECIAL SECTION

The role of theory in the analysis of large-scale data and the relationship between empirical learning analytics work and theory more generally are the focus of this special section of the *Journal of Learning Analytics*. In this section of the journal, we look at five studies, each of which connect their work to theory in some way. The papers provide examples of how learning theories are being used to craft analytics, but also in some cases how analytics are helping to advance learning theories. We think this collection of papers is particularly timely because, for the reasons outlined above, understanding the role of theory in the analysis of large-scale data is an urgent need for this young field. To help meet this

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

need, the special section takes a distinctive format intended to spark a larger conversation around the role of learning theory in learning analytics work. Each of the five research articles included in the section presents an exploration of how to move toward a theory-based approach to learning analytics. Following each article, a commentary has been invited to discuss the ways in which the paper draws on and/or contributes back to theory, as well as the challenges that were faced and what productive next steps forward might be. We believe that, collectively, the five articles and associated commentaries that make up this special section provide a fertile beginning for a larger conversation about the importance, role, and challenges for learning analytics in working with theory.

Kelly et al. (this issue) begin with the fundamental premise that applying learning theory to drive analytics is not a straightforward process but rather an artful one. Indeed, the construction of learning analytics is a design activity and thus an act of innovation that requires both deep familiarity with the theory and the context of application. Their work proposes a strategy for bridging theory and user needs through the development of first principles to guide function, behaviour, and structure, and provides an example of this approach in action in the context of a collaborative learning activity. In his commentary, Teplovs (this issue) acknowledges the value in developing process guidance for the application of theory to analytics design, and points out the need to extend such guidance to the complementary process of feeding back what is learned from the designed analytics to inform theory development. Here learning analytics researchers might look to the tradition of design-based research for methodological guidance (e.g., Barab, 2014; McKenney & Reeves, 2014; Reimann, 2011) as well as considering how they might infuse experimentation into their studies (Hewitt, this issue).

Both Miyamoto et al. (this issue) and Svihla et al. (this issue) engage in the activity of applying learning theory to analytics design, drawing their inspiration from the theory of distributed practice — a well-established psychological finding that memory retention is increased when rehearsal is spread out over time rather than massed. In a clear example of the generative nature of the dialogue between theory and situation, the way this construct is taken up by the two research groups is dramatically different. Working in the context of diverse massive open online courses (MOOCs), Miyamoto et al., examine the notion of "spaced study" (the degree to which time spent generally interacting with course material is concentrated or dispersed into some number of log-in sessions). In contrast, in the context of a classroom-supported, web-based inquiry learning environment, Svihla et al. probe students' practices of "revisiting" (whether and when they re-engage with particular learning environment elements).

The commentaries on these papers both speak to the challenges in applying carefully formulated lab findings in "the wild." As Hewitt notes (this issue) "a good deal of the learning theory we use today has emerged out of experimental studies where control groups were used to isolate variables. This bears little resemblance to much of today's research the learning analytics field, in which data tends to be collected from naturalistic learning settings" (p. 104). Further exacerbating the challenge, as learning analytics researchers, we often don't have tight control over the design of the learning environments we

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

are studying, working with post-hoc data generated from systems and data structures that we didn't create; we thus must rely on proxy indicators "that may only roughly approximate the phenomena of interest" (p. 104). Both Pardos (this issue) and Miyamoto et al. themselves (this issue) note that the mechanism behind the effects of space study they observed may be very different from the original ones involved in distributed practice, perhaps relating more to motivation than to memory retrieval.

Schneider and Pea (this issue), beginning with a different set of theories from the field of collaborative learning, investigate the potential for a variety of measures of collaborating pairs' language use to serve as a proxy for their degree of "common ground." Different from the Svihla et al. paper, they search not for actions contributing to learning processes, but for automatically trackable "markers" of productive (and unproductive) collaborative learning processes. In his commentary on the piece, Hoppe (this issue) explores the theoretical basis for this work, diving deeply into the literature on "common ground," and noting the great variety and contention in the exact meaning and use of this concept. This raises an important question for learning analytics researchers: How do we move forward to operationalize a concept when the theory itself is not fully agreed upon?

Finally, van Leeuwen (this issue) diverges from the other papers in the special section by proposing and presenting initial evidence supporting a theory of how analytics can support teacher regulation of collaborative learning via support for "noticing" that increases the specificity and confidence of a teacher's diagnosis of the situation. This contribution is notable for being one of first efforts in the field to develop a distinct theory of learning analytics. In his commentary, Chen (this issue) notes the importance of such work in helping learning analytics be "not merely the accepting side of a 'one-way bridge'" when it comes to theory, also helping to "shed light on learning theory and lead to theory building of its own" (p. 164). Specifically he highlights that, as a new field, we should take a generative stance, appreciating theoretical contributions for their ability to explain findings and stimulate new directions for research rather than focusing exclusively on verification and validation.

