Participatory Causal Modelling for Learning Analytics

Stage 1 Report

Ben Hicks

Supervised by Kirsty Kitto and Simon Buckingham Shum

October 2022

Abstract

Learning Analytics is not just about using data to understand learning and the environments in which it occurs. It also aims to improve the learning process into the future. This necessitates a causal understanding of learning processes; if I *do* this, how will it change the outcome? To develop a causal understanding from learning data we have to understand the context in which the learning occurs, and apply this to a rigorous statistical framework. One way in which to do this involves the use of causal inference frameworks. However, making use of this relatively new modelling technique requires that data experts and context experts need to form a shared understanding about the system they are analysing. This is a difficult epistemic gap to bridge, as developing a shared understanding between technical and non-technical stakeholders is challenging, as is translating contextual knowledge into an appropriate statistical apparatus. So despite the importance of making causal claims in order to improve learning a traditional statistical approach is generally adopted in Learning Analytics. This explicitly rules out causal claims unless they are accompanied by a randomised controlled trial (RCT), but RCTs are particularly difficult to run in education due to both practical and ethical concerns. Graphical Causal Models offer something new; a collaborative way to construct theories about the causal mechanisms of the world, that carries a direct interpretation for making causal claims from statistical associations. The visual formalism of the model requires little technical knowledge to engage with, whilst informing the statistical modeller about how to make stronger causal claims. In my thesis I will apply these models to two cases; a student retention model and a conceptualisation of student belonging. The models will be co-constructed between stakeholders with a wide range of expertise in the modelling process, using the visual formalism to help foster a shared understanding of the system. This process of participatory modelling has the potential to allow a wide range of experts to be actively involved in the co-creation of models that support making causal claims. This stage 1 document presents my progress to date, and my plans for the remainder of my candidature.

Contents

List of Figures

0 Published work

I have published several papers over the course of my candidature so far. The most directly relevant to this work was the following, which I presented at the Learning Analytics and Knowledge conference in 2022:

> [63] Ben Hicks, Kirsty Kitto, Leonie Payne, and Simon Buckingham Shum. Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 250–259, 2022

I have published several other works, not directly related to causal models, but all related to understanding the impact of various interventions in learning:

> Kelly Linden, Ben Hicks, and Sarah Teakel. Weathering the storm: Targeted and timely analytics to support disengaged students. *EDUCAUSE Review*, 2022
>
> Ben Hicks, Kelly Linden, and Neil Van Der Ploeg. Opportunities to improve learning analytics for student support when using online assessment tools. In *Back to the Future–ASCILITE '21*, pages 60–64. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021
>
> Robyn Brunton, Jasmine MacDonald, Nicole Sugden, and Ben Hicks. Discussion forums: A misnomer? examining lurkers, engagement and academic achievement. *Australasian Journal of Educational Technology*, pages 37–54, 2022
>
> Nicole Sugden, Robyn Brunton, Jasmine MacDonald, Michelle Yeo, and Ben Hicks. Evaluating student engagement and deep learning in interactive online psychology learning activities. *Australasian Journal of Educational Technology*, 37(2):45–65, 2021
>
> Nicole Sugden, Robyn Brunton, Jasmine McDonald, Ben Hicks, Cassandra Colvin, and Michelle Yeo. Delving deeper into student engagement: Conversations between academics and analysts. In *Australian College of Applied Psychology Conference: ACAP 2019*, 2019

Neil van der Ploeg, Kelly Linden, and Ben Hicks. Job ready graduates package: The difficulty in identifying 'genuine students'. In *HERDSA 2022*. HERDSA, 2022

Kelly Linden, Neil Van Der Ploeg, Ben Hicks, and Kirsten Locke. Increasing student grades in large online subjects: Combining tutorial support with technology. In *Back to the Future–ASCILITE '21*, pages 208–212. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021

Kelly Linden, Neil Van Der Ploeg, and Ben Hicks. Ghostbusters: Using learning analytics and early assessment design to identify and support ghost students. In *Back to the Future–ASCILITE '21*, pages 54–59. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021

Kelly Linden, Neil van der Ploeg, Ben Hicks, and Prue Gonzalez. Peering into the crystal ball of the disengaged: What happens to students that do not submit an early assessment item? In *37th International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education: ASCILITE 2020*, pages 48–53. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2020

Neil van der Ploeg, Kelly Linden, Ben Hicks, and Prue Gonzales. Widening the net to reduce the debt: Reducing student debt by increasing identification of completely disengaged students. In *Proc. ASCILITE 2020 in Armidale*, ASCILITE's First Virtual Conference, pages 54–59. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2020

1 Introduction

1.1 Making well informed interventions

Education, at its heart, is about improvement. Improving our society, our productivity, our culture, our knowledge, our understanding, and our thinking [67, 173]. Educational institutions want to know how to make the most effective interventions in order to improve learning outcomes [113, 173]. As a result education research also has a clear focus upon this problem, with 4,290 results in 2021 alone in response to the search query "improving student learning" in Google Scholar.

According to the systems theorist Midgley [118] an intervention in some system requires some *agent* to take purposeful *action* to create *change*, constrained by some agreed boundaries of the system. The judgements required for an agent to take action in this way require strong evidence, particularly in the high-stakes arena of education where an error can result in substantially different life outcomes for our students [64, 67]. At the scale of some individual teacher in a classroom the agent may be collating this evidence and making judgements according to their own internal mental models of the system. However, as learning increasingly moves online and to digital environments data and Learning Analytics (LA) are being used to provide *actionable insights* [74] that inform interventions and improve learning at a range of scales [43].

Successfully intervening in a system as complex as education is no easy task [25]. Two kinds of problem arise depending on whether we are informing future decisions or evaluating past actions: (i) gathering sufficient evidence to take action and intervene; and (ii) assessing the evidence of that intervention. Approaching either of these problems requires a deeper knowledge of the system than what is provided by the data alone [130]. Firstly, to take *action* on some insight gleaned from the data LA needs to make *causal claims* [121]. It is not enough to understand what is happening, one must also model how it can be changed, or how events would have changed under different circumstances. But to understand the effect that our actions may have, before taking them and seeing the result, requires a model of the world that describes causal mechanisms — we have to predict the consequences of our actions [64, 130]. Second, to measure the effect of an intervention we have taken we must model the hypothetical consequences of having *not* made the intervention (the counterfactual), and then compare this what actually happened. This also requires a model of the causal mechanisms within the system under study [130].

But LA lacks causal models and is reluctant to make causal claims (Section 2.2). This is in part due to a general historical aversion to making causal claims from data, across any field of study [130]. But the reluctance is also in part due to the challenge of modelling the complex systems that education generates [25]. A complex system is difficult to model as there can be multiple valid was to represent them [144]. Representing a student by any set of metrics has always been contentious and will likely always be, but representing a planet, as astrophysicists do, as a single point is uncontroversial. The key difference is the underlying complexity of the causal relationships of these systems at the scale we care about. It makes modelling how the causal mechanisms work in an educational environment a challenging task, and the models that do arise can be difficult for educators to understand, and reason with.

However some claim that modelling causation in the abstract and how we reason causally are two different creatures. Hitchcock [65] distinguishes two versions of causation; scientific and folk attributive. By *scientific* causation they mean the causation that can be represented by an abstract *causal model*. It is 'scientific' in the sense that it can be used for simulations, predictions and analysis. The *folk attributive* concept of causation is used to describe the causal reasoning of common folk. A similar distinction is made by Hastie [57], who describes the *causation of actual events* (the external world) compared to *how a person thinks about causation* (the cognitive world) [57]. This last distinction is of critical importance because how a person thinks about causation is how they will make decisions pertaining to acting on a potential intervention.

Well informed interventions based on data requires coherence between the causal reasoning of those taking action and the causal reasoning of the LA. My thesis will investigate the possibilities of using diagrams of the causal relationships, called Graphical Causal Models (introduce in Section 2.2.3.7), to make both stronger causal claims and to help promote a shared understanding of causation between the educators and LA developers.

1.2 Overview

This document begins with a review of the literature (Chapter 2) followed by a short section on the research approach (Chapter 3). As much of the research plan and methods is based on the prototyping and exploratory work I have already completed the next chapter covers the preliminary results (Chapter 4), before I elaborate on a more detailed research plan (Chapter 5).

Much of the work centres around the use of Graphical Causal Models (GCMs), often referred to as Causal DAGs. If you are unfamiliar with DAGs used for causal inference and want to know some of their technical details there is a more detailed tutorial at the end of this document (Appendix A) that I wrote for a paper presented at LAK22, *Thinking with Causal Models*, Hicks et al. [63].

The approach I will take to the entire project is depicted in my thesis map in Figure 1.1.

