JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 5(2), 47–59. http://dx.doi.org/10.18608/jla.2018.52.4

Embedding Experiments: Staking Causal Inference in Authentic Educational Contexts

Benjamin A. Motz [1], Paulo F. Carvalho [2], Joshua R. de Leeuw [3], Robert L. Goldstone [4]

Abstract

To identify the ways teachers and educational systems can improve learning, researchers need to make causal inferences. Analyses of existing datasets play an important role in detecting causal patterns, but conducting experiments also plays an indispensable role in this research. In this article, we advocate for experiments to be embedded in real educational contexts, allowing researchers to test whether interventions such as a learning activity, new technology, or advising strategy elicit reliable improvements in authentic student behaviours and educational outcomes. Embedded experiments, wherein theoretically relevant variables are systematically manipulated in real learning contexts, carry strong benefits for making causal inferences, particularly when allied with the data-rich resources of contemporary e-learning environments. Toward this goal, we offer a field guide to embedded experimentation, reviewing experimental design choices, addressing ethical concerns, discussing the importance of involving teachers, and reviewing how interventions can be deployed in a variety of contexts, at a range of scales. Causal inference is a critical component of a field that aims to improve student learning; including experimentation alongside analyses of existing data in learning analytics is the most compelling way to test causal claims.

Notes for Practice

- Learning Analytics, as a field, should ultimately strive to make strong causal inferences, identifying the specific interventions that optimize and improve learning.
- The most straightforward and compelling research method for supporting causal inference is experimentation.
- In this article, we advocate for embedding experiments within pre-existing learning contexts, in order to improve the strength of causal claims in learning analytics, and also to close the research/practice loop.
- We review practical matters in the design and deployment of embedded experiments and highlight the benefits of including experimentation in the learning analytics toolkit.

Keywords

Causal inference, experiments, research design, ethics , A/B testing.

Submitted: 16.10.18 — **Accepted:** 06.02.18 — **Published:** 05.08.18

Corresponding author [1]Email: bmotz@indiana.edu Address: Department of Psychological and Brain Sciences, Cognitive Science Program Indiana University, 1101 East 10th Street, Bloomington, IN, 47405, United States ORCID ID: 0000-0002-0379-2184.

[2]Email: pcarvalh@andrew.cmu.edu Address: Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, United States ORCID ID: 0000-0002-0449-3733

[3]Email: josh.deleeuw@gmail.com Address: Human-Computer Interaction Institute, Vassar College, 124 Raymond Avenue, Poughkeepsie, NY 12604, United States

[4]Email: rgoldsto@indiana.edu Address: Department of Psychological and Brain Sciences, Cognitive Science Program, Indiana University, 1101 East 10th Street, Bloomington, IN 47405

1. Causality in Learning Analytics

Learning analytics, as a field, is universally defined with a specific purpose in mind: optimizing and improving student learning. Towards this goal, research in learning analytics should not only explain learning processes within our educational systems, but should also bridge the research and practice gap to produce “actionable intelligence” (Norris, Baer, Pugliese, & Lefrere, 2008; Arnold, 2010; Elias, 2011; Clow, 2012, 2013) — developing systems, predictions, interventions, or insights to improve outcomes in authentic learning environments. As members of a learning analytics research community, we should aim to make strong, actionable, causal inferences: “my research suggests that if you do *this*, student outcomes will improve.”

Although a key goal of learning analytics is to ultimately make causal inferences, the conventional methods of learning analytics have excluded standard research tools for supporting such inferences. Until recently, most characterizations of learning analytics research methods were limited to observation of student data generated from real educational systems (Cope & Kalantzis, 2015b), with inferences gleaned primarily from statistical modelling, visualizations, and dashboards based on these extant data resources (Baker & Yacef, 2009; Bienkowski, Feng, & Means, 2012; Chatti, Dyckhoff, Schroeder, & Thüs, 2012; Siemens, 2012, 2013; Dietz-Uhler & Hurn, 2013; Khalil & Ebner, 2015; for a constructive critique of these characterizations, see Lodge & Corrin, 2017). The booming availability of large datasets, offering the ability to quickly search and summarize records across an entire student population's educational landscape, created enticing new research opportunities, typically emphasizing the discovery of relationships using exploratory data analysis (Enyon, 2013; Baker & Inventado, 2014) and predictive models of future outcomes (Macfadyen & Dawson, 2010). These analyses can reveal important and useful relationships that have previously been completely unobservable. But why stop there?

We suggest there is something missing, an epistemological gap, in the conventional view of the learning analytics toolkit. Analyses of existing datasets can play an important role in detecting and discovering causal patterns, but an indispensable aspect of this research, if we truly aim to create reliable actionable intelligence, is the conduct of *experiments*. In addition to harnessing data traces, learning analytics should rigorously explore ways of manipulating these traces, conducting experiments to evaluate an action's effect on intended outcomes.

We are not the first to voice this argument. Developing Kolb's (1984) theoretical work, Clow (2012) prominently asserted that, once learning analytics produces actionable intelligence, a critical next step is to develop this insight into an intervention, actively experimenting to examine whether an action causes a change in learner behaviour (see also Koedinger, Stamper, McLaughlin, & Nixon, 2013). Similarly, Reich (2015) argued that, without experimental intervention research, the causal links between aspects of course design and student performance are unclear. Some have also recently noted that online courses, in particular, provide researchers with the opportunity to easily implement experiments that clarify the relationship between design choices and student achievement (Williams & Williams, 2013), as well as addressing broader questions about educational practices (Kizilcec & Brooks, 2017).

The benefit of experimentation is that it represents the single most persuasive way to support a causal inference (Shadish, Campbell, & Cook, 2002). This is because, in an experiment, exposure to a causal antecedent (a learning activity, a new technology, an advising strategy, etc.) is manipulated by the researcher, enabling direct assessment of whether some consequence (e.g., a learning outcome) can be causally attributed to the specific change in treatment. The hallmark of an experiment is that the unit under observation (a student, a teacher, a class, etc.) should be randomly assigned to different conditions. In this way, there should be no differences between treatment groups other than the experimental treatment itself.

