JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 9(3), 183–199. https://doi.org/10.18608/jla.2022.7577

Causal Inference and Bias in Learning Analytics: A Primer on Pitfalls Using Directed Acyclic Graphs

Joshua Weidlich[1], Dragan Gašević[2], Hendrik Drachsler[3]

Abstract

As a research field geared toward understanding and improving learning, Learning Analytics (LA) must be able to provide empirical support for causal claims. However, as a highly applied field, tightly controlled randomized experiments are not always feasible nor desirable. Instead, researchers often rely on observational data, based on which they may be reluctant to draw causal inferences. The past decades have seen much progress concerning causal inference in the absence of experimental data. This paper introduces directed acyclic graphs (DAGs), an increasingly popular tool to visually determine the validity of causal claims. Based on this, three basic pitfalls are outlined: confounding bias, overcontrol bias, and collider bias. Further, the paper shows how these pitfalls may be present in the published LA literature alongside possible remedies. Finally, this approach is discussed in light of practical constraints and the need for theoretical development.

Notes for Practice

- If LA researchers want to investigate learning processes or make claims about the effectiveness of their interventions, they need causal knowledge.
- To arrive at causal knowledge from nonexperimental research data, a principled approach to reason about causal assumptions and make informed decisions is needed.
- This paper shows how directed acyclic graphs (DAGs) can be used for this purpose.
- Using examples from the published LA literature, potential pitfalls in causal inference are discussed and possible remedies are presented.

Keywords

Learning analytics, LA, causal inference, directed acyclic graphs, DAG, research design, observational research, bias

Submitted: 13/09/2021 — **Accepted:** 09/11/2022 — **Published:** 13/12/2022

Corresponding author: [1]Email: j.weidlich@dipf.de Address: Educational Technologies, DIPF — Leibniz Institut for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany. ORCID ID: https://orcid.org/0000-0002-1926-5127
[2]Email: dragan.gasevic@monash.edu Address: Faculty of Information Technology, Monash University, Clayton 3800, Australia. ORCID ID: https://orcid.org/0000-0001-9265-1908
[3]Email: h.drachsler@dipf.de Address: Educational Technologies, DIPF — Leibniz Institut for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany. ORCID ID: https://orcid.org/0000-0001-8407-5314

1. Introduction

The burgeoning field of Learning Analytics (LA) is concerned with improving student learning. Central to this endeavour is arriving at an *actionable insight* or *actionable intelligence* (Clow, 2013; Jørnø & Gynther, 2018), usually gleaned from data-intensive methods of modelling and predicting student learning in online environments. However, the step from predicting learning success via data-driven modelling to improving learning success via data-informed learning design necessitates causal knowledge about the effects of interventions. However, as others have noted, robust causal knowledge remains relatively scarce in LA (Motz et al., 2018; Wong et al., 2019).

Causal knowledge cannot be derived from mere analysis of data, as the mechanistic reality we are interested in, and the data generated from it are fundamentally distinct (Bareinboim et al., 2020). The leap from patterns in data to causal claims is a qualitative one, needing information from "outside" the data (Hernán et al., 2019). Such additional information comes in the form of reasoning about the data-generation process. Without it, claims of, for example, effective analytics implementations or improved understanding of learning processes must remain questionable (Prosperi et al., 2020). For example, a statistical comparison of two distributions becomes a vehicle for causal claims if we learn the data were generated in a randomized

experiment. On the other hand, lacking this information or finding reasons for doubt (e.g., indication of failed randomization or incomplete blinding), we decrease our belief in causal claims derived from this data. Causal assumptions are another type of "outside" information that helps us reason about the data-generation process. These types of theory- or experience-based assumptions (e.g., which variables causally affect which other variables) can be encoded in graphical models and generated with domain knowledge, which — under some conditions — allow for approximation of causal effects, even from observational data (Robins, 2001; Hernán et al., 2019).

The past decades have seen much progress in the field of causal inference, most prominently spearheaded by Judea Pearl (Pearl, 1995; Pearl, 2009; Pearl et al., 2016; Pearl, 2021). Central to this is the use of directed acyclic graphs (DAGs), a way of non-parametrically encoding causal assumptions within a set of variables of interest. Like the recent introduction of DAGs to the field by Hicks et al. (2022), this paper aims to give a non-technical introduction to how causal questions in LA can be graphically displayed in DAGs. While Hicks et al. (2022) show how DAGs can be used to communicate across discipline lines and provide an example from the area of student retention, our primer focuses on potential examples of bias in the published LA literature. We use these illustrative examples to show how they may be depicted in DAGs, how this may affect inferences and, finally, what principled methods of circumventing these pitfalls are available.

2. Actionable Insights and Causal Inference in Learning Analytics

Learning Analytics is frequently defined by referring to two underlying goals: understanding learning and optimizing learning (Siemens & Gašević, 2012). Similarly, the often-cited goal of arriving at actionable insights (Jørnø & Gynther, 2018) is rooted in the notion of LA being an applied field concerned with taking practical steps toward improving learning instead of stopping at the description and prediction of learning or student success (Cooper, 2012). In fact, where the "insight" component may be satisfied with modelling learning processes, the "actionable" component strongly implies the need for available measures known to be effective. This is also well represented in Clow's (2013) learning analytics cycle, where data lead to metrics, and metrics inform appropriate interventions.

However, what is not encoded in this cycle, is that the step from metrics to intervention is a particularly difficult one. Choosing an effective intervention based on certain metrics of learning implies that we have a causal model about 1) how the metric emerged in the first place and 2) how an intervention will affect the process responsible for the generation of the metric. One may be tempted to believe that this crucial information can be found in the data itself, but this is a fallacy, as description and prediction are epistemologically on a different plane than reasoning about cause and effect (Hernán et al., 2019). Thus, it would be a mistake to only use data-driven prediction models to arrive at causal claims of what can be done to improve learning. This mistake is increasingly recognized in epidemiology, health care, and other fields drawing on machine learning and related technologies (Prosperi et al., 2020; Pearl, 2019), and thus, logically extends to LA, as well as Educational Data Mining and Artificial Intelligence in Education (Baker et al., 2021).

A particularly well-known vehicle for causal claims is the simple information of whether the data was generated from a randomized controlled experiment (RCT: randomized control trial) or not. If the answer is "yes," then we are usually confident about drawing a causal conclusion from the data; if the answer is "no," we are hesitant, even though the data has not changed. Thus, the validity of claims hinges on this qualitative, outside information. However, as shown later in this paper, there is no magic ingredient that makes data from RCTs vehicles for causal inferences. In fact, even RCTs cannot deliver what we would *ideally* want for causal inferences: counterfactual data (Morgan & Winship, 2015). That is, data points from two parallel worlds, one in which a student has received a treatment and one in which the exact same student has not. In the absence of counterfactual data — which is impossible to obtain — we rely on an approximation via randomization. Thus, the RCT is merely a special case of a larger framework of causal reasoning; a special case, admittedly, with some effective embedded assumptions.

Despite RCTs being acknowledged as the gold standard for testing causal claims in many areas of the behavioural sciences, LA has been comparatively reluctant to adopt this tool for evaluating the effectiveness of its interventions, as reported in recent systematic reviews (Sønderlund et al., 2018; Viberg et al., 2018). Discipline-independent reasons may be found in a hesitancy to generalize from RCTs, given that imposing control over humans and social processes for the benefit of internal validity (strength of causal claims) is detrimental to external validity (transferability across target groups and contexts). Further, not all education researchers are equally convinced of the suitability of RCTs for certain aspects of the study of human learning, as RCTs may be considered limited in providing valuable information to learners as individuals (e.g., Winne 1982; 2017). Nonetheless, specific hallmarks of RCTs are generally accepted in the research community as a means to provide evidence for causal relations and the efficacy of treatments in particular.