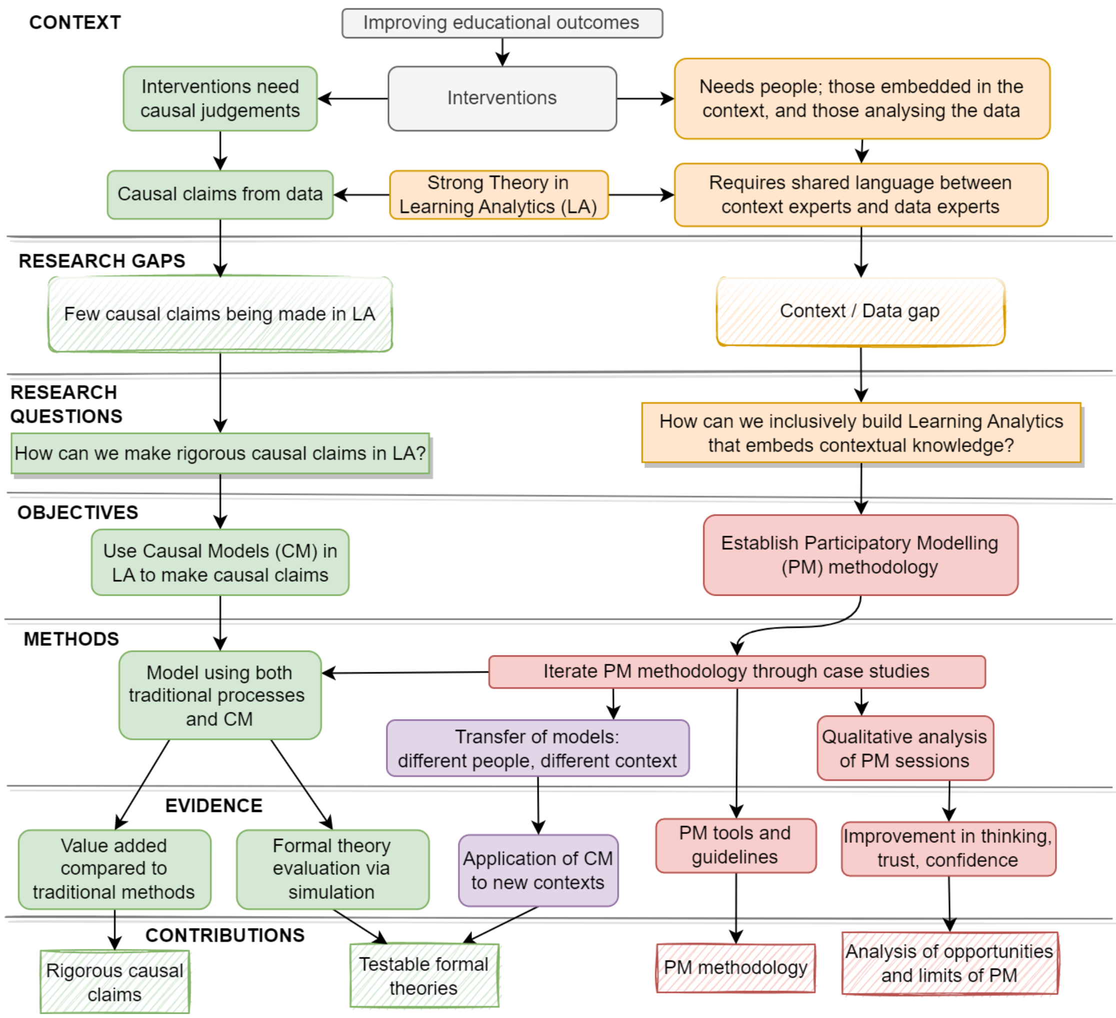


Figure 1.1: Thesis Map describing the proposed dissertation.

Several of the chapters (2, 4, and 5) follow the same structure of working down the left-hand side of the thesis map (in green, Figure 1.1) and then the right-hand side.

2 Literature review

This review of the literature begins with the problem of making well informed interventions in the Learning Analytics context (Section 1.1). I will identify two research gaps for the field: (i) the problem of making causal claims from data in order to take action; and (ii) the problem that those who take action need to understand, and trust, the resultant causal claims.

I explore the general problem of making causal claims from data in Section 2.1 and then specifically in the context of LA in Section 2.2. This leads to identification of the first research gap in the field, which motivates Research Question 1 (Section 2.3).

I then address the problem of understanding the causal claim by the relevant stakeholder, starting with outlining the problem in Section 2.4. Two approaches that LA has explored are examined in attempt to solve this problem: Participatory and Co-Design in Section 2.5; and then developing strong theory in Section 2.6. Both attempt to foster understanding by embedding contextual knowledge in the construction of LA. This leads to identification of the second research gap, motivating Research Question 2 (Section 2.7).

2.1 Making causal claims from data

To understand the challenges associated with using data to make causal claims it is necessary to understand some of the historical context surrounding statistics and causation. For a notion that is so central to how we understand the world causation has proven to be a slippery concept to analyse and as such it is heavily contested in several fields [41, 57]. In the early 20$^{\text{th}}$ century the likes of Karl Pearson, one of the founders of modern statistics, was a strong advocate of causality free statistics [130]. One could make claims about association between variables, that is they tend to move together, but not about causation. The one exception to this arose if a Randomised Control Trial was used to determine that a causal relationship between two variables really did exist (more on this later in Section 2.2.2). Over one hundred years later researchers are still challenged in making causal claims from data, often avoiding the word 'causal' entirely [54, 59]. This historical context, and the reluctance of many fields to make claims outside of a controlled experiment [54, 77, 120, 162], is important to keep in mind as we examine how causal claims have been made in Learning

Analytics (Section 2.2). However, the past few decades have seen a resurgence in interest in causality, and several frameworks have been developed for making causal claims outside of the experimental setting [128, 143]. Those pertinent to LA will discussed in Section 2.2.3, but for now we will examine what is in general required to move from our observations of the world (the data) towards a causal claim.

Pearl and Mackenzie [130] offer the *ladder of causation* as a way to assess different levels of statistical claims made from data, framed around exploring the relationship between two variables, X and Y. They are:

1. Seeing. Do X and Y move together? (**Association**)
2. Doing. What happens to Y when we actively change X? (**Intervention**)
3. Imagining. Does X cause Y? Would Y have changed if we did not do X? (**Counterfactual**)

Association is the lowest level in the ladder of causation, and it is worth pausing a moment to discuss what this means. Once we begin to add data from our observations of the world it is essential to talk about the statistical concept of *association*; if two variables tend to *move* together, then they are *associated*. The closer they follow each other the stronger the association (the term *correlation* is often used in place of association, although it is a specific case referring to the linear association between two continuous variables). Associations in the data can be analysed at different levels, based on the strength of the claim one wants to make; descriptive, predictive, or causal [54]. Descriptive claims amount to *seeing* patterns in the data, predictive claims can be used to make *predictions*, and causal claims can make statements about interventions, counterfactuals and causal mechanisms. The upper two rungs of the ladder of causation, Intervention and Counterfactual, are causal claims. Unfortunately the language around descriptive, predictive or causal claims using data are often used interchangeably in the literature, obscuring the actual claim being made [54]. The difficulty, then, is moving from seeing associations in the data to stronger, causal claims.

While 'association is not causation' (see Hernán and Robins [60, p11] for a thorough example), it is certainly a clue that *something* is going on. Attempting to extract causal claims from associations in data can lead to problems with *confounding variables* [104]. It can also be very difficult to extract information about the *direction* of the causal effect from data alone. I will now discuss these two problems in more depth, using some simple educationally inspired examples.

2.1.1 Causality and the problem of direction

Suppose we have two variables and we are trying to understand the causal connection between them; one is *passing an online quiz* (call this Y) and the other is *doing the required reading* (call this X). Most people would assume that people who do the required reading have a better chance of passing the quiz. This is a causal claim, and can be expressed mathematically [134] as:

$$\mathbb{P}(Y|X) > \mathbb{P}(Y|\neg X) \tag{2.1}$$

Equation 2.1 states that the probability of passing the quiz (Y occurring) is greater when a student does the required reading (X has happened), compared to when the student has not done the required reading (X has not happened). These probabilities can be estimated from the data.

However the estimation of these probabilities from data is symmetric, there is no intrinsic order to the joint probability distribution of the data. It can be shown (see Lemma 2.1.1) that the statement $\mathbb{P}(X|Y) > \mathbb{P}(\neg X|Y)$ is a mathematically equivalent statement; which translates back to the causal claim 'people who pass the test are more likely to have done the reading'.

Lemma 2.1.1. *The conditional probability statement of causality is symmetric.*

Proof. We are given:

$$\mathbb{P}(Y|X) > \mathbb{P}(Y|\neg X)$$

Applying Bayes rule ($\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$) to both sides gives:

$$\frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)} > \frac{\mathbb{P}(\neg X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

And then dividing both sides by $\mathbb{P}(Y)/\mathbb{P}(X)$:

$$\mathbb{P}(X|Y) > \mathbb{P}(\neg X|Y)$$

□

This does not mean that passing the test *causes* you to go back in time and do the required reading. Unfortunately, in the process of translating a causal claim such as 'doing the required reading leads to better quiz performance' to a mathematical equation has lost some important information — the direction of the flow of causation. Causal effects are inherently *asymmetric* whereas the language of mathematics, and by proxy statistics, gravitates towards a *symmetrical* interpretation, particularly where equivalence is involved. Finding two variables that are associated tells us nothing about which variable might be influencing the other. The direction of causation must generally be supplied; the data alone may not say.

Traditionally this has problem has been overcome by experimental design, which is explored in more detail in Section 2.2.2. The suspected cause, the X in Equation 2.1, is controlled by the researcher through randomisation, and the subsequent effect, Y, is then observed.

2.1.2 Causality and the problem of confounding

Next suppose we are interested in the effect that owning a *computer* has upon a student's academic *outcome*. We may find that owning a *computer* and academic *outcome* are associated directly from the data, but this is not necessarily due to a causal effect. There is something lurking in the background; a *confounding variable*. It is likely that coming from a more financially *privileged* background influences both the chance of owning a computer *and* a a student's academic outcome. The confounding variable *Privilege* distorts the measured effect; it is a source of *bias* in estimating the causal effect of owning a *computer* on academic *outcome*. This problem can be represented graphically using Figure 2.1.

$$\begin{tikzcd} & Privilege \arrow[dl] \arrow[dr] & \\ Computer \arrow[rr] & & Outcome \end{tikzcd}$$

Figure 2.1: The confounding variable *Privilege* obscuring the true causal effect of *Computer* on *Outcome*

This problem could have great importance if we were considering an expensive intervention of purchasing laptops for all students.

The traditional way of overcoming this problem is through experimental design, outlined in Section 2.2.2. The suspected cause, in this case *Computer* in Figure 2.1, is manipulated through randomisation. This blocks any influence that the variable *Privilege* has on *Computer*, and so removes any confounding effect that occurs.

2.1.3 Causal inference frameworks

Pearl [128] notes that traditionally causal claims from association are made in verbal interpretations of a statistical result. This is an appeal to reason or scientific thinking perhaps, but not part of the statistical apparatus. Statistical association is a clue that helps uncover evidence that a causal relationship might exist in a dataset, but it is insufficient on its own to make a statement about causation. The resulting association measure may be *biased*, pulled away from its causal interpretation in a systematic way. Something more beyond the data is needed, encapsulated in a *causal inference framework* [60, 127, 167]. Traditionally the Randomised Control Trial framework (Section 2.2.2) has been the dominant methodology but with new modelling approaches with causal inference frameworks it is now possible to move towards causal claims using different mechanisms. These frameworks must overcome these two problems of confounding and directionality in some way, a problem that is discussed in more detail by Hernán and Robins [60], Pearl [127], Pearl and Mackenzie [130]. Next we examine how these problems have been overcome in order to make causal claims in LA.