Nevertheless, experimenters should be sensitive to the possibility that some consequential difference other than the treatment could be lurking between randomly assigned comparison groups. Statistical analyses are used, in part, to quantify the likelihood of this error, and the possibility of imbalance can be minimized by using large samples and only accepting results that meet conservative statistical thresholds. Additional methods for randomly assigning treatments to subgroups within the sample (e.g., blocking; Higgins, Sävje, & Sekhon, 2016), or repeating random assignment until balance is achieved on pre-specified dimensions (rerandomization; Morgan & Rubin, 2012) may further mitigate the possibility of imbalance. These may be uniquely appropriate techniques in learning analytics, where researchers typically have more background data on research subjects than in other fields. Alternatively, a more common approach would be to include model-based estimators (e.g., regression adjustments) to control for other variables that might produce imbalance in the comparison groups. None of these techniques fully eliminates the possibility of error in random assignment, but with appropriate design and analysis choices, experimenters can minimize this risk.

In total, evidence from an experiment satisfies the strong requirements of causal inference by demonstrating that changes in treatment modify an outcome in a specific direction (ruling out reverse causality) while minimizing (by randomization and other methods) the possibility that some other factor caused changes in the outcome.

It bears mention that these conditions might also be satisfied (to some degree) using quasi-experimental or even non-experimental methods. Particularly when taking into account the temporal ordering of variables and causally relevant background variables, some observational analyses are able to provide distinguishing evidence for causal relationships over mere covariance (Pearl & Verma, 1995; Spirtes, Glymour, & Scheines, 2000; Russo, 2010; Murnane & Willett, 2010; Kumar, Clemens, & Harris, 2015). As Tufte (2003) pronounced, "Correlation is not causation, but it sure is a hint" (p. 4). Our goal is not to suggest that experimentation is the *only* way to offer empirical support for a causal claim or to suggest that it is infallible (Imai, King, & Stuart, 2008), but to assert that it is a uniquely powerful tool when assessing the effect of an intervention — particularly so, considering the goals of learning analytics and educational research in general (US Department of Education, 2016, 2017).

This assertion is not without historical controversy in the broader study of teaching and learning (Angrist, 2004). For example, theorists have questioned whether causality is a meaningful theoretical construct in education (e.g., Maxwell, 2004), whether control is possible in an educational setting (e.g., Barab & Squire, 2004), and whether it is feasible to identify individual causal relationships for complex problems in education (e.g., Morrison & van der Werf, 2016). These are reasonable concerns (which similarly apply to descriptive and correlational work), and programs of experimental research in learning analytics should certainly aim to make precise and meaningful theoretical claims (Wise & Shaffer, 2015), should utilize research implementations that have external validity (Lockyer, Heathcote, & Dawson, 2013), and should be sensitive to the complexity of educational systems (Koedinger, Booth, & Klahr, 2013). Experimentation that includes these features can be difficult to implement and is not always possible. However, the challenges of conducting experiments in education do not justify ignoring the epistemological value of experiments in education.

Even beyond providing strong evidence for a causal relationship, experiments can also help by pinpointing the precise conditions under which an outcome should be observed. As such, the details of an experiment can help researchers evaluate whether a causal relationship should generalize to new situations. It is likely that additional variables (such as learner demographics, the educational context, or the nuances of the situation) will moderate the effect of a treatment. The "mileage" of any intervention may vary between different situations, and controlled experiments can help prevent overgeneralization by providing clear estimates of a causal effect within a specific context (e.g., Kizilcec & Cohen, 2017). For these reasons, we see tremendous promise for the field of learning analytics researchers deploying experiments in a diversity of learning contexts.

Why, then, has experimentation only recently started to appear in catalogues of the methods of learning analytics? To our knowledge, no learning analytics researcher has ever voiced an argument *against* experimentation, but we can postulate a few concerns. Perhaps experiments, traditionally associated with laboratories, rigour, and control, seem incompatible with the opportunities afforded by the surge in big, messy, authentic student data. Perhaps the act of manipulating exposure to different educational interventions seems unethical in real classes. Perhaps an experimental operationalization of a learning treatment would be considered artificial or unrepresentative of natural instruction. And perhaps a randomly assigned learning intervention seems too challenging to implement at scale.

These hurdles are not insurmountable, and the benefits of explicitly including experimentation in the "learning analytics cycle" (Clow, 2012) greatly outweigh the challenges. In this article, we address each of these postulated challenges, and ultimately provide a framework to expand the scope of learning analytics research methodology, from pure extant data mining to the inclusion of embedded experimental research that aims to manipulate student outcomes and draw stronger causal inferences.

2. Embedded Experiments

Thus far, we have argued for the unique inferential power that experimental interventions have for determining causality, and that they should be a major component in the learning analytic toolkit. Assuming the acceptance of this general claim, a logical next question becomes: What would these experiments look like?

Consistent with the focus of learning analytics on measuring learner data within educational contexts such as classrooms, museums, online tutoring, and on-the-job training, we would like to advocate for *embedded experimentation*. By embedded experimentation we mean experiments conducted within pre-existing educational contexts, including both formal classrooms and informal learning settings, including both schools and workplace environments, and making use of authentic learning materials and assessment instruments that are relevant to the pre-existing learning goals.

The notion of embedded experimentation shares considerable common ground with proposals for *in vivo* experiments (Koedinger, Corbett, & Perfetti, 2012; Koedinger et al., 2013a), but we favour the "embedded experimentation" term because it emphasizes that learning is a major activity across the lifespan, in both educational and workplace training contexts, and in both informal and formal settings. Diverse and elaborate institutions have been established to foster learning, including classrooms, museums, studios, workshops, special interest groups, and online tutorials, and these offer unique opportunities to study societally relevant learning. By bringing experimentation to these contexts — by *embedding* experiments in these pre-existing institutions — we can assure that the results are pertinent to at least some naturally occurring situations, and we can take advantage of the learning infrastructures that have been created with much expense and time. While learning in both research laboratories and university classrooms is arguably *in vivo* in that it is taking place in an intact, whole organism, only learning in university classrooms would count as embedded learning. Embedded learning focuses on studying learning in the "wild" — in the natural, albeit socially constructed contexts in which it has developed on its own, independent of researchers' theories and paradigms.

To many ears, the very phrase "embedded experimentation" may sound like an oxymoron. Experiments may be assumed to be what researchers do within laboratory contexts: Learners are brought into a laboratory, settled into their own private cubicle, presented with artificial materials to be learned, and subsequently tested on their acquisition and generalization of these materials. Although this is the dominant paradigm within cognitive psychology, there is also a long, if sometimes forgotten history of conducting learning experiments in pre-existing contexts outside of the psychology laboratory (Bryan & Harter, 1899; Hall, 1891).