As a result, there have been calls for the increased application of experiments to improve the strength of causal claims in LA (Motz et al., 2018; Wong et al., 2019). A discipline-specific reason for the heretofore dearth of experiments may be found in the fact that LA is understood to be a rather applied field of research, thus, not lending itself readily to tightly controlled

experiments. Besides this, we can conjure two further reasons why the field of LA may, in the past, have been less worried about drawing causal inferences from observational data than, say, education research or psychology. First, LA frequently deals with larger amounts of data than education research or psychology. Researchers and practitioners may be compelled to believe in the validity of their inferences drawn from this data, as common constraints on validity, like lack of representativeness and low statistical power, appear less applicable. Second, LA research data are often behavioural data gleaned from the learning environment (e.g., LMS) instead of self-reporting. This lends credence to the notion that the collected data are objective and, thus, less prone to human biases (e.g., halo effect, social desirability, lack of metacognitive awareness), although there is evidence to suggest log and process data also provide challenges in terms of construct validity and measurement reliability (Zhou & Winne, 2012; Winne, 2020).

While using big sets of log and process data in LA may improve our confidence in the validity of the description and prediction of student learning, even the largest behavioural dataset and the most sophisticated algorithm will fall short of drawing valid causal inferences toward improving student learning without additional information. What is needed to draw these conclusions is information from outside of the data, and it is qualitative (Hernán et al., 2019). The framework by Pearl (2009) allows for reasoning about causal inference with this "outside" information. Thus, the following will outline a principled approach to improving causal inferences from data generated by different research designs, among them observational data. In this primer, we will focus on the "back-door criterion" (Pearl, 2009; Pearl et al., 2016; Pearl, 2021), the simplest way to use DAGs to assess the possibility of bias and to derive countermeasures.

3. Introducing Directed Acyclic Graphs

Causal inferences can be improved by reasoning about causal assumptions with the help of directed acyclic graphs (DAGs). DAGs are an intuitive way of visually representing causal assumptions, like in structural equation modelling (SEM), essentially by drawing variables, boxes, and arrows. Judea Pearl has been pioneering this approach since the 1990s (Pearl, 1995), and although Pearl, his collaborators, and subsequent researchers have now developed comprehensive frameworks for causal inferences using DAGs (Pearl et al., 2016; VanderWeele, 2015; Hernán & Robins, 2020; Morgan & Winship, 2015), we focus here on a few key ideas that can help us arrive at better causal inferences, no matter the research design at hand.

Importantly, and in contrast to typical SEMs, DAGs are nonparametric in that they make no assumptions about distribution or functional form (Elwert & Winship, 2014). For example, an arrow pointing from one variable to the other could indicate a linear, quadratic, or exponential function. Their primary function is not estimation — although estimation of causal effects is the ultimate goal — but identification of causal effects, which is done graphically by inspecting the causal assumptions represented in a DAG. Thus, nonparametric identification is akin to qualitatively asking the DAG, "Can we estimate the causal effect from the data (provided our causal assumptions are correct)?" In practice, however, it is unlikely that causal assumptions will be "correct" in a binary sense. Instead, researchers may aim for a high degree of correctness of their DAG. Crucially, the construction of a DAG (and its degree of correctness) relies on domain knowledge to convey the causal assumptions from "outside" the data. With the help of such substantive knowledge, e.g., experience and theory, experts can construct DAGs that convincingly depict the causal assumptions in the nomological net of the variables of interest. With this DAG at hand, we can then reason about sources of bias, the potential for confounding, and approaches for controlling or purposefully avoiding control of these factors, either statistically or by design.

The following subsections first provide an example from MOOC research, in which a DAG is used to reason about a simple but common research context with confounding present. Following this example, the three most fundamental causal configurations are described, how they are represented in a DAG and what they imply for causal inference. Then, using a more complex example from research on LA dashboards, it is shown how pitfalls stemming from these configurations can be overcome by controlling for the appropriate variables. A brief overview of statistical and experimental control concludes this introduction to DAGs, thereby laying the foundation for the remainder of this paper.

3.1. Motivating Example from MOOC Research

This section will provide a hypothetical example of how a DAG may be used for causal reasoning, demonstrating how employing relevant domain knowledge to construct DAGs may lead to improved causal inferences. Crucially, this is a fictitious and highly simplified scenario. We do not suggest that this and the following fictitious examples paint a comprehensive picture of the relationships at hand. Instead, they should be interpreted as illustrative examples to introduce causal reasoning with DAGs.

A fictitious team of researchers has collected data from a large MOOC and wants to understand which student behaviours predict MOOC completion to modify the learning design of the course to improve student success in future iterations. For this example, they collect observational data about MOOC completion and study behaviour (e.g., frequency of access to learning material, login times). The researchers draw a preliminary DAG detailing their hypothesis, as in Figure 1.

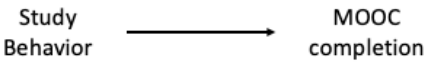

Figure 1. DAG representing the research interest of MOOC researchers in this example.

Grounding this in the adopted model of Joksimović et al. (2018), the hypothesized causal variable[1] study behaviour as it is conceptualized here is akin to behavioural engagement. Further, its indicators are likely continuous (e.g., number of discrete accesses, amount of time spent in the environment) and expected to be causally related to MOOC completion, a course-level outcome indicator (Joksimović et al., 2018), measured on a binary scale (completion y/n). Because DAGs are non-parametric, the level of measurement (binary, continuous) is not encoded visually. However, the causal assumptions are, so the researchers draw an arrow from this hypothesized causal variable to the dependent variable to indicate the causal assumption (this is the "directed" part of DAGs). At this point, it could be argued that most researchers would not make a causal assumption from observational data. Yet, this would be at odds to improve student success based on this research. So even though the researchers may not be able to conduct an RCT, it is still the causal effect they are ultimately interested in. Arguably, this is true for any research goal that attempts to go beyond prediction toward understanding and/or improving learning.

Once the necessary statistical analyses are conducted, the researchers obtain an estimate of the correlations and report the associations, taking care to contain their language so as not to suggest unwarranted causality, as is custom (Grosz et al., 2020). Yet, careful language does not help them here, as their ultimate goal is to learn how to improve course completion, a decidedly causal undertaking. Thus, they need to know to what extent the hypothesized causal variable would affect the outcome variable if they decided to redesign or modify the course accordingly.

A frequently plausible objection to causal claims from observational data is confounding. Because the researchers in this MOOC project are education researchers, they infer that self-regulated learning skills (SRL) may play a role in MOOC completion (Littlejohn et al., 2016; Moreno-Marcos et al., 2020). They hypothesize that SRL may be a confounder, a variable affecting what they expect to be the causal variable and the outcome, thereby introducing a spurious correlation between them. They draw a DAG to depict this assumption (Figure 2), in which SRL affects both study behaviour as well as MOOC completion. According to this, developing an intervention that merely changes aspects of study behaviour (e.g., higher frequency of access to material, increased time spent on the platform) could be a waste of resources because the causal effect of interest is still unknown. This is because a confounder introduces spurious associations between variables, and the researchers would not know whether to attribute the estimated correlation coefficient to the true causal effect or the spurious part of association. In the language of causal inference, this means that the causal effect is not *identified* (Pearl, 1993). A non-identified causal effect is the non-parametric equivalent of model misspecification. Model misspecification in multiple regression occurs when a relevant variable is omitted from the model, or an irrelevant variable is included, with varying effects on bias and error in the estimation, as demonstrated in Rao (1971). Similarly, when speaking of non-identified causal effects, we are stating in general terms that a statistical estimate will not reflect the true causal effect because our causal net of variables suggests a source of bias introducing spurious associations.

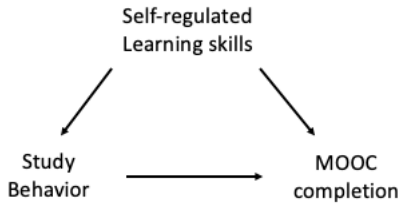

Figure 2. DAG representing the confounding situation in this MOOC research example.

Responding to this problem, the researchers could then try to statistically control for SRL. A straightforward way to do this is to collect self-report data on SRL from MOOC students in their next round of data collection. During analysis, statistical control would then be done, for example, by introducing the variable SRL into the regression model, in effect assessing the association between the hypothesized causal variable and the outcome variable while subtracting the common cause of SRL.