2.2 Causal claims in Learning Analytics

In the LA literature causal claims tend to centre around *effectiveness* and *evidence*. Effectiveness implies the influence of some thing (the cause) on a desired outcome (the effect). This means that a strong causal claim should result from a LA study if it has solid supporting evidence that it is effective.

Calls have increasingly been made in LA to improve the evidence that points to its effectiveness [35], particularly in the evaluation of interventions [90, 136]. Associated with this is an undercurrent of calls for a stronger link to causal claims [120]. This raises an interesting question: how precisely is LA currently addressing the challenge of making causal claims?

2.2.1 Associations without causal claims

LA tends to focus on statistical associations and avoids making causal claims. For instance, a review of student interventions using data found that only 11 of 689 papers assessed the effectiveness of the intervention [90]. Another review of student facing dashboards fared better, with just over 1 in 4 providing some kind of evidence for their effectiveness [8].

These results have led to criticism that the field is not adequately assessing its own effectiveness, and that claims are rarely made with strong supporting evidence [35, 68, 179]. While this is a poor result, Viberg et al. [179] point out more charitably, that it could be due to the potential of LA outpacing the rigorous evidence required to demonstrate its effectiveness. Nonetheless, it seems that the field needs to get better at generating causal claims that can be used to strengthen the claims that it can make about effectiveness.

2.2.2 Claims made by experiments

An *experimental* argument for causation is usually built around the execution of a Randomised Control Trial (RCT). Where causal claims are explicitly made in Learning Analytics they are commonly in the form of RCTs [9, 20, 70, 85]. RCTs are often referred to as the 'gold standard' in both evidence based science [35] and education [162], so much so that studies are commonly grouped into two categories; experimental and observational, with the latter essentially meaning you have not run an RCT.

An RCT experiment is built around measuring the effect of some *intervention* on some *outcome*. The experiment is attempting to remove any effect of confounding variables, as depicted in Figure 2.2.

The experiment is designed so that the assignment of intervention variable is randomised, effectively removing any potential causes of the intervention other than random chance. This can be seen in Figure 2.3 by the removal of the causal path from the *Confound* to the *Intervention*.

The removal of all other causes of the *Intervention* through its random assignment prevents possible common causes of the intervention and the outcome, so that any future association found between the intervention and the

$$\begin{tikzcd} & Confound \arrow[dl] \arrow[dr] & \\ Intervention \arrow[rr] & & Outcome \end{tikzcd}$$

Figure 2.2: Prior to the set up of a RCT experiment. The causal effect of *Intervention* on *Outcome* is obscured by the confounding variable(s) *Confound*.

$$\begin{tikzcd} Random \arrow[d] & Confound \arrow[dr] & \\ Intervention \arrow[rr] & & Outcome \end{tikzcd}$$

Figure 2.3: After to the set up of a RCT experiment — by randomising *Intervention* the potential effect of the confounding variable is broken; the only cause of *Intervention* is the *Random* process.

outcome variables can be interpreted as causal. The random assignment of *intervention* (and not *outcome*) also encodes the information about the *direction* of the causal effect being investigated, encoding the effect of interest as $intervention \rightarrow outcome$.

As an example, imagine a researcher is setting up an experiment to understand the effect of revising for a quiz on a student's quiz performance. In the experiment the students would not get to decide if they revised or not They would be randomly assigned into two different groups; one forced to revise (the treatment group) and the other prevented from revising (the control group). Any association between the two variables could then be interpreted as an unbiased estimate of the causal effect of revision on examination performance. Such an experiment would certainly make a strong causal claim, however formulating it raises a number of practical and ethical challenges. There are practical issues in trying to force a student to revise, who would not have otherwise, and there are ethical issues in preventing a student from revising who is committed to doing well in the subject. Even with those challenges overcome, it would also be incredibly difficult to ensure that the quality and quantity of the revision was comparable amongst the treatment group.

Despite their strong reputation, the effective use of RCTs in education is hampered by several factors: well designed experiments are not always practical; they can run into ethical issues [185]; and they may not adequately represent what happens outside of the experiment. Practically, an RCT requires careful design an execution so can be costly to run well. Authentic learning environments are often dynamic and evolving, and keeping exogenous variables constant can be challenging in this setting, as well as retaining trial participants as their motivations for learning change over time [20]. If some intervention is suspected to be helpful it may be deemed unfair and unethical to give it to some students and not others. RCTs are usually, by design, set in a restricted setting and may not adequately capture the messiness of the real world. They are, by their nature, made out of context [28, 162].

Because of these challenges RCTs are rare in an educational setting [27, 35, 90]. Yet practical and ethical blocks to the use of an RCT should not be considered an absolute barrier to making decisions around using an intervention. Parachutes, for instance, are still widely used despite the lack of RCT evidence of their effectiveness [156]. There are no RCTs demonstrating that surgery, attempting to stop drink-driving, or lecture attendance are effective. And yet we still make well informed interventions in each of these circumstances.

An alternative experimental design is the *Natural Experiment*, where the two groups are selected *naturally* rather than by *design*. Natural experiments can arise opportunistically when the assignment of the intervention is out of the control of the researcher and suspected to be effectively random. However, understanding if the assignment of the intervention is truly random is challenging. Although the assignment to the intervention being 'beyond the control of the researchers' is important, this does not exclude the possibility that other, potentially hidden, factors are influencing the assignment of the intervention and the outcome which can result in a biased estimate. For instance, participants may be 'self-selecting' into the intervention in some way, or there may still be potential confounding from unobserved variables that influence the exposure to the intervention and the outcome [24]. Often other variables need to be used for statistical control in order to minimise bias, arising from the causal structure of how the variables influence the selection process and the outcome [61]. Glymour [48] claims that understanding if the intervention assignment is essentially random requires knowledge of the causal structure of the system being investigated.

Natural experiments have been a popular alternative to RCTs in LA, with these two approaches to causal inference making up the vast majority of causal claims in the field. One advantage that natural experiments have over an RCT in the educational setting is that they are embedded in an authentic learning context, usually centred around attempting some intervention. In one such example Gašević et al. [42] compared the use of graded feedback with non-graded feedback on video annotations. This natural experiment was uncovered by looking at the instructor annotations in the video log data and realising that two concurrent courses, with some students enrolled in both, showed that one course used graded feedback and the other non-graded feedback. In a case testing a change in learning design, Mullaney and Reich [122] compared two versions of the same large online course, run in very close succession, with the only difference being the timing of the release of content. Loton et al. [102] compared immediately before and after the introduction of a new 'block model' of course delivery. The authors recognised this required additional statistical control, applying 'comprehensive statistical control in mixed-effect cross classified models' in order to account for changes in the student, teacher and course composition, with the choice of these variables based on their theoretical understanding of the system. This highlights the care required when running experiments outside of laboratory conditions.

Motz et al. [121] make the case that we should deliberately be embedding experiments more often in LA in order to draw stronger causal inferences. This does, however, raise ethical concerns; if the intervention is thought to be effective then surely this disadvantages the control group? Krause [89] noted this concern, and so experiment was achieved by staggering the intervention, with the control group simply receiving the intervention four weeks later in order to run the experiment and still ensure that all students could benefit. Roediger III et al. [139] and Carvalho et al. [18] point out a number of ways to address this concern, such as by focusing on small sections of a course with a large number of students, so as to have a small effect on the individual students' result but a measurable effect on the targeted intervention. Motz et al. [121] make the case that good teaching requires ongoing experimentation, which has certainly been my approach, and that perhaps "ethical considerations do not hinge on whether experiments should be embedded in classrooms, but whether well-designed, controlled experiments should be embedded".

Although experiments are robust way of making causal claims they are not always practical to execute or ethical to run. The experimenter must also strike a difficult balance between controlling the environment to satisfy the conditions of an RCT, and allowing the experiment to inhabit an authentic learning setting. Experimental methods will always form part of the knowledge building in LA, but they do not cover the wide range of circumstances and contexts in which LA wants to make causal claims.

2.2.3 Claims made by adjustments

Where RCTs are not possible confounding can potentially be accounted for using a *quasi-experimental* approach, which involves *adjustment* of the measure of association so that it can be interpreted as a *causal effect*. Without the adjustment the association measured between the intervention and the outcome is said to be a *biased* estimate of the causal effect. When each source of bias is *adjusted for* the remaining association is claimed to be causal. I have chosen to group these methods together as *adjustment methods*, as the term quasi-experimental is used inconsistently across different scientific fields.

The five main adjustment methods used in LA are outlined below.

2.2.3.1 Claims made by regression with covariates

The most common way to adjust the estimate of interest is by including additional variables, or covariates, as part of a regression model to marginalise out their confounding effect on the causal estimate. These extra variables are often said to be 'controlled for' or 'conditioned on'. This means that any influence these variables have on the outcome (say, passing the examination) and on the intervention (revising for the examination) has been considered and filtered out. Any resulting association between the intervention and outcome can then be interpreted as causation, with the proviso that this effect may still be accounted for by some unknown process.

However, this approach depends on choosing the right variables to condition on. For instance, if a covariate lies on a causal path between the intervention and outcome then including it in the model will mask the causal effect — the effect of the intervention on the outcome is mediated through the covariate. If a covariate is a common cause of the intervention and outcome then it should be included, but if it is a common effect then it should not [140]. Deciding what to do requires knowledge of the causal structure relevant to the effect of interest [104, 130]. There are many approaches to making these choices that attempt to adjust the way the association is measured to reduce confounding, however frequently they do so without explicitly mentioning the assumed underlying causal structure. Most often these choices are made for sensible, scientific reasons, but historically these reasons have often not been reported or justified [157]. There is also the temptation to 'control' for as much as possible, throwing as many variables into the model as possible. This approach has been critiqued by McElreath [117] as creating a 'causal salad', an opinion that I have some sympathy for.

A range of LA studies use covariates in regression models to estimate causal effects, with typically well considered selection of covariates based on strong, albeit varied, arguments. Some studies appeal to the scientific understanding of the system under study to identify suitable covariates, citing relevant literature as justification (e.g. Padgett et al. [125]), or providing a scientific argument (e.g. Beheshitha et al. [5], Dawson et al. [27], Grann and Bushway [52], Padgett et al. [125]). Other studies use existing theoretical frameworks to guide the choice of covariates (e.g. Beland and Kim [6]). Several studies tried to better capture the nested structure of the system by using multi-level (hierarchical) models [5, 132], with Dawson et al. [27] noting the change in the final inference when accounting for this structure and Pogodzinski et al. [132] using their framework to account for potential missing covariates.