Embedding experiments within already established learning contexts has several advantages over laboratory investigations. First, learners are less likely to be self-conscious and more likely to use the kinds of learning strategies that they normally employ. Laboratories are unfamiliar environments that almost inevitably put the learner at a disadvantage in terms of authority, control, and comfort. Second, if a researcher wants to better understand likely learning outcomes in a specific context, it is wise to study them within that context. There have been many well-documented cases in which learning processes and outcomes differ profoundly across cultures, schools, and contexts (see Medin & Bang, 2014). Third, the archetype of the solitary learner acquiring information in a generic context is never, in fact, realized (Greeno et al., 1998). Learning is always situated in a context, and whatever learning takes place is always an interaction between the learner and their context. Embedding experiments within those contexts allows a researcher to understand how an intervention affects the broader, distributed system of learning. For example, an intervention that encourages students in a class to talk to their peers about the course material may improve not only their own understanding, but the understanding of their peers as well (Crouch & Mazur, 2001). These indirect benefits would only be discoverable when the peer-instruction intervention is deployed in the context of a course complete with other students, and not when the students are isolated in their own laboratory cubicles.

The core characteristic of an embedded experiment is that some learners learn with one form of the intervention while other learners learn with another form of the intervention. This comparison between interventions may or may not resemble traditional laboratory experiments in which compared conditions are selected to differ in only one way. By virtue of this flexibility in choosing apt comparisons, we are more optimistic about the feasibility of conducting genuine experiments in embedded contexts than others who have emphasized the expense and difficulty in deploying randomized control trials, or RCTs (see Sullivan, 2011). Our optimism stems from an open, ecumenical stance towards experimental design. Different experimental designs are appropriate for different contexts, and if one permits oneself flexibility in terms of design choices, then one can usually find an embedded experimental design that warrants qualitatively stronger causal inference than is possible without intervening on the educational system (Pearl, 2000). Our optimism also stems from the surging availability of online data traces in contemporary educational systems; an experiment that randomly assigns different versions of an online homework activity can yield detailed behavioural data on-par with what had previously only been possible in a laboratory with specialized software (Cope & Kalantzis, 2015a).

One important design consideration concerns the choice of the treatment conditions to compare. For the purposes of isolating a key contributing factor in a learning context, establishing very similar groups that differ only on that factor is desirable. By keeping the materials and the student population constant across conditions, differences between even subtly different experimental conditions can be detected that would otherwise be missed. For example, Roediger, Agarwal, McDaniel, and McDermott (2011) conducted a series of embedded experiments to compare the benefit of frequent quizzing with the benefit of re-reading. For the study, the authors selected a subset of different materials covered in the students' curricula and normal class activities to be included in the study. Pre-test and post-test measures were specifically created for these materials and different materials were assigned to be quizzed or re-read for different individual participants (i.e., which subset of materials were re-read or quizzed varied across students). This strategy of designing minimally contrastive conditions is particularly useful when: 1) a researcher can identify and manipulate a key factor governing learning that is likely to arise in many different learning contexts, 2) the choices of factor levels (i.e., *quizzed* vs *re-read* for the factor "study type") along different factors (i.e., curriculum topic) are at least partially independent of each other, and 3) the difference in learning outcome likely to be found for different factor levels is small-to-moderate and may be swamped by variation along many other factors.

While minimally contrastive interventions are valuable for isolating the effect of a single contributing factor, they are by no means the only game in town, and other experimental designs are better in other contexts. One alternative, oftentimes effectively employed after several influential minimally contrastive interventions have been identified, is to compare a condition in which all empirically favourable levels of factors are combined on a "Dream Team" package of pedagogical changes and compared to a condition in which neutral or status quo levels of these factors are combined. A good example of this strategy was adopted by the National Research & Development Center on Cognition & Mathematics Instruction[1] in their

[1] https://www.iesmathcenter.org

effort to create an improved mathematics textbook by applying established principles of the cognitive science of learning (Booth et al., 2017). Although this approach — contrary to the minimal contrastive approach — does not allow a single factor to be unambiguously identified as impacting learning outcomes, it offers the countervailing advantage of determining whether a set of independent design decisions complement each other when combined so that the entire system confers pedagogical benefits. Furthermore, the "Dream Team" condition may often show large, statistically robust benefits even when each of the factors has only a small effect size. If the package of changes does show a robust benefit, then subsequent experiments employing minimally contrastive interventions can be deployed to isolate the most potent ingredients of the composite intervention.

Another possible way to choose the interventions to compare is inspired by the notion of "pragmatic trials" in medicine. Contrasted with "explanatory trials" designed to test if and how an intervention confers medical benefits compared to placebo controls using RCT, pragmatic trials investigate whether an intervention confers benefits in real life contexts compared to other viable alternatives (Patsopoulos, 2011). For example, in testing whether liposuction is an efficacious treatment for obesity, comparing its effects to those produced by putting patients on a regular schedule of exercise would count as a pragmatic trial. These strategies for treating obesity differ in a variety of important ways, and for that reason, even if one strategy, say exercise, is clearly superior to the other, one still would not know whether this is because it requires the active involvement of the patient, does not require invasive surgery, is persistent, or some other factor. Still, if one is a doctor trying to devise a sensible long-term policy for treating patients, the results from this pragmatic trial may be exactly what one is looking for. Likewise, teachers trying to choose between different curricula, tutoring systems, or textbooks may simply need an experimental "cook off" comparison of some of the most *prime facie* plausible possibilities, testing whether one reasonable instructional design is better than another. An example of this type of approach to embedded experimentation is the study conducted by Kirchoff, Delaney, Horton, and Dellinger-Johnston (2014) to test the efficacy of a computer-based perceptual training intervention. The authors tested whether training software aimed at improving student recognition of plants (that incorporated several design features known to benefit perceptual training) would improve student learning in a plant systematics course. To this end, they compared learning outcomes when students used the software and when they used status quo classroom practices. Although this study does not allow one to determine which feature(s) of the software contribute to improved performance, the results do suggest that perceptual training can contribute to improved conceptual learning.