[1] The term *hypothesized causal variable* is used throughout the manuscript because it communicates the causal assumptions that form the basis of causal reasoning. We use this label instead of *independent variable*, which presumes experimental manipulation, and the label *predictor variable*, which usually refers to non-manipulated variables. In contrast, hypothesized causal variables are agnostic to their methods of investigation. When explicitly talking about experiments or regression-based statistical modelling, the appropriate terms will be used instead of this label.

In an idealized world in which SRL is the only confounder, any remaining association between the variables of interest would be the causal effect. Then, depending on the presence of such an effect, a successful intervention could consist of prompting students to access learning materials more frequently or spend more time in the MOOC environment. This is often not the case, and de-confounding can be more difficult than this. Not only are there usually more potential confounders to consider, but there are also situations in which controlling for a variable may itself introduce spurious associations. However, if we conceive of the validity of our claims not as dichotomous but continuous, this does not invalidate efforts at reducing bias systematically. Indeed, making principled decisions about when to control, what to control for, and when to avoid control is important for making convincing causal claims in a diverse research setting and can be easily derived by inspecting DAGs.

3.2. Elemental Causal Configurations in DAGs

DAGs usually are more complex than the example presented above. However, even very complex DAGs consist of a small set of typical configurations with specific implications: chains, forks, and inverted forks.

3.2.1. Chains

In the example in Figure 1, we have seen the simplest configuration, A→B, or *access to learning material → MOOC completion,* with two nodes and one arrow. Slightly longer, a chain consists of three nodes and two arrows A→B→C or, for example, *access to learning material → materials studied → MOOC completion.* In this case, B is a mediator between A and C because completing the MOOC would be dependent on studying the material, which itself is dependent on accessing the materials. This means that if we assume that this DAG is exhaustive, the association between A and C represents the causal effect, which is "transmitted" via the mediator B. Crucially, should we decide to control for B in a chain A→[B]→C (controlled variables are indicated by a box), we effectively block the mediator from transmitting the causal effect. This makes sense because if we only look at students with identical amounts of material studied ([B]), the effect of accessing the material cannot be "transmitted" to MOOC completion. This is called overcontrol bias (Elwert & Winship, 2014) or overadjustment bias (Schisterman et al., 2009) and usually leads to an attenuation in the estimation of causal effects. Thus, under these conditions, the estimated effect would be smaller than the true causal effect.

3.2.2. Forks

Like our example in Figure 2, forks have the structure A←B→C, or, for example, *access to learning material ← self-regulated learning skills → MOOC completion*. In this case, B is a common cause of A and C, and if we estimate the effect of A on C, we encounter a non-causal association. This is because variables that share a common cause are associated, constituting bias due to confounding. As laid out in the example, controlling for B in A←[B]→C prevents this non-causal transmission because, in this situation, [B] is no longer a common cause of A and C. In other words, the researchers in this example could attempt to statistically control for self-regulated learning skills of students to remove confounding and estimate the effect of access to learning material on course completion. Under these conditions, the causal effect of A on C is "identified," and estimates of the true causal effect no longer contain spurious associations.

3.2.3. Inverted Forks

Finally, we have inverted forks, a particularly tricky and underappreciated configuration A→B←C, or, for example, *Self-regulated learning skills → MOOC success ← Prior knowledge.* In inverted forks, the B variable is called a collider because it is a variable with more than one incoming arrow. As colliders, or common effects, have the unique property of naturally blocking spurious associations, travelling along the inverted fork from A to C, we do not encounter a non-causal association. Considering the above example, this substantively makes sense because we have no reason to believe that SRL and prior knowledge (PK) should be associated just because they both have an influence on MOOC success. Crucially, however, any type of control on a collider introduces a spurious association, where A will be correlated with C despite there being no true causal connection. If we are not aware of the collider status of MOOC success, we may be tempted to control for this variable, thinking that we are controlling a confounding variable. Or, more likely, in this example, we may be controlling this variable without knowing it, for example, through selection bias via MOOC attrition. If we, for example, collect data by sampling only students who have successfully completed the MOOC, we are effectively looking only at a subset of the actual student population. If we assume that *either* self-regulated learning skills *or* prior knowledge are sufficient for completing the MOOC, then our subset includes students with high degrees of SRL and PK, students who have high SRL and low PK, as well as students with low SRL and high PK. What is missing from the sample are students low on both SRL and PK, as it is unlikely for them to complete the MOOC. This will yield a negative, yet entirely spurious, association between SRL and PK in the sampled population.

Conditioning on a collider, as it is called, or endogenous selection bias is an increasingly recognized pitfall that can have devastating effects on the validity of causal claims (Elwert & Winship, 2015; Munafò et al., 2018; Richardson et al., 2019). The dangers of conditioning on a collider lays bare the problematic practice of indiscriminately including any multitude of variables that are correlated with both predictor and dependent variable in a statistical model, ostensibly to avoid confounding.

This unprincipled approach, sometimes called garbage-can regression (Achen, 2005) was already criticized by Meehl (1970) and its perils can now be anticipated within the graphical causal inference framework presented here (Pearl, 2009; Lee, 2012; Rohrer, 2018).

Importantly, these sources of bias are independent of whether the effects are causally interpreted by researchers or not. This is because the underlying mechanistic reality itself is causal and, to the extent that these sources of bias are present, parameter estimates will similarly be biased. This makes them essentially uninterpretable (Antonakis et al., 2010). For these reasons, researchers from different disciplines increasingly argue that careful language (Grosz et al., 2020) and scientific euphemisms (Hernán, 2018) in observational research are non-solutions to the problem of causal inference.

3.3. More Complex Example from LA Dashboards Research

Another fictitious team of researchers has developed an ambitious LA dashboard that presents dynamic performance metrics to students, hoping that this information will influence studying behaviour, and ultimately, learning achievement. Because these researchers 1) do not have the resources to conduct an RCT and 2) have not collected data in the past semesters so that they could compare learning achievement before and after deploying the dashboard feature in a quasi-experiment, they too are limited to observational data. As earlier, they hypothesize that SRL may play a confounding role, since research indicates that SRL affects both dashboard use (Jivet et al., 2020) and learning achievement (Zimmerman, 1990). Moreover, they hypothesize that SRL not only directly affects learning achievement, but is partly mediated by students' study behaviour. The researchers now inspect the DAG shown in Figure 3 to assess if the desired causal effect is identified; that is, if all non-causal paths between IV and DV are blocked.

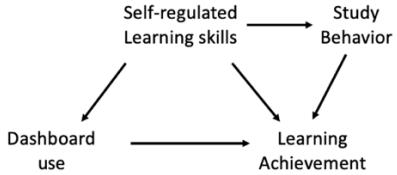

Figure 3. DAG representing the causal assumptions in a Learning Analytics Dashboard study.

Tracing the possible paths from dashboard use to learning achievement, we find two "back-door paths" (Pearl, 1993), 1) dashboard use ← SRL → learning achievement and 2) dashboard use ← SRL → study behaviour → learning achievement. These are back-door paths because they start with an arrow pointing to the hypothesized causal variable and end with an arrow pointing to the dependent variable. As both back-door paths in this example consist of only chains and forks, there is no inverted fork (with a collider) naturally blocking transmission of a spurious association. Due to this confounding situation, the researchers conclude that the causal effect is not identified and zero-order correlations between dashboard use and learning achievement would yield a nebulous mix of the true causal effect as well as additional non-causal associations travelling via back-door paths.