Whereas all these examples have generally strong arguments for the choice of covariates none explicitly described the assumed causal structure that these arguments are based on. This makes it hard to generalise the methodology to new contexts as each argument is specific to the particular case at hand. Furthermore, as the implicit causal structure is not described the underlying assumptions are difficult to interrogate.

2.2.3.2 Claims made by propensity scores

Propensity score methods operate by seeking to normalise the propensity of each individual data point to be selected into the intervention (or 'treatment') group by calculating a *propensity score*: the probability that an individual is assigned the intervention. The data is then adjusted in some way to effectively normalise the selection process so that it mimics random selection. Methods for this normalisation include matching, weighting, stratification and others, and all have a similar impact on emulating experimental conditions [149]. Matching, where data points are matched based on similar propensity scores and those without close matches are omitted, is most popular in education but has the challenge of biasing the estimate with excessive pruning of data points [82].

Propensity score methods have increased in popularity in education since the turn of the century [168], with calls for greater use for them in LA, particularly in the evaluation of dashboards [8] and intelligent tutoring systems [53]. Within LA, matching methods have been used to evaluate the effectiveness of various interventions [14, 31, 55, 181], ranging from adaptive learning implementations [119] to process focused feedback products [93]. In each case the aim was to minimise confounding due to selection bias, where the concern was that the two groups, those receiving the intervention and those not, may have slightly different attributes that could also affect the outcome. Other teams have used propensity score weighting to make stronger claims about the effect of an intervention [80, 91], such as dashboard use on academic outcome [177], or estimating the potential individual effect of a future intervention to prioritise resources [124]. Additionally, a recent paper by Karimi-Haghighi et al. [75] using a variety of methods, including propensity score matching and weighting, is of note for being very explicit in staking a causal claim.

These methods still require substantive knowledge of the causal structure of the system [56]. One must know what is affecting both the selection and the outcome in order to emulate a RCT. It is desirable to clearly state what variables are included to estimate the propensity score [168], a practice not embedded in any of the propensity score methods. This makes it difficult for teams to be sure that they have constructed a valid estimate of the propensity score.

2.2.3.3 Claims made by regression discontinuity design

Regression Discontinuity design (RD design) mimics an RCT by examining the data points around a relevant cut-off for the intervention (the discontinuity). For instance if scholarships are awarded for a test score of 80 or above, then one might examine all students with test scores from 77 to 82 and treat the assignment of the intervention (receiving a scholarship) as essentially random. The underlying assumption is that the students in this group, with scores from 77 to 82, are of a very similar academic ability and would, if the scholarships were not awarded to anyone, all have a similar chance of future success. This way those with scores of 77–79 are in the 'control' group and those with scores of 80–82 are in the 'treatment' group.

This approach has been widely used in education and shows some promise, with more testable assumptions than some other quasi-experimental methods [114]. Key to implementing, however, is the requirement that the problem can be framed around some kind of threshold, as the effect estimate should be viewed with more skepticism as we move further away from the threshold [22]. A good example of this technique being used in LA examines the effect of academic dismissal on academic outcomes by looking around the threshold for dismissal [23]. Another study looked at the students just below and just above the cut-off for a chemistry entrance exam, in order to estimate the effect of a gateway course on subsequent achievement in future, higher level, chemistry courses [152].

Some of the challenges for using RD design include compound treatment and sorting (see Eggers et al. [30] for details), which require an understanding of the causal structure to ameliorate. Additionally, this is only a viable technique in a small neighbourhood around of a continuous variable around a specific threshold, which decreases its utility for a number of authentic scenarios as LA models tend to apply to a whole cohort.

2.2.3.4 Claims made by difference in differences

The *Difference in Differences* approach exploits having two groups of similar composition at a similar moment around an intervention being introduced. Like a natural experimental design (Section 2.2.2) one group is assigned the treatment and the other is not. The outcome of interest is measured before and after the intervention, for both groups, and the difference is calculated for each group. Looking at the difference before and after the intervention attempts to control for confounding variables that are constant over time. The difference between these two differences is then calculated, which results in the causal estimate (see Fredriksson and Oliveira [38] for more detail).

In order to make these claims some conditions must be met. The data must satisfy the *parallel trends assumption*: that the groups would have otherwise followed roughly the same trend had the intervention not occurred [38]. Assessing if this assumption is met requires an understanding of the causal structure, parallel trends assumption is equivalent to assuming that the two groups are "equally affected by unobserved confounding" [190].

This method is relatively popular in the analysis of large educational datasets [22], and there are also a handful of recent examples of using a difference in differences approach in LA. For example, Zilvinskis et al. [191] looked at the effect of a conditional admissions policy on outcomes, utilising difference in differences to mitigate confounding. Others have used the approach to estimate the effect of seeding course forums with content on forum activity [154], or the effect of automated feedback on the academic performance of low achieving students [106].

However, difference in differences still requires an experimental design, and the assignment of an intervention and a control group. This means that all the ethical and practical concerns that apply to RCTs and Natural Experiments (see Section 2.2.2) still apply. Additionally knowing if the parallel trends assumptions is met requires the knowledge the causal structure underlying the system. The two groups must be sufficiently similar, and remain similarly impacted by external effects, for the duration of the experiment.

2.2.3.5 Claims made by instrumental variables

Instrumental Variables (IV) leverage the linear regression framework and an additional variable (the instrument) to help estimate the causal effect [3, 158]. The IV must satisfy certain conditions to be useful; primarily that it must influence the outcome only through the intervention, and also that it does not change direction in the way it influences the intervention[1]. This causal relationship is

[1]This condition is called *monotonicity*, or sometimes that there are no 'defiers'.

depicted graphically in Figure 2.4, for intervention X and outcome Y.

$$
\begin{tikzcd}
 & & U \arrow[dl] \arrow[dr] & \\
IV \arrow[r] & X \arrow[rr] & & Y
\end{tikzcd}
$$

Figure 2.4: The chosen IV must influence Y only through X. Here U is the unobserved confounds that are the source of the bias.

With the conditions met, the various regression models of $Y \sim X$, $Y \sim IV$ and $X \sim IV$ can be used to provide an unbiased estimate of the causal effect by looking at ratios of the coefficients estimated [158].

These techniques have been used in a small number of cases in LA. Most are using this technique to help eliminate bias when estimating a causal effect, such as the effect of: choosing to take online courses on degree completion [37]; using third-party software to help improve performance in language courses [33]; and the timing of taking technical courses on retention and behaviour [50]. Others used this method to examine multiple sources of bias in survey responses [165], while Davis et al. [26] used IV to minimise some bias but explicitly noted that the estimate remained non-causal because other sources of bias were still present.

There are several challenges particular to the IV method. Finding a 'strong' instrument is not always possible, as it must have a strong, direct relationship with the intervention that is also not confounded by other variables. Additionally, the models must be linear, or transformable to linear models, in order to compute the effect of interest. Finally the IV method is only valid for weak or moderate confounding [107].

2.2.3.6 An observation on the use of causal structure

The instrumental variables approach (above, Section 2.2.3.5) provides a noteworthy change in how the problem of confounding is addressed; the method does not try to reduce confounding effects by mimicking randomisation, it explicitly leverages the causal structure of a system in order to obtain an unbiased estimate. Up to that point all these adjustment methods required strong existing knowledge of the causal mechanisms involved but not in a structured fashion. Instead they require that causal knowledge be applied, in each particular context, to inform variable selection. By controlling for an appropriate collection of variables other sources of association between the intervention and the outcome are minimised. Despite relying on an understanding of the causal structure of the system, none of the approaches so far explicitly show what the assumed structure is. The IV approach utilised a very particular causal structure, but this approach can be generalised. The final adjustment method that I will introduce (Graphical Causal Models, Section 2.2.3.7) addresses the causal structure in a very direct and transparent way, however they are almost entirely absent from the field of Learning Analytics.

2.2.3.7 Claims made by Graphical Causal Models

Graphical Causal Models (GCMs) directly represent the causal structure with a diagram called a Directed Acyclic Graph (DAG). The nodes of the graph represent variables, and the directed edges (arrows) represent causal influence. A causal relationship such as "A causes B" can then be transcribed to the graph as $A \to B$.

For instance, if we think that time spent learning a topic (LT) affects test score (TS), but that intelligence (I) affects both these, we could represent it as a Directed Acyclic Graph (DAG), as in Figure 2.5:

$$\begin{tikzcd}
I \arrow[d] \arrow[dr] & \\
LT \arrow[r] & TS
\end{tikzcd}$$

Figure 2.5: Directed Acyclic Graph (DAG) modelling the influence of learning time (LT) on test score (TS), with the confounding variable intelligence (I). Example from Lübke et al. [104].

From this graphical causal structure we can identify if a particular causal effect is *identifiable*, which means that an unbiased estimate can be found by selecting a particular set of covariates to include in a statistical model [127]. Pearl [127] developed the *Calculus of Interventions*, or *do*-Calculus, that allows for the calculation of the set of variables required to adjust for to get an unbiased causal estimate, assuming the DAG sufficiently describes the system. Applied to the GCM depicted in Figure 2.5 the *do*-Calculus would inform the statistical modeller that they should include the covariate I in a regression model, along with the intervention and outcome variables LT and *TS*, in order to interpret any association between LT and *TS* as causal.

Boerebach et al. [10] provide an example of using multiple GCMs to compare competing theories, but this is an exception in the field of education, let alone LA. Structural Equations Models (SEMs) are closely related GCMs. They have a shared history [130] and represent the relationships between variables with a directed graph. However LA practitioners have tended to use SEMs primarily for confirmatory factor analysis rather than to make causal claims [184]. For instance Fincham et al. [36] utilise SEMs for confirmatory factor analysis to help validate a theoretical model of engagement. Liu [100] also uses SEMs for factor analysis, but additionally leverages existing theory and SEMs to make claims about the relative influence of several variables on blended learning student acceptance.

Many papers use examples from education to help explain how these models work [54, 140], including the example from Lübke et al. [104] shown in Figure 2.5. Pearl and Mackenzie [130] also use an educational example, the Berkley's Admission Paradox, to describe a particular case of confounding and how it can be addressed. It is curious then that claims made in LA from graphical causal models are exceedingly rare. One barrier in the use of GCMs is the initial creation of the model. Generally this needs to be built from existing knowledge, not learned from the data [138].