One advantage of embedded over laboratory experiments is that they encourage researchers to consider comparing interventions that make sense in real world contexts. For example, cognitive psychologists studying concept learning in the laboratory often make the assumption that learners must learn a set of concepts via pure induction — by seeing examples, attempting to categorize the examples, and then receiving feedback on the correctness of their categorization (Goldstone, Kersten, & Carvalho, 2017). Perhaps this assumption is a vestige from early animal learning research (in which it would be impossible to provide verbal instruction to a rat, for example, that shape but not brightness is relevant), but in educational contexts this represents a rather ineffective pedagogical strategy. By contrast, teachers, coaches, and parents have all found that even though wisdom cannot always be directly told to learners (Bransford, Franks, Vye, & Sherwood, 1989), well-crafted words, rules, and instructions can often be used to dramatically expedite both performance and understanding (Klahr & Nigam, 2004; Ellis, 2005). Laboratory-focused researchers might end up comparing artificial learning conditions, such as perfect alternation between concepts to be learned (e.g., sequencing the examples of two concepts in the order ABABABAB) versus perfectly blocked concepts (e.g., AAAABBBB), without adequate acknowledgment of the possible irrelevance of this comparison for real world learning environments. Researchers engaging in embedded experimentation are more likely to consider interventions that generally conform to educational best practices such as well-timed instructions, informative feedback, verbal help, and hints.

The general point is that choosing minimally contrastive interventions to compare is indeed an appropriate experimental design strategy, but it is not the only important consideration. It is also appropriate to consider the real-world relevance of the interventions to actual instructional practice, and the current state-of-the-art in teaching of the discipline. For example, if it is generally appreciated in a teaching community that simple re-reading is not an effective study method, then an experiment that compares re-reading as part of a group versus independent re-reading will risk being largely irrelevant to practice. The choice of interventions to compare should be based on their prevalence, demonstrated efficacy, and practicality, in addition to their precision in isolating single factors, depending on the research question.

In sum, embedded experiments can support a variety of causal inferences. In some cases, the inference will be specific to a particular factor that affects learning outcomes. In other cases, the inference will be about a general approach or strategy without isolating the factor(s) responsible for the improvement. The nature of a particular embedded experiment will depend on the theoretical goals of the research and the practical constraints of the educational situation.

3. Ethical Considerations

The notion of intentionally manipulating a learner's educational experience for the purpose of research raises an important ethical question: What if condition B is reliably inferior to condition A? Has the research harmed the learners in this case? Before addressing this question specifically, consider for a moment that teachers, at all levels, are encouraged, if not expected, to experiment in their classrooms routinely. Experimenting with different instructional methods is viewed as a positive feature of teachers' professional development and growth (Guskey & Huberman, 1995), where a teacher tries new things (on a full student cohort) and reflects on the efficacy of the new approach. Under this scheme, whether new tactics "work" can only be judged by subjective reflection, because there is no balanced comparison condition to make valid analytical contrasts. Thus, unbeknownst to them, students in practically all classrooms are participants in a vast enterprise of uncontrolled experimentation. This enterprise carries the same risk of inferior treatment as what we are proposing (and perhaps more, because negative effects might not be readily apparent to subjective reflection) but affords none of the benefits of causal inference. Perhaps ethical considerations do not hinge on whether experiments should be embedded in classrooms, but whether well-designed, controlled experiments should be embedded.

At the most basic level, a manipulation that is known to negatively impact learning would be of no use as a comparison condition in an embedded study. Similarly, unnaturally deprived control conditions would be inappropriate for experimental contrast, as these would overestimate the manipulation's performance against realistic alternatives (as discussed in the previous section). At the very least, an embedded experiment should contrast sensible design options, and should not administer a treatment known to or believed to potentially cause decrements in learning outcomes.

Even so, a practical way to avoid any possibility that a group will experience disproportionate risk is to administer all treatments to all groups but staggered in time. A crossover or delayed treatment design allows one to compare a group that received a treatment with a group that has not *yet* received that treatment. For example, in examining the benefits of instruction using library archives, Krause (2010) embedded an experiment in an undergraduate history class: one half of the class initially received the experimental exposure to archival instruction, and the other half received the same instruction and assignments four weeks later. Incidentally, in addition to addressing potentially ethical issues, this approach might also improve the statistical and exploratory power of the study, potentially allowing replications within the same cohort (Heath, Kendzierski, & Borgida, 1982).

Rather than *avoiding* risks (by balancing the different treatments within comparison groups), another option would be to simply *minimize* possible risks of different treatment. An embedded experiment could focus on a relatively small aspect of the course, so that any differences between groups are practically negligible for an individual student. For example, an intervention could be designed to only affect performance on just a few target questions on a single test, such as in the study we mentioned above testing the benefits of frequent quizzing (Roediger et al., 2011). One of the benefits of scale (see next section) is the opportunity to embed experiments with a very large number of students, making it possible to measure reliable differences in treatment, even with small effect sizes. With unknown consequences of treatment, it is best to keep modest aims when embedding an experiment in a real learning context. For example, in an embedded experiment including over 2,000 students, Carvalho, Braithwaite, de Leeuw, Motz, & Goldstone (2016) tested whether the way students choose to organize their study influenced their learning outcomes. The authors did this by choosing a single class topic (measures of central tendency) for their intervention and included only four test questions (on a single exam pertaining to that topic) as a post-test measure. The large sample allowed inferences to be drawn from a short intervention with a small effect size.

How do the risks of embedded experimentation compare with laboratory experimentation? Arguably, generalizing from small-scale laboratory studies with limited samples carries a bigger potential of deploying detrimental interventions. Instead, by embedding experiments in authentic contexts, interventions are tested in natural settings using appropriate sample sizes that represent the diverse population of students. This means that embedded studies have the potential benefit of a more inclusive study setting capable of reaching populations not typically studied in the laboratory.

In our view, with proper care as described above, this type of study risks no greater harm than any number of pedagogical decisions that teachers make every day. By working with teachers to create truly embedded studies, using appropriate tools and large-scale data collections, we believe the benefits to the research participants can be maximized. Under these terms, one might argue that conducting an experiment is *more* ethical than uncontrolled pilots in educational environments, particularly when one is uncertain about which of several plausible interventions to implement. If an intervention is worth doing, it's worth systematically testing its effects against reasonable alternatives. By conducting an actual experiment on an intervention's efficacy, its benefits will be more convincing to other researchers and teachers. We suggest there is risk in *not* conducting experiments: genuinely beneficial instructional innovations will be ignored if they are not supported by compelling, rigorous data.

4. Embedded Experiments at Scale

It is possible to conduct embedded experiments at all scales. Whereas small scale “drop-in” studies that involve one small manipulation in a single classroom are common (Arnold et al., 2017; Butler, Marsh, Slavinsky, & Baraniuk, 2014), it is also possible to create carefully controlled studies embedded in educational contexts that span several classes, schools, populations, and geographical areas. For example, it is possible to perform the same experimental manipulation in different content areas (Cantor & Marsh, 2017), across different classes of the same course (Carvalho et al., 2016), in large-scale massive online courses (Chen, Demirci, Choi, & Pritchard, 2017; Zheng, Vogelsang, & Pinkwart, 2015; Kizilcec, Pérez-Sanagustín, & Maldonado, 2016; Williams & Williams, 2013), or across multiple schools (Fyfe, 2016; Koedinger & McLaughlin, 2016).