Knowing that they can block a back-door path — thus preventing it from transmitting non-causal information — by controlling any one variable on this path, they start looking for candidate variables to statistically control for. The researchers figure out that controlling for study behaviour would not suffice as this would only block *dashboard use* ← *SRL* → *study behaviour* → *learning achievement* but leave open *dashboard use* ← *SRL* → *learning achievement.* Controlling SRL, on the other hand, would do both by blocking *dashboard use* ← *SRL* → *study behaviour* → *learning achievement* and *dashboard use* ← *SRL* → *learning achievement.* If the researchers assume that their DAG in Figure 3 exhaustively captures the causal web of variables with a high degree of veracity, controlling for SRL would suffice to identify the causal effect of dashboard use and learning achievement so that the remaining correlation would be an estimate of the true causal effect. Section 3.4 will briefly outline the different ways that researchers may go about controlling confounders in their research.

3.4. Statistical and Experimental Control

We can distinguish between two approaches to control: statistical control and experimental control. Though they can sometimes lead to similar results in terms of causal inferences, they differ in terms of the underlying mechanism of control (Knight & Winship, 2013). Common methods of statistical control described here include group-specific analysis, third variable inclusion, and matching. For each of these methods, strengths and shortcomings will be briefly discussed. Importantly, statistical controlling for a variable requires that it has been measured in some way, highlighting the need for working with DAGs in the design stage of a research study.

3.4.1. Statistical Control

Group-specific analyses work by stratifying the sample and then conducting analyses within each subset, or *strata* (Tripepi et al., 2010; Kahlert et al., 2017). A classic example of this is estimating an effect for men and women separately. We can conceive of research hypotheses where biological sex is a confounding variable. Stratifying our analysis along biological sex would, in essence, block this non-causal path because only women are compared to women and we then no longer are worried that gender may be an alternative explanation for our effect. Alternatively, as reported in Russell et al. (2020), at-risk status (moderate vs. high-risk) may be considered a confounder, such that the analyses of the effects of an LA application on outcome variables are conducted separately for these student groups. This approach works best for categorical variables but becomes less useful with more continuous variables, as defined strata may be the result of arbitrary cutoffs. Also, with more stratification variables, the number of strata quickly increases, limiting the effective sample size and, thus, statistical power.

A second popular approach is the inclusion of one or multiple third variables in a regression model, and examining the regression coefficient of the predictor variable beyond what the third variables (or covariates) explain. An example of this can again be found in Russell et al. (2020), where researchers controlled potentially confounding variables like prior learning outcomes, demographics, and study skills via covariates in a regression model to estimate the effect of their LA intervention on outcome variables. This approach avoids the issues of stratification by allowing for the inclusion of many confounders. However, this approach comes with strong assumptions; for example, the linearity between all confounders and the dependent variable (McNamee, 2005). Two other practical issues can emerge when attempting to control via third variable inclusion: the Table 2 fallacy and distortions of construct validity. The Table 2 fallacy (Westreich & Greenland, 2013) refers to incorrectly interpreting multivariate-adjusted associations that were included to control for confounding. While confounding may be removed, leading the causal effect of interest to be identified, estimates for the confounders themselves should not be interpreted in the same way, that is, as predictor variables. Distortion of construct validity refers to the phenomenon that residualization of the variables in the statistical model implies a transformation of the construct itself, as can be seen in changed discriminant and convergent validity (Winne, 1983). Thus, their parameters should not be interpreted at face value.

Matching is another common approach to statistical control and works by purposefully creating groups that are similar with respect to potentially confounding variables. In other words, researchers aim to approximate the effect of randomization by assigning participants to groups based on variables that are expected to exert a confounding effect. The matching procedure is data-based and can be based on a multitude of variables. In a situation of perfect matching, these variables could no longer be the cause of any differences in the dependent variable. A widely used method of this approach is propensity score matching (Rosenbaum & Rubin, 1983; Austin, 2011), which has also been employed in the LA literature (Lim et al., 2021). However, propensity score matching has been associated with shortcomings (King & Nielsen, 2019), while other matching procedures like nearest neighbour matching (using Mahanalobis Distance) and (coarsened) exact matching having been shown to avoid these shortcomings (King et al., 2011). Indeed, a comparison of matching methods in the context of assessing the efficacy of automated feedback showed that nearest neighbour matching using Mahanalobis Distance was the most suitable approach and outperformed the more common Propensity Score approach (Mousavi et al., 2021).

Finally, regression discontinuity can be used when treatments are assigned based on a threshold or cutoff value of the so-called *running variable.* Observations close to either side of the cutoff will be comparable with respect to the running variable as well as all variables that determine the running variable. In these cases, the causal effect can be assumed to be identified, because all differences in the dependent variable will likely be due to the treatment. As an example, if the provision of an educational remedial program is based on standardized math and reading tests (as reported in Jacob & Lefgren, 2004), students who just barely were included and those who were barely excluded can be used as observations to identify the causal effect of the remedial program. An example of this in the area of LA can be found in the study by Klenke et al. (2021), where the authors assess the efficacy of an early warning system for students with the help of regression discontinuity. An extension of this is *fuzzy* regression discontinuity design, where the cutoff value does not fully determine the treatment but rather the likelihood of treatment (Bloom, 2012). A major drawback of this approach is the limited statistical power that results from considering only the slice of data around the cutoff value. On the other hand, the further we move away from the cutoff (i.e., to increase power), the less defensible the assumption of comparability becomes. Further, in many research contexts, a natural threshold of the running variable will not be available. In these cases, researchers must decide on a cutoff themselves, a practice that, similar to median splits, may be deemed artificial and difficult to defend.

Common to all approaches to statistical control reviewed here is that their utility in reducing bias is proportional to their ability to measure a given variable validly and reliably. Thus, attempts to reduce bias from observational data are influenced by measurement error. For example, if a putative confounding variable cannot be measured with high reliability, this negatively affects our ability to reduce spurious association via statistical control (Westfall & Yarkoni, 2016). Practically, this suggests that statistical control will be a more appealing option in situations where well-understood variables can be enlisted via well-established psychometric methods.

3.4.2. Experimental Control

Experimental control is different from statistical control because it does not block or unblock causal paths; it deletes them from the DAG entirely. By manipulating the hypothesized causal variable (i.e., making it a *truly* independent variable), we take control of the mechanism that generates it and this — in theory — leaves no other causes remaining (Lee, 2012). For example, if we are worried that student characteristics influence both the usage of a technological intervention (e.g., a learning analytics dashboard) and the dependent variable the intervention is targeting (e.g., learning achievement), we can randomly assign students to two conditions, one in which students use the intervention as part of their learning activities and one in which the intervention is not present at all.

This was done in Hellings & Haelermans (2020), where computer science students were randomly assigned to two groups that either received weekly emails with access to a learning analytics dashboard or did not receive these updates. The goal of this experimental assignment was to delete all arrows pointing toward the usage of the intervention because there are no longer any causes for the intervention aside from randomness which, by definition, is not a systematic cause. Thus, the de-confounding power of RCTs lies in the ability to eliminate back-door paths by deleting arrows into the independent variable. In Hellings & Haelerman (2020), however, this was not entirely successful, as some students in the control group refrained from accessing their dashboard, in effect remaining without intervention, for reasons unknown but unlikely to be random. For this reason, unobserved causes influence both dashboard use as well as the dependent variables, thus presenting a confounding situation similar to Figure 2.

Another example of experimental control is seen in natural experiments. Here, randomization occurs in real-world settings without researchers deliberately controlling this randomization. An example of this can be found in Mullaney & Reich (2015), where a MOOC platform changed their content release strategy and the effect of this change was assessed in terms of outcome variables. As this example shows, *actual* randomness is not strictly necessary to generate causal inferences from natural experiments; it suffices to find observations whose treatment status is not generated through a systematic mechanism that may serve as a back-door path. In this case, the content-release strategy of the provider reasonably did not affect who decided to subscribe to a course. Thus, we expect no arrows pointing at the hypothesized causal variable, yielding an experimental control without actual randomness.

It should be noted that although experimental control circumvents many challenges that plague statistical control, the effectiveness of a manipulation is determined by the extent to which stringent operationalization of the independent variable is available. In other words, while randomized experiments allow a degree of confidence that a causal effect is present, further, more restrictive assumptions are needed to claim that it was indeed this (and only this) independent variable that affected this (and only this) dependent variable.