2.3 Research Gap 1: Few causal claims are being rigorously made in Learning Analytics

Section 2.2 has demonstrated that while some causal claims are being made in LA, they tend not to address the underlying causal structure of a system. They are also relatively uncommon. Despite the need for strong causal claims in order to enable actionable LA (Section 1.1) Ifenthaler and Yau [68] and Viberg et al. [179] have both noted that use of techniques to enable this level of evidence is rare. Reviews by Larrabee Sønderlund et al. [90] and Bodily and Verbert [8] found a very low level of evidence, on the whole, to assess the effectiveness of interventions. Where models are used these are typically for prediction, and not causal explanation [99]. Several authors have called for stronger evidence in LA [35, 120, 136], which I have framed as making stronger, more rigorous, causal claims. This brings me to my first research question.

Research Question 1: How can Learning Analytics make rigorous causal claims?

2.4 Interventions require a shared understanding between interdisciplinary stakeholders

For LA to be actionable it requires more than just a rigorous causal claim; the causal claim needs to convince a stakeholder to act. However, addressing the challenges in understanding a system as complex as learning requires the integration of many different viewpoints and a range of expertise [142]. As such LA operates in a 'middle space' between the learning and analytical sciences [87, 163] which necessitates bridging epistemic boundaries [46]. This means the field can suffer from problems of communication, where people from very different cultural backgrounds talk past one another. Often, we see educational experts excluded from the conversation, unable to judge or evaluate highly technical approaches that rely upon advanced statistics, machine learning, and other methods that quickly come to resemble a black box [126]. As a result, educational stakeholders often find it difficult to trust the results of a data analysis enough to act upon it. This problem is likely to be contributing to the failure of LA to influence institutional decision-making processes [105].

A key reason for this difficulty is the underlying complexity of learning, learning environments, learning institutions and the broader context in which learning happens [25]. One view of complex systems claims that they require multiple models for a complete understanding [103, 144]. This implies that there will be multiple ways to abstract, analyse and understand any learning system. While this richness of perspectives is important for LA [142] it creates a situation where aspects of the construction of LA may be more or less accessible to different stakeholders, depending on their realm of expertise. Being able to bring a wider variety of expertise into the design process of LA is strongly desirable [187], and the emergence of Co-Design methods (see Section 2.5) has the potential to build a shared understanding between LA users and developers.

In order to proceed with this discussion, I will define two broad groups that must work towards a shared understanding in order to achieve actionable LA: *context experts*, and *data experts*.

Context experts are those that will be using LA products and technologies; they reside close to the *context* in which the LA product operates. These people may be students, practitioners, teachers, administrators, leadership; the common thread being they will be hopefully utilising LA to support them in making decisions, and they are generally *not* experts in data analysis. I will examine this group, and the problems they face, in Section 2.4.2.

In contrast I take data experts to be those who understand the technical details and are often performing a data analysis and/or developing LA products, but not necessarily using them to make decisions. These people may be developers, analysts, or data scientists. Their common characteristic is that they understand the data side of LA well, but not necessarily the context in which it will be used. I will examine this group, and the problems they face, in Section 2.4.1.

This data / context division is no doubt a simplification, and may need revision and a more nuanced approach in the future. It nonetheless helps to explore tensions that arise in communication between these two "tribes" in LA.

2.4.1 Data experts and the context barrier

I am defining the data experts as those who have the technical capacity to analyse data, but may not have the experience or contextual knowledge to translate what they find in the data to how it fits with the lived experience of those for whom the LA is designed. Despite early promises of Artificial Intelligence, Machine Learning algorithms and Big Data to revolutionise and 'disrupt' education [76, 180], a growing narrative suggests that the data experts can only learn so much from the data alone [44, 186]. As seductive as it may be to believe in the independence and objectivity of data, numerous decisions are made by people with their own opinions, assumptions and prejudice, at numerous points in the collection, analysis and dissemination of data [47]. In order to account for these data dependencies this subjectivity should to be represented in the modelling somehow, but this requires knowledge of the educational context which is something that the data expert may need to seek out [45, 111, 153].

Without a proper appreciation of the needs and capabilities of those who intend to use the analytics, data experts run the risk of those without their expert knowledge misinterpreting and misusing the data, or designing systems that do not improve outcomes. One of the examples examined by Kitto et al. [83] highlights the danger of systems being designed to optimise for internal accuracy, which, while a desirable feature for the data expert, has very little impact on real outcomes of interest to stakeholders. Boysen et al. [15] found that use of averages lead to faculty and administrators making incorrect judgements of teacher performance, where small, insignificant differences were interpreted as meaningful. It is then natural to question the utility of including such metrics, such as the mean in LA reports and dashboards, if judgements should not be based on them. Even when the choice of metrics seems well thought out, the choice of comparisons can have unintended consequences when the context is not fully understood. When students are comparing themselves with other students, for instance, they run the risk of moving towards a more competitive motivation instead of content mastery [71]. Gasevic et al. [45] provide another example by highlighting the way in which internal and external threats to data quality are incredibly difficult to address without contextual knowledge. Take the challenge of measuring the time a student is spending 'on task' in an online learning environment. The online learning management systems may not effectively stop tracking time 'on site' when a browser window is left open for an extended period of time. This means that the consistency of the data is compromised because large values are meaningless and dramatically distort aggregations of time data. And then there are external threats to data quality, for instance as there is no way to realistically know when a student is engaged in the online content they have open on their computer, or if they are off making a cup of tea. Understanding these threats to the quality of the data comes from understanding the context in which a student is learning, not from the data itself.

This issue affects more than just the field of LA. Whatmore [183] discusses the challenge of a 'scientific' group, performing modelling on flooding risk, informing the 'vernacular' group, who are experiencing the flooding first hand. There was a perceived failure of the scientific modellers to "learn from the locality or consult those who live" in the area, those with the contextual knowledge. This results in a knowledge controversy between the first hand experience of those living where the water is rising and those building the mathematical models of flooding risk.

So despite the burgeoning development and deployment of sophisticated analytical tools, methods and products, there is only so far that the data experts can go. Eventually they will run into a *context barrier*, where the continued use of analytics without the appropriate contextual knowledge may lead to irrelevance and misuse. The data alone is not a true window into natural processes of learning, but instead must be considered *in context* in order to understand their relationship with what they mean [69].

In short, the data experts cannot build effective LA alone.

2.4.2 Context experts and the data barrier

I am defining the context experts in LA as anyone embedded in the learning system in some way with particular knowledge of their context, such as the students, the teachers, the administrators, or the institutional leaders. What links this diverse group together is their need to make evidence based decisions, but without the in-depth technical training to analyse relevant learning data for themselves. By definition these stakeholders have knowledge relevant to their context in the learning system, but will not necessarily have expertise in analysing data. Helping context experts to understand and act upon LA is made challenging by this lack of expertise [15, 45].

This challenge seems to arise every time educational decisions need to be made, at every scale, and with a diverse range of non-technically experienced stakeholders. University leadership have consistently highlighted that gaps in perception between stakeholders, data literacy of leadership and the technical capability of staff are barriers when engaging with LA innovations [171]. This lack of understanding of what can be achieved with analytics can diminish its potential at a strategic level [45]. Moreover, the inability of decision makers to be able to critique and challenge an algorithm or statistical result seriously hampers efforts towards transparency and fairness [84]. Faculty and administrators also often lack the the technical expertise to correctly interpret data. This has led to erroneous conclusions drawn from statistical analysis regarding the performance of teachers. They simply do not have the expertise or training to interpret the data, as presented, correctly. Boysen et al. [15] found the lack of expertise in reading the of uncertainty in point estimates, in the form of error bars, led to incorrect judgements about teacher performance. Misinterpretation and misuse of dashboards is common with many stakeholders, however learners in particular find it challenging to interpret them, and can become less motivated from viewing dashboards depending on their frame of reference [92]. This creates a challenge for students to make the most out of their learning as new LA innovations arise, now needing a level of data literacy for learning [182]. Together, these issues raise concerns about a data and algorithmic literacy divide, where those that can leverage the data intensive world can participate more fully than those that cannot [51].

Equipping users with the skills to interpret LA would help but their are often insufficient opportunities to do so [170]. Even when opportunities are provided it can be challenging to help context experts meaningfully participate in the shaping of the technical, algorithmic elements of LA systems [66]. For good or ill the datafication of education has arrived [69], and context experts trying to make sense of the deluge of information will inevitably run into the *data barrier*; where the complexity of the system they are trying to analyse is beyond ‘out-of the box’ analytical solutions or their capacity to interpret.

In short, the context experts cannot correctly interpret the data alone.

2.4.3 The context / data divide

This sensemaking ‘gap’ between the context experts and the data experts [112] I will call the *context / data divide*. It is within this space that we need to look for possible ways to create a shared understanding.

2.5 A shared understanding via Co-Design

The good design of LA is not an individualistic pursuit, with many claiming that it requires collaboration between stakeholders and a human-centred approach [17, 29]. A variety of techniques, under the umbrella of Participatory Design and Co-Design methods, attempt to address this through the implementation of probes, generative toolkits, and prototyping. Probes are used to find out initial impressions, generative toolkits for refining and moving towards a protoyped product, and prototyping for constructing an artefact that can be evaluated [145]. The aim is to bring the multitude of voices together in a collaborative design space. Each of the techniques generates artefacts that then need to be interpreted by the designer. They are often built upon similar techniques in the field of knowledge elicitation, such as concept sorting, semi-structured interviews and concept mapping, where the *elicitor* is trying to make explicit the tacit knowledge of the *expert* [148]. Knowledge elicitation is framed as a one-way flow of information; the elicitor is trying to extract knowledge from an expert. Whereas knowledge elicitation from context experts, such as students and teachers, to the designers of intelligent tutoring systems [111] and LA frameworks [7] has been highlighted as critically important, this process is focused on adequately embedding the contextual knowledge in the LA system, not necessarily promoting a shared understanding between the stakeholders. This contrasts with the use of generative toolkits within a Co-Design framework that attempt to design *with* instead of design *for* the end user [145]. So although there are similarities in some of the methods of knowledge elicitation with Co-Design, the intent and ethos of Co-Design, particularly at the post-ideation and pre-prototype phase, is more likely to promote a shared understanding between the users (context experts) and the designers (data experts).