One of the powers of embedded experimentation lies in combining it with institution-level data collection in the learning analytics tradition, commonly by using online learning platforms or massive courses (e.g., Renz, Hoffmann, Staubitz, & Meinel, 2016; Heffernan & Heffernan, 2014). Larger studies integrating across multiple populations will support more sensitive comparisons and/or more robust causal inferences. Moreover, when outcome measures are joined with existing institutional data, these can also provide better information about demographic factors that correlate with observed effects. Still, although large-scale embedded experiments have great potential, scaling up to a large coordinated experiment across multiple populations can present substantial challenges.

The internet is an obvious tool for solving the scaling challenge. The most straightforward use of the internet is as a distribution platform. For example, the PhET Interactive Simulations project (Wieman, Adams, & Perkins, 2008) has developed dozens of simulations for teaching concepts in STEM fields. These simulations can be accessed by any teacher or researcher through a web browser, and used as part of a classroom activity or an embedded experiment (Finkelstein et al., 2005; Moore, Herzog, & Perkins, 2013). Going a step further, the internet can also be used to create efficient coordination for collecting and aggregating data across multiple classrooms. One example of this approach is the ASSISTments platform (Heffernan & Heffernan, 2014). Researchers can use ASSISTments to develop student activities that contain a manipulation of one or more factors and collect data from students in classrooms. Teachers are (partially) involved in the process because they can choose which activities in ASSISTments are relevant for their class.

A more flexible approach is to create custom experiment materials using web-friendly technology so that the experiment can be deployed online and yet retain the flexibility of traditional classroom activities. Increasingly, cognitive scientists are utilizing online tools to conduct experiments over the internet (Stewart, Chandler, & Paolacci, 2017), and several platforms have been created to make the development of custom online experiments easier (de Leeuw, 2015; Henninger, Mertens, Shevchenko, & Hillbig, 2017). Embedded experiments using online survey platforms (Day, Motz, & Goldstone, 2015), and custom JavaScript (Carvalho et al., 2016) highlight the utility of this approach. However, building an online experiment still requires a relatively burdensome amount of technical knowledge, and so is presently only available to researchers and teachers who themselves have expertise, or a substantial budget.

Even when studies themselves are performed at small scale in isolation, open-science tools like DataShop (Koedinger et al., 2010) and LearnSphere[2] — where data from embedded studies can be stored, shared, combined, and analyzed — can facilitate the kinds of statistical power that would be possible with large-scale experiments (e.g., Koedinger & McLaughlin, 2016; Koedinger, Booth, & Klahr, 2013). These tools exemplify an alternative approach to scaling up embedded experiments: individual researchers and teachers conduct experiments at relatively small scales, but data collection is aggregated across research sites to realize the power of large-scale experimentation. In psychology, a series of *ManyLabs* projects have allowed researchers to pool resources in this way to investigate questions best answered with distributed large-scale experiments (Klein et al., 2014; Ebersole et al., 2016; Frank et al., 2017).

In an ideal world, the technology for creating large-scale embeddable experiments would be user-friendly enough that teachers and researchers can use it as they would use any other piece of software to create classroom activities. For example, we imagine a world where a teacher could decide that she wants to compare different strategies for practicing factoring polynomials and is able to create a homework assignment that randomly assigns different strategies to students. Perhaps the teacher deploys this experiment in one or two classes of 30 students, finds the exercise to be generally useful, but the sample size too small to draw any robust conclusions. The teacher shares the materials with her colleagues, who do no additional work other than assigning the work to their students, perhaps by directing them to a website or a module in a learning management system. The data from multiple classes is then aggregated and made available for analysis using hierarchical models that take into account different levels of variability in this type of nested data (for example, independent variation in individual student knowledge, as well as classroom differences, teacher differences, etc.).

[2] http://learnsphere.org/

This scenario is certainly possible with today's technology (Severance, Hanss, & Hardin, 2010), but it is neither easy nor commonplace to run embedded experiments. For this vision to become a reality, we need technology that enables highly customizable experiments and instructional materials, that is accessible to teachers and researchers without substantial additional training or expertise, and that can operate at any scale. Currently available options often have one or two of these features, but not all three. This, of course, does not mean that embedded experiments are not feasible with current technology, but rather that there is no universal solution yet.

5. Involvement of Teachers

An experiment involves comparing different treatment conditions, and considering that these treatments are, fundamentally, educational tools and interventions, teachers ought to be involved in the design and analysis of these embedded experiments. This may seem a blatantly obvious and unnecessary statement, seeing as how many learning analytics researchers are teachers ourselves. But in a field that has traditionally defined itself by analysis of second-hand observational data, a shift toward experimental manipulation within live learning settings requires additional involvement of teachers as content experts and as users in the areas being investigated. Specifically, we advocate for increased involvement of teachers in embedded experiments so that interventions are authentic and feasible.

The results of an embedded experiment must be feasible in order to advance the goal of optimizing and improving learning. Just because an experiment is embedded in an authentic pre-existing learning environment does not mean that the experimental manipulation is useful. The results of a well-controlled experiment that pays money to real students for time spent studying, for example (see Fryer, 2011), would be inapplicable to the vast majority of learning environments, because they wouldn't be able to afford to monetize self-regulated studying behaviours at scale. Effectively embedding an experiment in a learning environment means more than just administering treatment to real students; it also means developing a treatment that fits within the constraints of the learning environment. Teachers should be involved to help identify these constraints, ensuring that the experimental manipulation could be realistically implemented in similar environments, which is an important aspect of providing actionable intelligence. Again, an analogy can be drawn to medicine, where many practicing physicians are also involved in research, and the participation of doctors who are also seeing patients is a good thing, benefiting the clinicians, as well as the quality of the research (Lader et al., 2004; Rahman et al., 2011). It helps identify implementation challenges, and bottom-up ideas for treatments. The same is analogously true for teachers.