4. Examples of Pitfalls in the Learning Analytics Literature

In the following subsections, we will review selected examples from the published LA literature to point out how the three basic pitfalls may manifest in actual research studies. We present one study where confounding may be present, one study with a potential endogenous selection bias, and two studies where we suspect a type of collider bias. Importantly, the DAGs constructed and the causal inference implications derived are tentative, in that they represent a good-faith effort of extrapolating a causal structure from the information provided in the publication, both explicit and implicit, alongside some basic domain knowledge. Given that some assumptions needed to be made, we want to clearly point out that actual data and statistical analyses in the future may prove our conclusions wrong. The goal of this section is not to generate novel scientific insights in these different lines of research but instead to show how reasoning about bias with DAGs can be used to improve study design and statistical analyses for sounder inference.

4.1. Confounding

The Course Signals study by Arnold & Pistilli (2012) had a big impact on the then-still-young field of LA. In September 2021, almost a decade later, Google Scholar reported 1048 citations to this LAK conference paper. A big part of the success of this research may be attributed to the impressive effects reported. In 2007, Purdue University piloted Course Signals, a student early warning system that allowing instructors to learn be alerted about students who might be at risk of failing classes and to act appropriately to inform students of their at-risk status. As a typical example of applied research, the systems effectiveness was evaluated observationally by exploiting the fact that not all courses employed the technology. Due to this, retention of students encountering Course Signals was compared with that of students who did not encounter the system. Notably, the retention rate was more than 20% higher for 4th-year students who had encountered Course Signals in two or more of their classes (93.24%) than those who had not encountered the system at all (69.4%). Later cohorts showed similarly impressive numbers. Moreover, the notable jumps in retention rates of students who had encountered the system in two or more classes led the authors to hypothesize about this significant threshold above which changes “for the rest of their academic careers” may occur (Mathewson, 2015).

However, as skeptical researchers soon began pointing out, there are reasons to believe that the effect may be confounded (Caulfield, 2013; Clow, 2013; Ferguson & Clow, 2017). To reason about the identifiability of the causal effect of interest, we can again refer to available information from "outside" the data. Because we know that the treatment was not randomized, we need to consider variables that point toward the outcome variable. A plausible contender for this is the number of classes taken. The more classes a student has taken, the less likely it becomes that they have not encountered the treatment. Confounding would be present under the additional assumption that this variable — number of classes taken — also affects the outcome of interest, student retention. Of course, this assumption must be true because a student who has dropped out will always have taken fewer classes than a student who has not. The resulting DAG shown in Figure 4 leads us to the conclusion that the effect of Course Signals → Student retention is unidentified due to the open back-door path Course Signals ← Number of Classes → Student Retention. Blocking this back-door path by controlling for the number of classes taken, Course Signals ← Number of Classes → Student Retention would stop the transmission of non-causal association, leading to the causal effect being identified. In this case, the remaining association would more closely approximate the true causal effect of interest. As this confounding variable is easily measured, causal inferences regarding the effects of Course Signals, presumably, could have been greatly improved, even without dramatic changes to the research design.

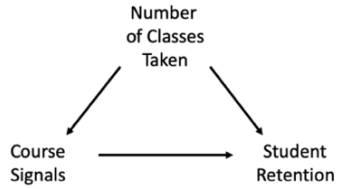

Figure 4. DAG displaying the confounding situation of Arnold & Pistilli (2012), making the causal effect unidentified without further control.

In general, confounding is likely the most prevalent bias in any empirical undertaking, also in LA. Outside of tightly controlled randomized variables, fully squashing confounding is difficult as there are conceivably always additional "lurking" variables that are not amenable to control. Thus, a reduction of confounding should be the goal here. Reducing confounding is necessary to reduce false positives and the identification of confounding variables can only be done with substantive domain knowledge.

4.2. Blocking a Mediator

The following is a somewhat more complex example, in which one incidental control may have led to overcontrol bias and the attenuation of effects, while the deliberate statistical control may have prevented a collider bias. Beheshitha et al. (2016) investigated the effects of three different visualization types on student participation in online discussions, controlling for achievement goal orientations. Notably, they were able to randomize visualization type provisions so that we can be sure no path is pointing to the independent variable, which suggests no confounding. As the authors further expect differential effects due to individual dispositions, they control for achievement goal orientations (AGO). These causal assumptions are encoded in a DAG in Figure 5. As there is no back-door path connecting the independent variable with the dependent variable, we would assume that the causal effect is identified. This type of control is also called "neutral control" (Cinelli et al., 2020), as it does not block a back-door path while also not opening any previously blocked back-door paths. Thus, in terms of the identification and estimation of the causal effect, it is neutral.

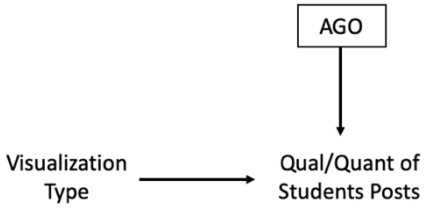

Figure 5. Possible DAG for RQ1 and RQ2 of Beheshitha et al. (2016). AGO stands for achievement goal orientation variables and the box surrounding it indicates (statistical) control. The authors also test visualization type X AGO interactions so that the dependent variable would strictly speaking have another arrow incoming from this interaction term, but we left this out for simplicity's sake.

However, the authors then explain that they restricted their sample to those students who have actually engaged with the visualization more than once. Intuitively, this makes sense because why should students be considered who did not experience the treatment? What is not made explicit though is that the authors, in doing so, introduce a mediator — visualization use — that transmits the association between the intervention and the dependent variable (Figure 6). This mediating configuration is logically necessary because if the actual use of visualizations is necessary to experience the effects of the visualization type, it must be on a path from independent to dependent variable. Also, we must assume that the type of visualization presented will, to some extent, affect the likelihood that a given student will use it. To be clear, this mediating mechanism was, of course, part of the data-generation mechanism before the authors mention it in their report. This is true for any number of potential mediating variables between our main variables of interest. However, if the variable is conditioned on, explicitly or implicitly — the latter as is the case here — we must consider its position within the causal graph.

In essence, by restricting their sample to students who have used the visualizations, they block the mediating path *Visualization Type* → *Visualization Use* → *Qual/Quant of Student Posts,* leading to overcontrol bias (Figure 6). We can presume that this led estimates to be smaller than the true causal effect. If we further assume that AGO affects both visualization use and the dependent variable — both theoretically highly plausible — then we have the further complicating situation that visualization use is a collider, potentially opening up the non-causal path *Visualization Type* → *Visualization Use* ← AGO → *Qual/Quant of Student Posts,* because conditioning on a collider always opens otherwise naturally blocked paths. Fortunately, the authors already control for AGO, so that this non-causal back-door path actually remains blocked, *Visualization Type* → *Visualization Use* ← AGO → *Qual/Quant of Student Posts.* In this case, the arguably more severe collider bias — frequently a Type I error — is avoided, whereas the Type II-leaning overcontrol bias remains present. Having the DAG at hand, we can now reason that simply statistically controlling for AGO would have sufficed to avoid both pitfalls.

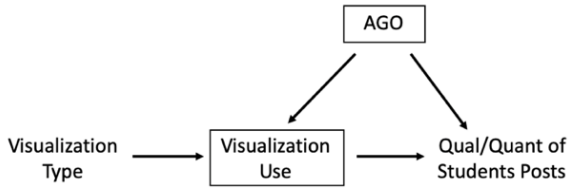

Figure 6. DAG representing possible causal structure of Beheshitha et al. (2016). In this situation, visualization use is a collider.

Mediating variables are the mechanisms by which causal effects are transmitted. Thus, limiting its variance in any way reduces this transmission and, as a result, attenuates the estimate of the true causal effect. In other words, blocking the mediator reduces the association of interest (Schistermann et al. 2009). Researchers should take care not to engage in such overcontrol as this leads to false negatives (Elwert & Winship, 2014). Identification of what is a mediator and what is a confounder necessitates a theoretical understanding of the variables themselves as well as their interrelations.