Participatory and Co-Design techniques have been growing in popularity in LA [146] as have calls for their use [29, 88]. There are several good examples in LA, however as Sarmiento and Wise [146] note they are recent and not prolific. Card-based generative tools have been used in several cases for informing LA design choices. Vezzoli et al. [178] developed one such tool to help in the development of a language game app, but they can have a broader scope of understanding the overall LA system requirements [1]. Martinez-Maldonado et al. [108] used blended card sorting, focus groups and low-fidelity prototyping, both with students and teachers, to understand users' perceptions around complex multi-model LA. Both student and teacher knowledge was also utilised in another case, where the participatory design sessions were used to uncover themes that were unexpected by the designers [135]. These authors analysed the qualitative data from the sessions, with multiple coders, to then synthesize the results into themes to be guiding principals for dashboard design. Holstein et al. [66] implement an end-to-end Co-Design approach embedded with a variety of Co-Design methods at various stages. They use a probe early on in the form of a 'teacher superpowers' prompt, 'speed dating' with teachers to rapidly think about possible futures, through to low to high fidelity prototyping.

The Co-Design processes used in LA typically elicit tacit knowledge from the user (the context experts) into explicit knowledge that the designer (the data experts) can use. Although this may be entirely appropriate for designing for the user, it may not be as useful for developing a shared understanding. An exception to this one-directional flow of knowledge, when Vezzoli et al. [178] utilised a Co-Design phase that "aimed to generate a shared understanding between participants and researchers". Here the context experts and the data experts are both being set up by the researchers to gain knowledge in this situation, though the authors did raise concerns that the context experts may "lack the data literacy background" [178] to fully assist in the Co-Design process beyond the initial problem-identification stage. Kitto et al. [84] claim that to truly participate all parties must be able to interrogate the system being designed. The asymmetry in the flow of knowledge may contribute to a power imbalance between the different participants in a Co-Design process, which can be further exacerbated in contexts that require technical knowledge [29]. Sarmiento et al. [147] noted a concern for these potential power imbalances and developed specific protocols to help ameliorate this issue when using small groups of students in their participatory design workshops, although the aim was to help students feel comfortable and valued, rather than to facilitate a two-way flow of knowledge. The adoption and development of these techniques demonstrates that many LA researchers and practitioners understand that getting the stakeholders into the same room together is not enough. We also require an awareness of the power balance and value alignment [29].

Co-Design looks promising for the design of quality LA through a wider participation of key stakeholders, but the tooling and techniques as they stand seem not to, at least practically as reported so far, facilitate a lasting shared understanding of the challenges of the system being designed. This might be improved with different generative toolkits, the phase of Co-Design with the strongest emphasis on 'designing with' [145]. Addressing the flow of knowledge imbalance may also be a way forward. As it stands the existing generative toolkits used do not directly inform the abstract data models, there is still considerable work for the data expert to do in order to implement the context experts elicited knowledge [29, 178]. If both parties have a similar amount to learn in order to collaborate then perhaps this is a way towards a shared understanding.

2.6 A shared understanding with strong theory

Whereas Co-Design shows potential to facilitate a shared understanding between context experts and data experts on a case-by-case basis, the development of strong theory may provide a more generalisable shared understanding. A program of well coordinated theory construction allows modelling by means of mathematics or simulation, which helps reduce the reliance on pure empiricism on one hand [12] but also helps explain phenomena on the other [13, 164], and facilitates knowledge transfer within a community. This ambition has been identified in LA, with a growing number of calls for a more theory-based construction of LA [45, 73, 88, 109, 159, 172, 186].

Here I will follow the lead of Sutton and Staw [164] and begin with what good theory is *not* (Section 2.6.1), and from there examine a methodology for constructing strong theory (Section 2.6.2), before discussing how strong theory might support a shared understanding between context experts and data experts (Section 2.6.3).

2.6.1 The symptoms of weak theory in Learning Analytics

Sutton and Staw [164] approach the problem of understanding what theory is by articulating five examples of what theory is *not*:

1. References are not Theory. What more is needed is for authors to "explicate which concepts and causal arguments are adopted" from their sources.

2. Data are not Theory. The data, on its own, describes associations, but "theory explains *why* empirical patterns were observed".

3. List of variables or constructs are not Theory. "A theory must also explain why variables or constructs come about or why they are connected".

4. Diagrams are not Theory. A theory must also justify the causal connections in the diagram.

5. Hypothesis (or Predictions) are not Theory. These "are concise statements about what is expected to occur, not why it is expected to occur."

Sutton and Staw [164] note that authors in the social sciences routinely use these five examples in lieu of theory, despite consensus that they do not constitute strong theory. This is an understandable oversight, as building theory is difficult due to the challenge in defining theory, understanding what good theory is [164], and constructing robust theory [13]. However, without strong theory, problems can arise.

In various fields the problem of meaningful transfer of results and ideas across the context / data divide has been described as a lack of robust theory [32, 39, 123, 164] stemming from a lack of coordinated theory construction [13]. It has been claimed that the field of LA is not in a position of having developed robust, strong theory [4, 43, 44, 159, 186]. A deficiency in strong theory can hinder progress in several ways:

- It becomes very challenging to know how to intervene in a system to achieve a desired outcome due to the lack of causal explanations provided by a strong theory [13]. This impacts the effectiveness of LA by preventing it from becoming reliably actionable [186] and isolated from the context [44, 86].

- Data can become the primary driver of evidence in the absence of strong theoretical explanations, letting what we can measure drive our values instead of what we value drive what we measure [86].

- Values may be driven by those important to one side of the context / data divide. For instance the value of accuracy highly prized by machine learning may be favoured at the expense of contextual relevance [83].

- A lack of cohesion can arise on either side of the context / data divide because there is no common language for describing phenomena of interest to both sides [40, 137], exacerbated by the lack of thinking tools to adequately navigate the space of possibilities [13].

The creation of good theory is a challenging problem. The issues raised by Sutton and Staw [164] about what theory is not still resonate nearly three decades later. The field of psychology has been very active in to understand the problem of theory [123], and it is from there that I adopt a framework developed by Borsboom et al. [13] for constructing strong theory, outlined below in Section 2.6.2.

2.6.2 Constructing strong theory

Before we introduce a method for constructing theory it is worth outlining exactly what we are trying to create. To describe what I mean by 'theory' I will use the following model, introduced by Borsboom et al. [13], and shown in Figure 2.6.

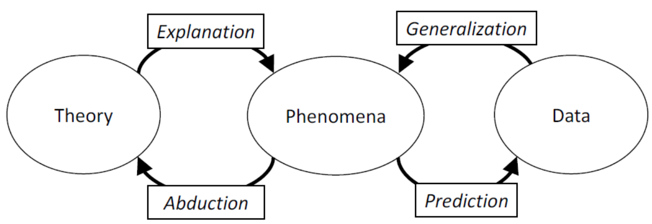


Figure 2.6: The relation of theory to observed phenomena and data (Figure 1 from [13]).

Borsboom et al. [13] delineates between *theory*, *phenomena* and *data*. The theory itself does not explain the data directly, it tries to explain the phenomena of interest. These are taken to be the stable features of the world that we seek to understand [189]. Ideally these *phenomena* should be observable in the data and in the world itself [13], which means they are observable by both the context experts and the data experts. Borsboom et al. [13] claims a good theory should explain the phenomena of interest. The theory should be generated from the phenomena by some abductive reasoning, that is assessing the many possible explanations for the phenomena and settling on the 'best'. This forms a *prototheory*, or *verbal theory*. Much theory development stops at this point, with the resulting informal description of the phenomena. However, claims have been made that without further abstraction into a *formal theory*, verbal theories cannot provide a common language to lend themselves to collaborative development [40, 137].

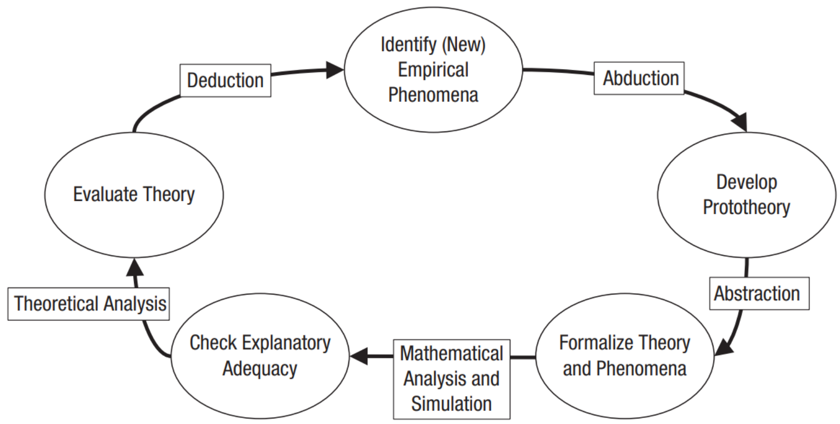


Figure 2.7: Theory Construction Methodology cycle (Figure 2 from [13]).

The *Theory Construction Methodology* (TCM) cycle [13] provides a way of moving from the prototheory towards a more robust theory that can adequately explain phenomena and provide a common language. This process is outlined in Figure 2.7. For example, imagine we are noticing the phenomena that scores on cognitive tasks tend to correlate positively with each other. A common explanation for this is that there is some underlying 'general intelligence' possessed by each person tested [49]. However Van Der Maas et al. [174] formulate a different prototheory, that this effect emerges from the positive interactions between different cognitive processes, and not some single underlying factor. They then *abstract* this into a formal theory, utilising assumptions about the underlying cognitive processes and their assumption that the different processes are mutually beneficial. The end result is expressed most succinctly in Equation 2.2 and also graphically in Figure 2.8.

$$\frac{dx_i}{dt} = a_i x_i (1 - x_i/K_i) + a_i \sum_{\substack{j=1 \\ j \neq i}}^{W} M_{ij} x_j x_i / K_i \quad \text{for } i, j = 1 \ldots W \tag{2.2}$$

With the formalised theory Van Der Maas et al. [174] are able to run simulations to explain several phenomena such as the low predictability of intelligence from early childhood performance and the hierarchical factor structure of intelligence. This forms their evaluation of the theory and checks its explanatory power, the last stages of the TCM cycle (Figure 2.7).