Teachers, as content experts, can also crystalize and constrain our assumptions about how experimental interventions are appropriately embedded into our courses' and institution's educational goals (Gašević, Dawson, & Siemens, 2015; Bakharia et al., 2016). In this way, the involvement of experienced teachers will help learning analytics researchers avoid overgeneralization and build precision, tailoring experiments to address precise and practical questions about learning. This is important because different learning goals require different teaching moves; an experiment demonstrating a reliable effect in one domain may not generalize to other domains, and this is true at many levels of granularity (Gašević, Dawson, Rogers, & Gašević, 2016). For example, at a very coarse level, how we teach skills is not the same as how we teach declarative knowledge. At a very fine level, there may be uniquely useful models for teaching specific topics, like teaching fractions with pizza slices, or teaching the genetics of inheritance using Punnett Squares. Experimental interventions should be sensitive to these contingencies, avoiding manipulations that are orthogonal to the learning goals, while leveraging best-practice teaching approaches within the discipline so that the treatment is authentic and broadly feasible.

For STEM educators, the contingencies of what "works" when teaching different forms of knowledge (e.g., computer science, biology, engineering, physics, etc.) have catalyzed the emergence of a new field — discipline-based educational research (DBER; National Research Council, 2012). Among the tenets of DBER is the view that disciplinary expertise is a core component of learning research, and that consideration of how students learn in different disciplines does not lead to the degradation of this research. After all, there is not a one-size-fits-all approach to education. Similarly, embedded experiments may aptly uncover different causal patterns in different learning contexts. As such, increasing teacher involvement in learning analytics experimentation can help yield more precise theories, hypotheses, and inferences.

Another product of learning analytics (besides actionable knowledge of learning processes) may be institution-wide data-driven dashboards and visualizations to inform teachers, advisors, and students themselves about learning behaviours and student properties (Govaerts, Verbert, Duval, & Pardo, 2012; Motz, Teague, & Shepard, 2015; Duval, 2011; Tervakari, Silius, Koro, Paukkeri, & Pirttila, 2014). In this mode, learning analytics is responsible for providing an analytical viewport to improve teaching and learning, enabling a user to become better-aware of student propensities (Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). Such a lens might be useless if teachers were not involved in its development. For example, as Lockyer and colleagues (2013) observed, the value of learning management system (LMS) data to predict student success is limited to academic disciplines that make heavier use of digital infrastructure for coursework. When the goal of learning analytics is to produce such a lens, teachers should be involved, both as designers of the system and as users in an embedded experiment pilot, so that the visualization tool is congruent with classroom practice, and so that the tool augments teaching and learning effectively (Plaisant, 2004).

6. Conclusion

The understanding that comes from embedded experiments at scale is an indispensable element of a research enterprise that aims to improve learning. It allows us not only to understand the causal relationship between an intervention and the learning outcomes, but also to uncover its limitations — when it might work differently in different implementations. By embedding experiments in real educational contexts, one can also uncover treatment effects that were not suggested by previous theory or by laboratory experimentation, and test predictions suggested by exploration of existing data. In the end, embedded large-scale experimentation should play a fundamental role in the learning analytics toolkit, bridging research and practice, and helping to identify better learning interventions, better models of learning, and better suggestions for teaching and advising practice.

Acknowledgements

The authors are grateful to two anonymous reviewers whose expert commentary elevated the quality of this article.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Support for this work comes from the “Learning: Machines, Brains, and Children” Emerging Area of Research Initiative at Indiana University.

REFERENCES

Angrist, J. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20*(2), 198–212. http://dx.doi.org/10.1093/oxrep/grh011

Arnold, K. E. (2010). Signals: Applying academic analytics. *EDUCAUSE Quarterly, 33*(1). https://er.educause.edu/articles/2010/3/signals-applying-academic-analytics

Arnold, K., Umanath, S., Thio, K., Reilly, W., McDaniel, M., & Marsh, E. (2017). Understanding the cognitive processes involved in writing to learn. *Journal of Experimental Psychology: Applied*, *23*(2), 115–127. http://dx.doi.org/10.1037/xap0000119

Baker, R., & Inventado, P. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61–75). New York: Springer. http://dx.doi.org/10.1007/978-1-4614-3305-7_4

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

Bakharia, A., Corrin, L., de Barba, P., Kennedy, G., Gašević, D., Mulder, R., Williams, D., Dawson, S., & Lockyer, L. (2016). A conceptual framework linking learning design with learning analytics. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK ’16), 25–29 April 2016, Edinburgh, UK (pp. 329–338). New York: ACM. http:dx.doi.org/10.1145/2883851.2883944

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14. http://dx.doi.org/10.1207/s15327809jls1301_1

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Washington, DC: US Department of Education, Office of Educational Technology.

Booth, J., McGinn, K., Barbieri, C., Begolli, K., Chang, B., Miller-Cotto, D., Young, L., & Davenport, J. (2017). Evidence for cognitive science principles that impact learning in mathematics. In D. Geary, D. Bearch, R. Ochsendorf, & K. Koepke (Eds.), *Acquisition of Complex Arithmetic Skills and Higher-Order Mathematics Concepts* (Vol. 3, pp. 297–327). Cambridge, MA: Academic Press.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can’t be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 470–497). New York: Cambridge University Press.

Bryan, W., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological*

Review, 6(4), 345–375. http://dx.doi.org/10.1037/h0073117

Butler, A., Marsh, E., Slavinsky, J., & Baraniuk, R. (2014). Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review, 26*. http://dx.doi.org/10.1007/s10648-014-9256-4

Cantor, A., & Marsh, E. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory, 25*(2), 220–230. http://dx.doi.org/10.1080/09658211.2016.1152377

Carvalho, P., Braithwaite, D., de Leeuw, J., Motz, B., & Goldstone, R. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLOS ONE, 11*(3), e0152115. http://dx.doi.org/10.1371/journal.pone.0152115

Chatti, M., Dyckhoff, A., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5/6), 318–331. http://dx.doi.org/10.1504/ijtel.2012.051815

Chen, Z., Demirci, N., Choi, Y.-J., & Pritchard, D. (2017). To draw or not to draw? Examining the necessity of problem diagrams using massive open online course experiments. *Physical Review Physics Education Research, 13*, 010110. http://dx.doi.org/10.1103/PhysRevPhysEducRes.13.010110

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK ’12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 134–138). New York: ACM. http://dx.doi.org/10.1145/2330601.2330636

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683–695. http://dx.doi.org/10.1080/13562517.2013.827653

Cope, B., & Kalantzis, M. (2015a). Sources of evidence-of-learning: Learning and assessment in the era of big data. *Open Review of Educational Research, 2*(1), 194–217. http://dx.doi.org/10.1080/23265507.2015.1074869

Cope, B., & Kalantzis, M. (2015b). Interpreting evidence-of-learning: Educational research in the era of big data. *Open Review of Educational Research, 2*(1), 218–239. http://dx.doi.org/10.1080/23265507.2015.1074870

Crouch, C., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*, 970–977. http://dx.doi.org/10.1119/1.1374249

Day, S., Motz, B., & Goldstone, R. (2015). The cognitive costs of context: The effects of concreteness and immersiveness in instructional examples. *Frontiers in Psychology, 6*, 1876. http://dx.doi.org/10.3389/fpsyg.2015.01876

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. http://dx.doi.org/s13428-014-0458-y

Dietz-Uhler, B., & Hurn, J. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of Interactive Online Learning, 12*(1), 17–26.