4.3. Conditioning on a Collider

This section will provide examples of two different types of collider bias that recur in the published LA literature: collider bias by sample truncation and collider bias by nonresponse.

4.3.1. Sample Truncation

Gašević et al. (2016) conducted a large analysis (N=4134) of how student learning behaviours in the LMS predict their grades and their likelihood of passing/failing the course. Building an argument against generalized predictive models, they demonstrate that such predictions of student success vary substantially between different courses. Therefore, they argue, predictions of student success that fail to consider effects of differential instructional conditions place limitations on the potential of LA to improve teaching and learning. To build their case, they sample student data from nine courses, ranging from accounting to graphic design to biology. These courses were chosen due to their participation in an initiative designed to enhance student retention (ESAP). Crucially, eligibility to the ESAP initiative was based on a cutoff of student success in the previous five years (less than 80% or 85% retention, depending on the course). Thus, we can infer that the final sample was truncated not at random but based on relatively stable factors at the course level that lead to consistent underperformance.

As in previous examples, to assess the possibility of bias, we must, to some extent, theoretically reason about these factors; if and how they relate to the hypothesized causal variable and dependent variables of interest. In this case, we expect there to be some relatively stable course factors that may lead to eligibility to the retention initiative. These could be, for example, complexity of course content, difficulty of assessment tasks, alignment in assessment, quality of instructional design, ability

of instructor, or others. Importantly, although the reported analysis included several control variables in the regression models, they are all individual variables like age, gender, previous enrollment, and more. Thus, no course-level variables are accounted for so that they remain completely unobserved. Collider bias looms if we can make the additional assumption that some of these unobserved factors influence not only eligibility to ESAP but also have a causal effect on student behaviour in the LMS (hypothesized causal variable) and student success (dependent variable). Upon consideration, both assumptions seem not only plausible but logically necessary. Figure 7 depicts such a possible configuration.

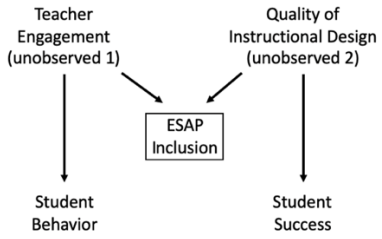

Figure 7. DAG representing possible causal structure of Gašević et al. (2016). Unobserved variables are hypothesized factors determining ESAP inclusion as well as current student behaviour and student success, respectively. For purposes of this example, potential additional, crossing arrows from the unobserved to the observed variables are excluded.

The resulting causal configuration yields an M-bias, a specific type of conditioning on a collider named after its distinct shape (Pearl, 2009; Flanders & Ye, 2019). In this DAG, we have an open back-door path *student behaviour* ← *teacher engagement* → *ESAP inclusion* ← *quality of instructional design* → *student success*. This would introduce spurious associations between student behaviour and student success and could help explain some of the difficult-to-interpret negative associations between student behaviour and student success, e.g., negative association between accessing the LMS feature "book" and student percent marks for some courses (Table 5, p. 77). Paradoxically, this negative association is consistent with the respective behaviour generally being a strong positive predictor in a non-truncated sample.

There are different approaches to circumvent collider bias by sample truncation. As it is likely too difficult to explicitly model all the causal factors for ESAP inclusion in order to control for them — the nuanced effects of teacher engagement and instructional design on student behaviour and student success being the focus of ongoing research efforts — the remaining approaches should be based on ceasing to condition on the collider ESAP inclusion. Therefore, this is an example of bias that likely cannot be remedied by statistical control with the tools laid out in the current paper. Instead, experimental control seems most appropriate. Thus, the effects of student behaviour on student success should be estimated based on a sample that is not truncated based on the unobserved antecedents of ESAP inclusion. An extension of the initiative to all courses or a randomized subset of these (i.e., not only underperforming ones) would remove the collider bias and allow for the estimation of the true causal effect of student behaviour on student success.

Moving away from this specific example to the more general issue of sample truncation, an important goal of LA (Verbert et al., 2012) is identifying and helping at-risk student populations (e.g., Akçapınar et al., 2019; Russell et al., 2020; Foster & Siddle, 2020). From an external validity perspective, sampling such a subset of students can bring the limitation that inferences will not be applicable to the larger, non-at-risk student populations — a limitation that may be deemed acceptable, depending on the underlying goal of learning analytics interventions. However, as outlined above and perhaps counterintuitively, sample truncation may also lead to inferences that themselves are biased, thus lacking not only external but also internal validity (Munafò et al. 2018). This is because the causes for the at-risk status of students are themselves causal factors for the variables of interest and a sample truncated based on these causes must be, by definition, biased with respect to the variables of interest. In these cases, the obtained parameter estimates will not reflect the true causal effect, which in the obvious cases can lead to results that are hard to interpret, for example, due to sign changes. In the less obvious cases, results may simply be attenuated or inflated.

4.3.2. Nonresponse

Zhu et al. (2016) conducted a longitudinal analysis in a MOOC on Big Data and Education. Specifically, they looked at network metrics derived from forum activity and looked for associations between the social connectedness of a given student and learning engagement, as measured by acquired scores, lecture views, and other extracted behaviours. The longitudinal aspect becomes salient as they conduct these analyses several times (once per week for an 8-week course) to account for 1) developments over time and 2) network metrics from a previous week influencing learning engagement later. At the time of data collection, this MOOC has had more than 48,000 participants over its duration, with only a small subset of students actively participating and completing, as is common in MOOCs (Reich & Ruiperez-Valente, 2019). Thus, in line with their

research questions, the authors restricted their sample to students who had posted or commented at least once in the forum, reducing the sample to 770 students. Further, we learn that of this subset, only 155 (or 20%) of students earned a certificate. The longitudinal design of this study thus implies that the sample progressively shrunk from week 1 to week 7 (week 8 was not analyzed). This means that each weekly regression model is conducted on an increasingly limited sample. Our substantive knowledge of the MOOC literature suggests that this attrition process is not random but may be determined by a multitude of factors, two of which may be the variables of interest themselves or closely related, i.e., social connectedness (e.g., Galikyan et al., 2021; Wang et al., 2019) and learning engagement (e.g., Tseng et al., 2016; Joksimović et al., 2018). Or, conversely, given our substantive knowledge, it appears highly unlikely that MOOC attrition is entirely independent from these student variables. These resulting causal assumptions are encoded in the DAG in Figure 8 and suggest that the authors have conditioned on the collider MOOC attrition. Because colliders become unblocked if they are conditioned on, there is now an open back-door path between the causes of attrition, introducing spurious associations.

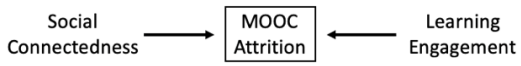

Figure 8. DAG representing possible collider bias in Zhu et al. (2016).

Through consulting information from outside the data, we can hypothesize that the authors have conditioned on MOOC attrition by 1) initially restricting their sample to students who have participated in the forum and 2) by conducting analyses on a progressively shrinking subset of these initial students (which we find additional evidence for in the increasing sizes of standard errors). This attribute of the research design may help explain the "sometimes puzzling effects" (Zhu et al., 2016, p. 1) discovered in the data. The authors note that, for example, there were even sign changes of some effects between weeks. Instead of substantive reasoning, these inconsistent findings may be consistent with an increasingly severe collider bias due to the longitudinal design of the study. Unfortunately, in this case, there seems to be no easy fix as the collider is conditioned on by design. Due to the research questions formulated, it is not a reasonable solution to extend the analysis to the total sample of MOOC participants. One possibility for mitigating the inconsistency of bias in this case would be restricting the sample to the subset of students who completed the course from the beginning. However, this would not affect the severity of collider bias but only yield more consistently biased estimates.