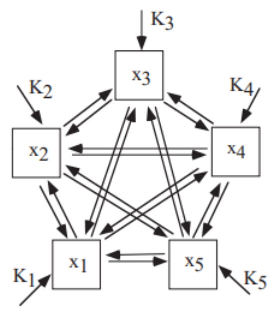


Figure 2.8: Mutualism model of intelligence (Figure 1 from Van Der Maas et al. [174]).

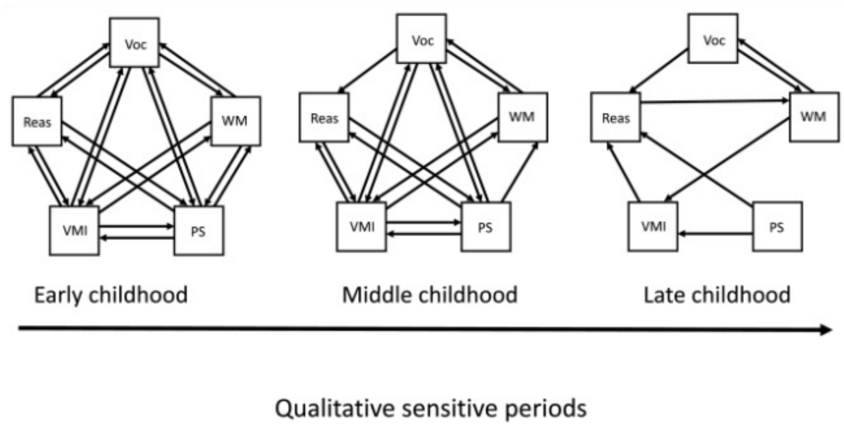


Figure 2.9: Adaptation of the mutualism model of intelligence to qualitative sensitive periods of childhood cognitive development (Figure 3 from Kievit [81]).

The abstract form of the formal theory enables adaptation to different contexts. Kievit [81] implemented the model developed by Van Der Maas et al. [174] within the context of critical times in childhood cognitive development, the *sensitive periods*. One of the proposed models is shown in Figure 2.9 and enabled Kievit [81] to recast the sensitive periods in cognitive development as times when the dynamic processes underlying cognition are changing in kind.

The extra work in moving towards a formal theory involves abstracting the prototheory into a "set of model equations that encode explanatory principals" [13], and can be challenging. The end result is a system of rules that help describe the causal structure. This structure opens up a data generating process, enabling simulation to replicate observed phenomena, which can then be used to evaluate the theory.

This approach may not be easily applicable to all theory across LA, but with the calls for stronger theory [186] adopting a process like TCM might be useful.

2.6.3 Instantiated theory creating a shared understanding

The 'formal theory' stage of the TCM (see Figure 2.7) is also referred to as a *formal model*. Formal models can be seen as a particular instantiation of a more generalised theory, narrower in scope and more applied, serving as a bridge between the theory and the real world [40]. It is critically important to distinguish these types of models, which represent phenomena, from *statistical models*, which represent data. Somewhat understandably, these two types of models are often conflated, with some 'theories' actually being statistical models [21]. In TCM formal models are used as an intermediatory step between the less structured and more descriptive prototheory and the *Mathematical Analysis* step (bottom box of Figure 2.7), which is the natural step towards the implementation of statistical models.

Borsboom et al. [13] see the construction of formal theories as a critical step moving between the contextual knowledge of phenomena and the mathematical apparatus that describe data. Robinaugh et al. [137] claim a benefit of formalisation is that the theory becomes a tool for collaboration and integration. Formalisation of theory make the theory explicit, transparent and allows development "across domains of expertise". As noted at the beginning of Section 2.6 there have been calls for stronger theory in LA, perhaps this will help context experts and data experts communicate across the divide?

2.7 Research Gap 2: There is a context / data divide in Learning Analytics

There appears to be a gap in LA between the sense-making of the context experts and that of the data experts. This can result in LA becoming isolated from their context [44, 86]. LA has attempted to address this through the use of Co-Design methods but different generative toolkits may be required to better cultivate a shared understanding between participants (Section 2.5). Responding to the call for stronger theory in LA may help close this gap, however LA is not in a position, as yet, of having developed strong theories that help translate between prototheory and mathematical and statistical models (Section 2.6).

I propose to focus on tools that may help close this gap through the construction of LA.

Research Question 2: How can we inclusively build Learning Analytics that embeds contextual knowledge?

3 Research approach

3.1 Research questions

As outlined above, I will be investigating two research questions:

RQ1: **How can Learning Analytics make rigorous causal claims?**

RQ2: **How can we inclusively build Learning Analytics that embeds contextual knowledge?**

3.2 Opportunities afforded by graphical causal models

Graphical Causal Models (GCMs) are a way of encoding the causal structure through a graph (briefly introduced in Section 2.2.3.7). This graph is often referred to as a DAG (Directed Acyclic Graph), but technically a DAG need not represent a causal model [1]. These models come with a calculus of interventions, the *do*-calculus [127], which translates a graphical structure into a statistical model that should provide an unbiased estimate of the causal effect of interest [127]. They are closely related to Structural Equations Models (SEMs), which also articulate the data generating process through a system of equations, but GCMs are more overtly focused on making causal claims than SEMs [11, 128, 129]. Despite having roots in the development of Artificial Intelligence, somewhat outside the world of inferential statistical sciences, GCMs have become increasingly popular causal inference framework across many fields due to their flexibility in illuminating and overcoming the obstacles to robust inference [104, 110, 116, 130, 141]. The flexibility of the GCM framework has allowed it to equivalently model the Potential Outcomes Framework [128], SEMs [11, 127], RD Design [158], Propensitiy Scores [158], Instrumental Variables [127, 158] and Validity Typology [110]. This means that GCMs provide a direct and general approach to causal inference.

[1] The terms DAG, GCM and SCM (Structural Causal Model) are often used interchangeably. DAG is more common, although technically insufficient, SCM is used by Pearl but GCM has been chosen here to highlight the visual component.

As noted in Section 2.2.3.7 GCMs are rarely used in LA. However, many illustrative examples from outside of LA do use educational scenarios to build GCMs [104, 130, 140]. I believe that they show the potential to help answer both research questions, and they are a good fit for the field of LA more generally. One of the criticisms, or barriers, in using GCMs more broadly is that they require input from other stakeholders to build [138]. I would claim that for LA there are some strong positives that stem from this requirement. Although moves have been made towards more context informed (Section 2.5), theory informed (Section 2.6) and human centred LA, the field has yet to find a structured way of helping context experts to encode their knowledge into our statistical models directly. GCMs could provide a mechanism for achieving this encoding of educational expertise.

The DAG component of the GCM serves as an artefact, a generative tool in a Co-Design process (Section 2.5), to help facilitate collaboration across disciplinary boundaries and so leverage a wider variety of expert knowledge, potentially inviting a greater breadth of expertise to the LA development table. DAGs are a way of reducing artificial complexity in understanding relationships between variables. For instance $A \to B \to C$ is rich in data about conditional independence between the variables A, B and C. While it is possible to state that C and B are not independent, B and A are not independent, and C is independent of A conditional on B (with notation $C \perp A|B$), and those statements do not capture the postulated *direction* of the effect. These are also more foreign mathematical concepts and symbols for a non-technical audience to interpret and challenge. The causal model and DAG seems to keep the natural complexity of the model intact, while minimising its artificial complexity.[2]

For readers unfamiliar with DAGs and GCMs I have included the Causal DAG tutorial section I wrote for the paper Hicks et al. [63] in Appendix A. Alternatively Lübke et al. [104], Rohrer [140] both have excellent introduction to Causal DAGs.

3.3 Research objectives

The aim of my research is to provide a method to make stronger causal claims in LA through Co-Design of causal models. The gaps identified above will be addressed by pursuing the following two research objectives:

1. Use Graphical Causal Models in Learning Analytics to make causal claims, and

2. Establish a methodology for Participatory Modelling with Graphical Causal Models.

[2]This paragraph is adapted from my LAK22 paper, Hicks et al. [63].

4 Preliminary results

Work completed so far can be organised into 3 parts:

1. Prototyping the participatory modelling methodology with a case study in student retention, described in Section 4.1. This was also published in Hicks et al. [63].

2. Development of participatory modelling prompts to help support the participatory modelling methodology, described in Section 4.2.2.

3. Prototyping the participatory modelling methodology with a case study in student sense of belonging, described in Section 4.3.

4.1 Case study: Student retention

In what follows I make use of the widely studied problem of student retention and support [68]. The rest of this section is adapted from a paper I led and had published at the Learning Analytics and Knowledge Conference, 2022 [63]. In the spirit of "thinking with diagrams", I will present this work in a stylised form of analyst's narrative, as though we were thinking aloud with colleagues, using the causal diagrams to clarify my thinking.

4.1.1 Causal models in the retention space

Not everyone succeeds in their academic studies, but a better understanding of what helps a student persist is important. In LA, student retention normally centres around the prediction of "at risk" students who need to be supported in some way via an intervention. While some early attempts used student dashboards [131], it has by now become common to contact students identified as at risk via a phone call that offers them support, and potentially discuss changing their study load. Finding the best way to intervene in this complex scenario is a challenge [27], it remains hard to know which lever to pull. The effectiveness of an intervention requires a model that can offer explanation as well as prediction. Since this area of LA has been extensively studied, we need to demonstrate what more could be gained through the use of a causal model.

Interestingly this space has already been approached from a causal perspective, years before the formalisation of methods involving DAGs [127] emerged. Kember proposed a causal model of student persistence [78, 79], that operationalises earlier work by Tinto on a conceptual model of student dropout [169]. Woodley et al. [188] has since pointed out that dropout is complex and comes in many forms, possibly requiring a different model for each.

The example discussed here will centre around a particular case of a student support system at Charles Sturt University (CSU) aimed at early intervention, primarily via a phone call [94]. This example was chosen as it is a central part to the work I perform at CSU. Students are identified in week 3 or 4 of the session and put on a list to be called by a student outreach team. Students on the "at-risk" list are all contacted in some form (via sms or email) and phoned multiple times, but a conversation with an outreach team member can only happen if the student answers the phone. The overall aim is to reduce failure rates in subjects, particularly due to the non-submission of assessments. As a result, 'success' may be due to one of two possible outcomes: guiding the student to existing support structures, or towards withdrawing from the subject. In the scenario to be modelled here, CSU was trying to ascertain what effect the intervention has on the academic outcome of a student identified as at risk of failure.