Duval, E. (2011). Attention please!: Learning analytics for visualization and recommendation. In P. Long, G. Siemens, G. Conole, & D. Gašević (Eds.), *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK ’11), 27 February–1 March 2011, Banff, AB, Canada (pp. 9–17). New York: ACM. http://dx.doi.org/10.1145/2090116.2090118

Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Allen, J., Banks, J., Baranski, E., … & Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. http://dx.doi.org/10.1016/j.jesp.2015.10.012

Elias, T. (2011). *Learning analytics: Definitions, processes and potential.* http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Ellis, R. (2005). Principles of instructed language learning. *System, 33*, 209–224. http://dx.doi.org/10.1016/j.system.2004.12.006

Enyon, R. (2013). The rise of big data: What does it mean for education, technology, and media research? *Learning, Media and Technology, 38*(3), 237–240. http://dx.doi.org/10.1080/17439884.2013.771783

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., Reid, S., & LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Physics Education Research, 1*(1), 1.010103. http://dx.doi.org/10.1103/PhysRevSTPER.1.010103

Frank, M., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435. http://dx.doi.org/10.1111/infa.12182

Fryer, R. G., Jr. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics, 126*(4), 1755–1798. http://dx.doi.org/10.3386/w15898

Fyfe, E. (2016). Providing feedback on computer-based algebra homework in middle-school classrooms. *Computers in Human Behavior, 63*, 568–574. http://dx.doi.org/10.1016/j.chb.2016.05.082

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. http://dx.doi.org/s11528-014-0822-x

Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*(1), 68–84. http://dx.doi.org/10.1016/j.iheduc.2015.10.002

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2017). Categorization and Concepts. In J. Wixted (Ed.) *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 4th ed., Volume Three: Language & Thought* (pp. 275–317). New Jersey: Wiley.

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The student activity meter for awareness and self-reflection. *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 869–884). New York: ACM. http://dx.doi.org/10.1007/978-3-642-25813-8_20

Greeno, J. G., & Middle School Mathematics through Applications Project Group. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*(1), 5–26. http://dx.doi.org/10.1037/0003-066X.53.1.5

Guskey, T., & Huberman, M. (1995). *Professional development in education: New paradigms and practices*. New York: Teachers College Press.

Hall, G. (1891). The contents of children's minds on entering school. *The Pedagogical Seminary, 1*(2), 139–173. http://dx.doi.org/10.1080/08919402.1891.10533930

Heath, L., Kendzierski, D., & Borgida, E. (1982). Evaluation of social programs: A multimethodological approach combining a delayed treatment true experiment and multiple time series. *Evaluation Review*, *6*(2), 233–246. http://dx.doi.org/10.1177/0193841X8200600205

Heffernan, N., & Heffernan, C. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. http://dx.doi.org/10.1007/s40593-014-0024-x

Henninger, F., Mertens, U. K., Shevchenko, Y., & Hillbig, B. E. (2017). lab.js: Browser-based behavioral research. http://dx.doi.org/10.5281/zenodo.597045

Higgins, M., Sävje, F., & Sekhon, J. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7369–7376. http://dx.doi.org/10.1073/pnas.1510504113

Imai, K., King, G., & Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502. http://dx.doi.org/10.1111/j.1467-985X.2007.00527.x

Khalil, M., & Ebner, M. (2015). Learning analytics: Principles and constraints. In S. Carliner & N. Ostashewski (Eds.), *Proceedings of the World Conference on Educational Media and Technology* (EdMedia 2015), 22–24 June 2015, Montréal, Canada (pp. 1789–1799). Waynesville, NC: Association for the Advancement of Computing in Education (AACE). www.learntechlib.org/results/?q=Khalil&source=EDMEDIA%2F2015%2F1

Klahr, D. & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–667. http://dx.doi.org/10.1111/j.0956-7976.2004.00737.x

Kirchoff, B. K., Delaney, P. F., Horton, M., & Dellinger-Johnston, R. (2014). Optimizing learning of scientific category knowledge in the classroom: The case of plant identification. *CBE Life Sciences Education*, *13*(3), 425–436. http://dx.doi.org/10.1187/cbe.13-11-0224

Kizilcec, R., & Brooks, C. (2017). Diverse big data and randomized field experiments in MOOCs. In C. Lang, G. Siemens, A. Wise, and D. Gašević (Eds.), *Handbook of Learning Analytics* (pp. 211–222). Society for Learning Analytics Research. http://dx.doi.org/10.18608/hla17.018

Kizilcec, R., & Cohen, G. L. (2017). Eight-minute self-regulation intervention improves educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), 4348–4353. http://dx.doi.org/ 10.1073/pnas.1611898114

Kizilcec, R., Pérez-Sanagustín, M., & Maldonado, J. (2016). Recommending self-regulated learning strategies does not improve performance in a MOOC. *Proceedings of the 3rd ACM Conference on Learning @ Scale* (L@S 2016), 25–28 April 2016, Edinburgh, Scotland (pp. 101–104). New York: ACM. http://dx.doi.org/10.1145/2876034.2893378

Klein, R., Ratliff, K., Vianello, M., Adams Jr., R., Bahník, Š., Bernstein, M., Bocian, K., … & Nosek, B. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. Baker, *Handbook of Educational Data Mining.* Boca Raton, FL: CRC Press.

Koedinger, K., Booth, J., & Klahr, D. (2013a). Instructional complexity and the science to constrain it. *Science, 342*(6161), 935–937. http://dx.doi.org/10.1126/science.1238056

Koedinger, K., Corbett, A., & Perfetti, C. (2012). The knowledge–learning–instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798. http://dx.doi.org/10.1111/j.1551-6709.2012.01245.x

Koedinger, K., & McLaughlin, E. (2016). Closing the loop with quantitative cognitive task analysis. In T. Barnes et al. (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 412–417). International Educational Data Mining Society.