Extending this to the broader literature, as a high degree of student attrition is a known feature of MOOCs (Reich & Ruiperez-Valente, 2019), LA research using MOOC data should be aware of the effects of non-response on causal inferences. Again, from an external validity perspective, it can be problematic to draw inferences based on those students who persevered, as this amounts to a survivorship bias, the fallacy of deriving broad conclusions based on a sample that has overcome a critical selection process. What is less recognized is that from the perspective of internally valid causal inferences, a failure to account for the possibility of student attrition being a collider increases the likelihood of biased parameter estimates (Munafò et al., 2018), which can lead to inconsistent or puzzling results. This is because the causes of student attrition are themselves the variables of interest and, thus, samples selected in this way are *per se* biased with respect to these focal variables.

5. Discussion and Conclusion

Educational research is a complex undertaking (Berliner, 2002), not least due to the many potential sources of bias that undermine attempts of arriving at answers to causal questions from observational research (Gustafsson, 2013). By association, this is also true of the field of LA, where the same issues prevail, albeit embedded in the complexity of socio-technical systems (Dawson et al., 2019). As a result, outside of randomized controlled experiments, it is unlikely that bias can be circumvented entirely. In these cases, a principled approach to reducing bias to approximate true causal effects is needed.

Using the graphical methods of inspecting DAGs to assess the identifiability of causal effects, this paper has focused on the back-door criterion, where non-causal paths are discovered and selected for potential adjustment. With this, we have outlined a non-parametric and straightforward tool to reason about bias in the LA literature. However, the back-door criterion does not exhaust the approaches of estimating causal effects from observational data. For example, in the unfortunate situation where unobserved or even unobservable variables are expected to introduce bias, we can turn to another approach, the *front-door criterion* (Pearl, 1995; 2009). This approach works if an observed variable fully mediates between the hypothesized causal variable and dependent variables but is unrelated to the unobserved confounder. Then, the causal effect can be identified, even in the presence of such an unobserved confounder (Bellemare et al., 2019). However, this approach is not yet incorporated widely into the toolkit of empirical researchers, mainly due to the rarity of situations where these assumptions hold. For this reason, this primer focuses on the much more widely used back-door criterion, which can be applied to a variety of research situations in the social sciences, such as education and psychology, and to LA in particular.

A hallmark of this approach is the reliance on expert or "outside" knowledge, i.e., theory, to arrive at DAGs that are as valid and comprehensive as possible. Although the importance of theory has been noted (e.g., Dawson et al., 2015; Wise & Shaffer, 2015; Wong et al., 2019), for example, to guide the choice of variables, to help interpret findings, or to reason about generalizability, the role of substantive theory to avoid pitfalls in causal inference has received little explicit attention in the LA literature. This translates into an important implication of this paper: Substantive knowledge — about how constructs and variables interact — is the most valuable lever for causal claims (Shrier & Platt, 2008) and, in essence, the foundation for discovery of actionable insights to understand and optimize learning, the stated goal of LA (Clow, 2013; Siemens & Gašević, 2012).

Unfortunately, resolving this theoretical impetus may quickly become challenging in practice, as substantive knowledge (theory) itself consists of assumed causal structures. However, as demonstrated in this primer, with the DAG approach to causal inference and only skeletal, outside information at hand, we can plausibly reduce bias in observational research. In some cases, there may be competing, equally convincing DAGs with diverging causal implications. In these cases, points of divergence should themselves become the focus of future research endeavours, such that — akin to a bootstrapping procedure — a causal foundation is built based on existing knowledge. Crucially, a DAG derived based on substantive and rational deliberation, no matter how uncertain or preliminary, will likely yield better estimates than remaining in the darkness of spurious associations. In this, we wholeheartedly agree with Hicks et al. (2022) that causal reasoning with DAGs provides a valuable non-technical tool to incorporate knowledge from different sources — for example non-research stakeholders or researchers from different disciplines — to arrive at actionable insights for substantive questions. They demonstrate this for the challenge of student retention but, of course, other research lines may also profit similarly.

As a specific implication of this paper, we suggest that future research studies should make use of DAGs to reason about the strength of causal claims and the possibility of bias. Ideally, this should be done in the planning phase so that potential pitfalls may be mitigated ahead of time, for example by collecting additional data needed for statistical control or improving the research design toward tighter experimental control, or both. It is hoped that this relatively simple non-parametric approach to causal reasoning — back-door criterion using DAGs — becomes more widespread in the field of LA, as has been the case in many other fields, with epidemiology being a striking example of speedy integration of this approach (Tennant et al., 2021). Transparent communication of causal assumptions may not only yield better estimates of causal effects in single studies but may overall have the added benefit of making more salient the status of causal knowledge in the field, highlighting gaps and open questions regarding actionable knowledge in LA. It is our hope that this introduction helps kickstart and extend the conversation about causal claims, potential pitfalls, and the use of DAGs in LA research.

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

The authors declared no financial support for the research, authorship, and/or publication of this article.

Acknowledgements

The first author would like to thank Julia M. Rohrer for her generous support and guidance in constructing plausible DAGs and discussing potential sources of bias in the early stages of this manuscript. Further, we extend our thanks to the three reviewers who helped us improve this manuscript over two review rounds with their critical and valuable comments.

References

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327–339. https://doi.org/10.1080/07388940500339167

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*(40). https://doi.org/10.1186/s41239-019-0172-z

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 267–270). ACM Press. https://doi.org/10.1145/2330601.2330666

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Baker, R. S., Gašević, D., & Karumbaiah, S. (2021). Four paradigms in learning analytics: Why paradigm convergence matters. *Computers and Education: Artificial Intelligence*, *2*, 100021. https://doi.org/10.1016/j.caeai.2021.100021

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2020). On Pearl's hierarchy and the foundations of causal inference. Technical Report R-60, Causal AI Lab, Columbia University. https://causalai.net/r60.pdf

Beheshitha, S. S., Hatala, M., Gašević, D., & Joksimović, S. (2016). The role of achievement goal orientations when studying effect of learning analytics visualizations. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 54–63). ACM Press. https://doi.org/10.1145/2883851.2883904

Bellemare, M. F., Bloem, J. R., & Wexler, N. (2019). *The paper of how: Estimating treatment effects using the front-door criterion*. Working Paper.

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18–20. https://www.jstor.org/stable/3594389

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82. https://doi.org/10.1080/19345747.2011.578707

Caulfield, M. (2013, Sept 26). *Why the Course Signals math does not add up.* Hapgood. https://hapgood.us/2013/09/26/why-the-course-signals-math-does-not-add-up/

Cinelli, C., Forney, A., & Pearl, J. (2020, October 29). A crash course in good and bad controls. http://dx.doi.org/10.2139/ssrn.3689437

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

Cooper, A. (2012). What is analytics? Definition and essential characteristics. *CETIS Analytics Series*, *1*(5), 1–10. http://publications.cetis.org.uk/wp-content/uploads/2012/11/What-is-Analytics-Vol1-No-5.pdf

Dawson, S., Mirriahi, N., & Gašević, D. (2015). Importance of theory in learning analytics in formal and workplace settings. *Journal of Learning Analytics*, *2*(2), 1–4. https://doi.org/10.18608/jla.2015.22.1

Dawson, S., Joksimović, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (LAK '19), 4–8 March 2019, Tempe, AZ, USA (pp. 446–455). ACM Press. https://doi.org/10.1145/3303772.3303784

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 56–65). ACM Press. https://doi.org/10.1145/3027385.3027396

Flanders, W. D., & Ye, D. (2019). Limits for the magnitude of M-bias and certain other types of structural selection bias. *Epidemiology*, *30*(4), 501–508. https://doi.org/10.1097/EDE.0000000000001031

Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, *45*(6), 842–854. https://doi.org/10.1080/02602938.2019.1682118

Galikyan, I., Admiraal, W., & Kester, L. (2021). MOOC discussion forums: The interplay of the cognitive and the social. *Computers & Education*, *165*, 104133. https://doi.org/10.1016/j.compedu.2021.104133

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*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, *24*(3), 275–295. https://doi.org/10.1080/09243453.2013.806334

Hellings, J., & Haelermans, C. (2020). The effect of providing learning analytics on student behaviour and performance in programming: A randomised controlled experiment. *Higher Education*, *83,* 1–18. https://doi.org/10.1007/s10734-020-00560-z

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, *32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health, 108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC.