7 CONCLUSION

All of the papers in this special section begin with accepted educational or psychological constructs.[2] As noted by Schneider and Pea (this issue), this constrains the ways in which the data can be analyzed, and is thus powerful in reducing the risks of finding an effect due to chance. Equally powerfully, taking theory as a starting point helped these researchers to move past the simple time-on-task and activity

[2] Both Pardos (this issue) and Teplovs (this issue) highlight the importance of this — that is, working with well-established learning theories, rather than peripheral ones or theoretical premises chosen in a piecemeal fashion; however, Chen (this issue) also warns of the danger of adhering so tightly to one theoretical doctrine that other relevant ones are ignored. He goes on to comment that to address this concern we may need to go beyond taking a theoretical stance to "articulate why competing theories are less fruitful for a given scenario" (p. 164).

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

count metrics commonly used in the field (and which imply a de facto "theory" of more-is-better) to explore more nuanced metrics of learning processes. Nonetheless, a critical question raised in many of the commentaries is this: To what extent do the papers simply evoke theory as a point of departure versus carefully using theoretical constructs to inform the specific analytics created? Theory-inspired learning analytics research is certainly an improvement over dustbowl empiricism, but it is not enough to build a body of knowledge nor to sustain a field. Thus the question of *how* theory is operationalized in a given analytics effort and the justifications for this become important in assessing the work. One notable example of this appears in the paper by Svihla et al. (this issue) who include a chart that lays out a theoretical justification for each of their revisiting metrics (p. 86). This precision in how theory informed their different analytics (and the specification of more than one potential operationalization that could then be tested) made it possible for them to then speak back to their nascent idea of revisiting in a meaningful way — in other words, Svihla et al. do conceptual work in proposing and empirically testing how the theory of distributed practice might be *productively adapted* to the context of student-driven activity in a web-based inquiry learning environment.

With the goal of doing such conceptual work, collectively the papers in this special section both provide powerful examples of how to move towards theory-informed and theory-contributing learning analytics work and raise a number of important challenges for researchers to consider. We hope that together, the set of papers and associated commentaries provoke a productive dialogue in the field about the ways in which learning analytics research can draw on and contribute to theory.

REFERENCES

Arastoopour, G., Buckingham Shum, S., Collier, W., Kirschner, P., Knight, S. K., Shaffer, D. W., & Wise, A. F. (2014). Analytics for learning and becoming in practice. In J. L. Polman et al. (Eds.), *Learning and becoming in practice: The International Conference of the Learning Sciences (ICLS) 2014: Vol. 3* (pp. 1680–1683). Boulder, CO: International Society of the Learning Sciences.

Barab, S. A. (2014). Design-based research: A methodological toolkit for engineering change. In K. Sawyer (Ed.), *Handbook of the Learning Sciences: Vol. 2.* (pp. 233–270). Cambridge, MA: Cambridge University Press.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment, 8*(1), 13–25. Retrieved from http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF2.pdf

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics & Knowledge (LAK 2012),* 134–138. http://dx.doi.org/10.1145/2330601.2330636

(2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. http://dx.doi.org/10.18608/jla.2015.22.2

Cui, Y., & Wise, A. F. (2015). Identifying content-related threads in MOOC discussion forums. *Proceedings of the 2nd ACM Conference on Learning @ Scale*, 299–303. http://dx.doi.org/10.1145/2724660.2728679

Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2015). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicating academic success. *The Internet and Higher Education*. Advance online publication. http://dx.doi.org/10.1016/j.iheduc.2015.10.002

Hattie, J. A. C. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement.* New York: Routledge.

Hill, A. B. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine, 58*(5), 295–300. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. http://dx.doi.org/10.1080/01621459.1986.10478354

Johnson, B. (2001). Toward a new classification of nonexperimental quantitative research. Educational Researcher, 30(2), 3–13. http://dx.doi.org/10.3102/0013189X030002003

Leek, J. T., & Peng, R. D. (2015). Statistics: *p* values are just the tip of the iceberg. *Nature*, *520*, 612. http://dx.doi.org/10.1038/520612a

McKenney, S., & Reeves, T. C. (2014). Educational design research. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 131–140). New York: Springer.

Reimann, P. (2011). Design-based research. In L. Markauskaite, P. Freebody, & J. Irwin (Eds.), *Methodological choice and design: Scholarship, policy and practice in social and educational research* (pp. 37–50). New York: Springer.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Shaffer, D. W. (2011, December). Epistemic network assessment. Presentation at the National Academy of Education Summit, Washington, DC.

Van der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology, 118*(1), 29–60. Retrieved from http://www.jstor.org/stable/30039042

Wise, A. F. (2014). Data archeology: A theory informed approach to analyzing data traces of social interaction in large scale learning environments. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses* (pp. 1–2). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W/W14/W14-4100.pdf