Koedinger, K., Stamper, J., McLaughlin, E., & Nixon, T. (2013b). Using data-driven discovery of better student models to improve student learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 421–430). Springer.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development.* Upper Saddle River, NJ: Prentice Hall.

Krause, M. (2010). Undergraduates in the archives: Using an assessment rubric to measure learning. *The American Archivist, 73*, 507–534. http://dx.doi.org/10.17723/aarc.73.2.72176h742v20l115

Kumar, V., Clemens, C., & Harris, S. (2015). Causal models and big data learning analytics. In Kinshuk & R. Huang (Eds.), *Ubiquitous learning environments and technologies* (pp. 31–53). Springer.

Lader, E., Cannon, C., Ohman, E., Newby, L., Sulmasy, D., Barst, R., Fair, J., Flather, M., Freedman, J., Frye, R., Hand, M., Van de Werf, F., Costa, F., & American College of Cardiology Foundation (2004). The clinician as investigator: Participating in clinical trials in the practice setting. *Circulation, 109*, 2672–2679. http://dx.doi.org/10.1161/01.CIR.0000128702.16441.75

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action aligning learning analytics with learning design. *American Behavioral Scientist, 57*(10), 1439–1459. http://dx.doi.org/10.1177/0002764213479367

Lodge, J., & Corrin, L. (2017). What data and analytics can and do say about effective learning. *npj Science of Learning, 2*(5). http://dx.doi.org/10.1038/s41539-017-0006-5

Macfadyen, L., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588–599. http://dx.doi.org/10.1016/j.compedu.2009.09.008

Maxwell, J. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11. http://dx.doi.org/10.3102/0013189X033002003

Medin, D., & Bang, M. (2014). *Who's asking? Native Science, Western Science and Science Education.* Cambridge, MA: MIT Press.

Moore, E. B., Herzog, T. A., & Perkins, K. K. (2013). Interactive simulations as implicit support for guided-inquiry. *Chemical Education Research and Practice, 14*, 257–268.

Morgan, K., & Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics, 40*(2), 1263–1282. http://dx.doi.org/10.1214/12-AOS1008

Morrison, K., & van der Werf, G. (2016). Large-scale data, "wicked problems," and "what works" for educational policy making. *Educational Research and Evaluation, 22*(5/6), 255–259. http://dx.doi.org/10.1080/13803611.2016.1259789

Motz, B., Teague, J., & Shepard, L. (2015). Know thy students: Providing aggregate student data to instructors. *EDUCAUSE Review, 3*. https://er.educause.edu/articles/2015/3/know-thy-students-providing-aggregate-student-data-to-instructors

Murnane, R., & Willett, J. (2010). *Methods matter: Improving causal inference in educational and social science research.* New York: Oxford University Press.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering.* Washington, DC: National Academies Press.

Norris, D., Baer, L., Pugliese, L., & Lefrere, P. (2008). Action analytics: Measuring and improving performance that matters in higher education. *EDUCAUSE Review, 43*(1), 42–67. https://er.educause.edu:443/articles/2008/1/action-analytics-measuring-and-improving-performance-that-matters-in-higher-education

Patsopoulos, N. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience, 13*(2), 217–224.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, J., & Verma, T. (1995). A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics, 134*, 789–811. http://dx.doi.org/10.1016/S0049-237X(06)80074-1

Plaisant, C. (2004). The challenge of information visualization evaluation. *Proceedings of the 2nd International Working Conference on Advanced Visual Interfaces* (AVI '04), 25–28 May 2004, Gallipoli, Italy (pp. 109–116). New York: ACM. http://dx.doi.org/10.1145/989863.989880

Rahman, S., Majumder, M., Shaban, S., Rahman, N., Ahmed, M., Abdulrahman, K. B., & D'Souza, U. (2011). Physician participation in clinical research and trials: Issues and approaches. *Advances in Medical Education and Practice, 2*, 85–93. http://dx.doi.org/10.2147/AMEP.S14103

Reich, J. (2015). Rebooting MOOC research: Improve assessment, data sharing, and experimental design. *Science (Education Forum), 347*(6217), 34–35. http://dx.doi.org/10.1126/science.1261627

Renz, J., Hoffmann, D., Staubitz, T., & Meinel, C. (2016). Using A/B testing in MOOC environments. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 304–313). New York: ACM. http://dx.doi.org/10.1145/2883851.2883876

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. http://dx.doi.org/10.1037/a0026252

Russo, F. (2010). *Causality and causal modeling in the social sciences.* Springer.

Severance, C., Hanss, T., & Hardin, J. (2010). IMS learning tools interoperability: Enabling a mash-up approach to teaching and learning tools. *Technology, Instruction, Cognition, & Learning*, *7*, 245–262.

Shadish, W., Campbell, D., & Cook, T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 4–8). New York: ACM. http://dx.doi.org/10.1145/2330601.2330605

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. http://dx.doi.org/10.1177/0002764213498851

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge, MA: MIT Press.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. http://dx.doi.org/10.1016/j.tics.2017.06.007

Sullivan, G. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Training, 3*, 285–289. http://dx.doi.org/10.4300/JGME-D-11-00147.1

Tervakari, A., Silius, K., Koro, J., Paukkeri, J., & Pirttila, O. (2014). Usefulness of information visualizations based on educational data. *Proceedings of the 2014 Global Engineering Education Conference* (EDUCON 2014), 3–5 April 2014, Istanbul, Turkey (pp. 142–151). IEEE Computer Society. http://dx.doi.org/10.1109/EDUCON.2014.6826081

Tufte, E. (2003). *The cognitive style of PowerPoint.* Cheshire, CT: Graphics Press.

US Department of Education. (2016). *Using evidence to strengthen education investments (Non-regulatory guidance).* Washington, DC. https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

US Department of Education. (2017). *What works clearinghouse standards handbook, Version 4.0*. Washington, DC: Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509. http://dx.doi.org/10.1177/0002764213479363

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. *Science, 322*(5902), 682–683. http://dx.doi.org/10.1126/science.1161948

Williams, J., & Williams, B. (2013). Using randomized experiments as a methodological and conceptual tool for improving the design of online learning environments. http://dx.doi.org/10.2139/ssrn.2535556

Wise, A., & Shaffer, D. (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

Zheng, Z., Vogelsang, T., & Pinkwart, N. (2015). The impact of small learning group composition on student engagement and success in a MOOC. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 500–503). International Educational Data Mining Society.