Hicks, B., Kitto, K., Payne, L., & Buckingham Shum, S. (2022). Thinking with causal models: A visual formalism for collaboratively crafting assumptions. *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK '22), 21–25 March 2022, Online (pp. 250–259). ACM Press. https://doi.org/10.1145/3506860.3506899

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, *86*(1), 226–244. https://doi.org/10.1162/003465304323023778

Jivet, I., Scheffel, M., Schmitz, M., Robbers, S., Specht, M., & Drachsler, H. (2020). From students with love: An empirical study on learner goals, self-regulated learning and sense-making of learning analytics in higher education. *The Internet and Higher Education, 47*, 100758. https://doi.org/10.1016/j.iheduc.2020.100758

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C., & Brooks, C. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research*, *88*(1), 43–86. https://doi.org/10.3102/0034654317740335

Jørnø, R. L., & Gynther, K. (2018). What constitutes an "actionable insight" in learning analytics? *Journal of Learning Analytics*, *5*(3), 198–221. https://doi.org/10.18608/jla.2018.53.13

Kahlert, J., Gribsholt, S. B., Gammelager, H., Dekkers, O. M., & Luta, G. (2017). Control of confounding in the analysis phase: An overview for clinicians. *Clinical Epidemiology*, *9*, 195–204. https://doi.org/10.2147/CLEP.S129886

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. Unpublished manuscript, Institute for Quantitative Social Science, Harvard University, Cambridge, MA.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis 27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Klenke, J., Massing, T., Reckmann, N., Langerbein, J., Otto, B., Goedicke, M., & Hanck, C. (2021). Effects of early warning emails on student performance. https://doi.org/10.48550/arXiv.2102.0880

Knight, C. R., & Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In *Handbook of Causal Analysis for Social Research* (pp. 275–299). Springer. https://doi.org/10.1007/978-94-007-6094-3_14

Lee, J. J. (2012). Correlation and causation in the study of personality. *European Journal of Personality, 26*(4), 372–390. https://doi.org/10.1002/per.186

Lim, L.-A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., & Dawson, S. (2021). What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction, 72,* 101202. https://doi.org/10.1016/j.learninstruc.2019.04.003

Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education*, *29*, 40–48. https://doi.org/10.1016/j.iheduc.2015.12.003

Mathewson, T. G. (2015, August 21). Analytics programs show 'remarkable' results — and it's only the beginning. Higher Ed Dive. https://www.highereddive.com/news/analytics-programs-show-remarkable-results-and-its-only-the-beginning/404266/

McNamee, R. (2005). Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*, *62*(7), 500–506. http://dx.doi.org/10.1136/oem.2002.001115

Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Pérez-Sanagustín, M., Alario-Hoyos, C., & Kloos, C. D. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. *Computers & Education*, *145*, 103728. https://doi.org/10.1016/j.compedu.2019.103728

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Meehl, P. E. (1970). Nuisance variables and the *ex post facto* design. University of Minnesota Press. https://hdl.handle.net/11299/184638

Motz, B. A., Carvalho, P. F., de Leeuw, J. R., & Goldstone, R. L. (2018). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, *5*(2), 47–59. https://doi.org/10.18608/jla.2018.52.4

Mousavi, A., Schmidt, M., Squires, V., & Wilson, K. (2021). Assessing the effectiveness of student advice recommender agent (SARA): The case of automated personalized feedback. *International Journal of Artificial Intelligence in Education*, *31*, 603–621. https://doi.org/10.1007/s40593-020-00210-6

Mullaney, T., & Reich, J. (2015). Staggered versus all-at-once content release in massive open online courses: Evaluating a natural experiment. *Proceedings of the 2nd ACM Conference on Learning @ Scale* (L@S 2015)*,* 14–18 March 2015, Vancouver, BC, Canada (pp. 185–194). ACM Press. https://doi.org/10.1145/2724660.2724663

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Smith, G. D. (2018). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, *47*(1), 226–235. https://doi.org/10.1093/ije/dyx206

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science, 8*(3), 266–269. https://www.jstor.org/stable/2245965

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.2307/2337329

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

Pearl, J. (2021). Causal and counterfactual inference. In M. Knauff & W. Spohn (Eds.), *The Handbook of Rationality* (pp. 427–438). The MIT Press.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*, 369–375. https://doi.org/10.1038/s42256-020-0197-y

Rao, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician, 25*(5), 37–39. https://doi.org/10.2307/2686082

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science*, *363*(6423), 130–131. https://doi.org/10.1126/science.aav7958

Richardson, T. G., Smith, G. D., & Munafò, M. R. (2019). Conditioning on a collider may induce spurious associations: Do the results of Gale et al. (2017) support a health-protective effect of neuroticism in population subgroups? *Psychological Science*, *30*(4), 629–632. https://doi.org/10.1177/0956797618774532

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, *12*(3), 313–320. https://doi.org/10.1097/00001648-200105000-00011

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/25152459177456

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.2307/2335942

Russell, J.-E., Smith, A., & Larsen, R. (2020). Elements of success: Supporting at-risk student resilience through learning analytics. *Computers & Education*, *152*, 103890. https://doi.org/10.1016/j.compedu.2020.103890

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, *20*(4), 488–495. https://doi.org/10.1097/EDE.0b013e3181a819a1

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*. https://doi.org/10.1186/1471-2288-8-70

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: Review and recommendations. *International Journal of Epidemiology, 50*(2), 620–632. https://doi.org/10.1093/ije/dyaa213

Sønderlund, A. L., Hughes, E., & Smith, J. (2018). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Siemens, G., & Gašević, D. (2012). Guest editorial: Learning and knowledge analytics. *Educational Technology & Society*, *15*(3), 1–2.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2010). Stratification for confounding — Part 1: The Mantel-Haenszel formula. *Nephron Clinical Practice*, *116*(4), c317–c321. https://doi.org/10.1159/000319590

Tseng, S.-F., Tsao, Y.-W., Yu, L. C., Chan, C.-L., & Lai, K. R. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning*, *11*. https://doi.org/10.1186/s41039-016-0033-5

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society, 15*(3), 133–148. https://www.jstor.org/stable/jeductechsoci.15.3.133

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

Wang, W., Guo, L., He, L., & Wu, Y. J. (2019). Effects of social-interactive engagement on the dropout ratio in online learning: Insights from MOOC. *Behaviour & Information Technology*, *38*(6), 621–636. https://doi.org/10.1080/0144929X.2018.1549595

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Westreich, D., & Greenland, S. (2013). The Table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. https://doi.org/10.1093/aje/kws412

Winne, P. H. (1982). Minimizing the black box problem to enhance the validity of theories about instructional effects. *Instructional Science, 11*, 13–28. https://doi.org/10.1007/BF00120978

Winne, P. H. (1983). Distortions of construct validity in multiple regression analysis. *Canadian Journal of Behavioural Science*, *15*(3), 187–202. https://doi.org/10.1037/h0080736

Winne, P. (2017). Leveraging big data to help each learner and accelerate learning science. *Teachers College Record*, *119*(3), 1–24. https://doi.org/10.1177/016146811711900305

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior, 112*, 106457. https://doi.org/10.1016/j.chb.2020.106457

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. https://doi.org/10.18608/jla.2015.22.2

Wong, J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., Houben, G.-J., & Paas, F. (2019). Educational theories and learning analytics: From data to knowledge. In D. Ifenthaler, D.-K. Mah, & J. Yin-Kim Yau (Eds.), *Utilizing learning analytics to support study success* (pp. 3–25). Springer. https://doi.org/10.1007/978-3-319-64792-0_1

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction, 22*(6), 413–419. https://doi.org/10.1016/j.learninstruc.2012.03.004

Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., & Paquette, L. (2016). Longitudinal engagement, performance, and social connectivity: A MOOC case study using exponential random graph models. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 223–230). ACM Press. https://doi.org/10.1145/2883851.2883934

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17. https://doi.org/10.1207/s15326985ep2501_2