On ethical grounds, a randomised controlled trial was not possible, since the goal of the project was primarily to support as many students as possible; understanding the projects' effectiveness was a secondary level goal, but not the principle *raison d'être* for the project. Nonetheless, we see the potential for using two subgroups of students to make claims about the effectiveness of the project; students who received the intervention (by answering a phone call) and those who do not (they did not answer any of the calls). Membership of either group is not randomly assigned, however, so any attempt to assign a *causal* effect of the intervention on the outcome would need to also understand and account for how students are assigned to the two groups. In what follows I will illustrate how causal modelling can help us to reason about this scenario.

4.1.2 Thinking with causal models

A key insight of Hicks et al. [63] is that causal models can help us to think more clearly about a system under study. This section will attempt to demonstrate how this works, taking the form of a dialogue, with each subsection following the same structure: answer to the previous problem; *thinking* about the new graphical causal model and its implications; framing a new *problem* with the current graphical causal model that needs to be addressed.

4.1.2.1 Starting somewhere

The model construction begins simply with a hypothesis that: $Intervention \rightarrow Outcome$. In this case, $Intervention$ denotes whether or not the student had a conversation about their options with an outreach caller, while $Outcome$ denotes if the student passed all their subjects they remained enrolled in or not. We immediately encounter a problem in our modelling.

Problem 1: The intervention does not *directly* change the outcome, rather, it changes some unmeasured attributes of the student that drive their academic decision making.

4.1.2.2 What don't we know?

To address this problem, we can add a variable in between *Intervention* and *Outcome* that represents the changeable academic attributes of the student that affect their chance of success. It is these changeable attributes that we are trying to influence with our phone intervention. Let us call this variable *Mutable*, to represent that these student attributes are malleable. Our resulting enhanced model is depicted in Figure 4.1.

$$\begin{tikzcd} \text{Intervention} \arrow[r] & \text{Mutable} \arrow[r] \arrow[l, bend left, dashed] & \text{Outcome} \end{tikzcd}$$

Figure 4.1: A DAG relating an *Intervention* to an *Outcome*, via a mediator of *Mutable* student academic attributes.

Thinking: In Figure 4.1 we have started with a near minimal DAG. Although far from ideal, once drawn it can be interrogated by others, and the building of the causal model becomes a collaborative process of thinking with the model. However, in making explicit our assumptions, we quickly hit a new problem.

Problem 2: What are we not seeing? This modelling exercise began because an educational expert had voiced a suspicion that the selection of students into the *Intervention* group (i.e. those who answer the call) is not random. But if this the case, and we have wrapped these student attributes into *Mutable*, then *Mutable* influences *Intervention* as well as the other way around. We have a loop. This is not allowed in DAG (see Section A).

4.1.2.3 Making time acyclic

We can avoid loops in a DAG by carefully splitting variables into 'before' and 'after' some epoch, in this case we can split $Mutable$ into before the phone call, $Mutable_{t=0}$, and after the phone call, $Mutable_{t=1}$, an update that is now reflected in Figure 4.2.

$$\begin{tikzcd} & \text{Intervention} \arrow[d] & \\ \text{Mutable}_{t=0} \arrow[r] \arrow[ur] & \text{Mutable}_{t=1} \arrow[r] & \text{Outcome} \end{tikzcd}$$

Figure 4.2: Adding an epoch, before and after the phone call.

Thinking: The maxim that causes should precede effects is a useful device to help sharpen thinking about a causal model. However, working out exactly *where* to divide the model of data into before and after can require some finesse.

Problem 3: At the moment, because the *Mutable* variables are unmeasured, there is no way to adjust for the confounding inducted by the path $Intervention \leftarrow Mutable_{t=0} \rightarrow Mutable_{t=1} \rightarrow Outcome$. This means there is no approach, with the current model, to find the causal effect of *Intervention* on *Outcome*.

4.1.2.4 Leveraging educational theory

We have placed unknown (i.e. latent) student attributes into our model, now we add further structure to our model using the insight that there are student attributes that we cannot change. We denote these *Fixed*, and in Figure 4.3 we use them to represent any student qualities that we can measure prior to the intervention that: (i) remain fixed throughout the study session, (ii) are measurable, (iii) influence the chance of the student answering the phone, (iv) influence student outcomes through the *Mutable* attributes. The most important thing to note when augmenting a graphical causal model like this is the *lines that are not drawn*.

$$\begin{tikzcd}
\text{Fixed} \arrow[r] \arrow[d] & \text{Intervention} \arrow[d] & \\
\text{Mutable}_{t=0} \arrow[r] \arrow[ur, dashed] & \text{Mutable}_{t=1} \arrow[r] & \text{Outcome}
\end{tikzcd}$$

Figure 4.3: Adding *Fixed*, a variable of known (relatively fixed) student attributes.

In the case of Figure 4.3, we should note that there is no direct path between $Fixed$ and $Outcome$. This amounts to using the model to claim that the effect of $Fixed$ on $Outcome$ is always mediated by $Mutable$. The dashed path from $Mutable_{t=0}$ to $Intervention$ indicates some doubt over including that path. Indeed, this dashed path indicates our concern that there may be some student characteristics in *Mutable* that impacts directly upon the likelihood of a student receiving an intervention (i.e. in this case answering the phone call) that is not explained through the common cause, *Fixed*.

Thinking: We have split the variables describing our students into two parts: mutable ones that cannot be measured, but which we are trying to influence with our intervention, and a separate set of fixed, measurable variables that are likely to be useful in segmenting our student population. At this point it is not necessary to determine exactly what *Fixed* or *Mutable* are comprised of; these are place holders for everything we know about the student that we think affects the other variables in a particular way. This level of abstraction helps move forward in discussing the causal structure quickly, without getting stuck in details that someone unfamiliar with the system may not (care to) know about. However, an educational expert could work with the modeller to decide what belongs in them. *Fixed* might include variables relating to: socio-economic status, gender, and first language, whereas *Mutable* could include variables relating to: employment status, study load, grit, and work ethic.

Having worked through this exercise, we can now see how to measure the causal effect of interest, but how we do this will depend on our assumptions around the path $Mutable_{t=0} \dashrightarrow Intervention$. If this path is weak enough to be ignored, meaning we believe the association between *Mutable* and *Intervention* arises mostly from their common cause *Fixed*, then there is only one non-causal path to block: $Intervention \leftarrow Fixed \rightarrow Mutable_{t=0} \rightarrow Mutable_{t=1} \rightarrow Outcome$. This can be blocked by adjusting for *Fixed* (see Appendix A for technical details).

Problem 4: On the other hand, if we believe the path $Mutable_{t=0} \dashrightarrow Intervention$ should be included in the model then there is no way to use adjustment to block the non-causal path $Intervention \leftarrow Mutable_{t=0} \rightarrow Mutable_{t=1} \rightarrow Outcome$ as we cannot adjust for the unmeasured $Mutable$ variables. In this case there is no way to get an unbiased estimate of the effect of the *Intervention* on the academic *Outcome* of the student. This would be reflected in the data, and so we now have a testable set of claims that can be used to investigate this scenario further.

4.1.3 What was gained by this approach?

The final model provided two clear pieces of knowledge: (i) it helped us to clearly scope the assumptions required for estimating a causal effect, and (ii) it provided us with a method for estimating this effect if the path $Mutable_{t=0} \dashrightarrow Intervention$ can be ignored. This model will be developed further in order to estimate a causal effect. I will outline my proposed future work on this problem in Section 5.1.1.

The *process* of constructing the model through dialogue and diagram also provided me with insights that helped clarify my thinking about this scenario. The first was the introduction of the *Mutable* variable and thinking about what exactly is it that the intervention was trying to change. The second was a clearer understanding of *why* the system was complex — the many confounding factors are challenging to place in the model structurally and temporally. These insights were gained *prior* to looking at the data, through the careful incorporation of concepts from educational theory.

This dialogue was reconstructed after the fact, but closely followed the flow of ideas discussed over several meetings with one of the co-authors, Leonie Payne. These discussions are, in effect, the prototype for the participatory modelling methodology.

4.2 Participatory modelling

In what follows I have called the process of a data expert and context expert collaborating to construct a Graphical Causal Model *Participatory Modelling*

(PM), and the workshop sessions where this is done *Participatory Modelling sessions*.

4.2.1 Construction of graphical causal models

The goal of the sessions in the student retention case study (Section 4.1) was to transcribe the space of relevant causes in the form of a diagram. The sessions also informed my thinking about how this might be approached in a discussion with actual context experts in an authentic setting. The little advice I found about *how* to construct a causal DAG with an expert was not from the formal literature, but from online courses and tutorials [58, 115], and was quite generic in nature. This advice tends to follow the following structure [115]:

1. Begin with $X \to Y$ (X causes Y).
2. Think about other causes of the variables, and their possible connections.
3. Think about other effects of the variables, and their possible connections.
4. Think about any unmeasured confounders.

Additionally, I have incorporated components of the initial three steps of the Theory Construction Methodology (TCM) [13] cycle (see Figure 2.7 in Section 2.6.2) in subsequent sessions, detailed in Section 4.3. These steps are (i) *identifying relevant phenomena*, (ii) *formulating a prototheory*, and then (iii) *developing a formal model*. The concept of a 'prototheory', with fewer constraints, as an intermediate step between the context experts knowledge of phenomena and the formal model has been a practical framing of the problem. The Participatory Modelling sessions focus on eliciting the context knowledge experts initial understanding of the phenomena as a prototheory, followed by incremental formalisation [151] of the prototheory into a formal theory. Whilst developing a prototheory the GCM constraints are loosened; loops are allowed, many models are drawn (possibly disconnected), unobserved confounding variables can be ignored, the abductive reasoning is allowed to take precedence over the formal constraints. From here the requirements of a GCM help guide the experts in formalising the model by suggesting natural questions to ask of the graphical representation of the prototheory. For instance the GCM requirements that (i) all possible confouning variables are included, (ii) missing edges imply mutual independence of variables, and (iii) there are no cycles in the DAG, suggest the experts interrogate the prototheory with the following questions, respectively:

(i) Are there any unidentified common causes not drawn (possibly unmeasured)?

(ii) Are all edges *not drawn* justifiable?

(iii) Can the graph be restructured to eliminate any causal loops?

To help understand the combination of rhetorical and graphical moves used during a participatory modelling session I have developed a collection of Participatory Modelling prompts, outlined below in Section 4.2.2.

4.2.2 Participatory modelling prompts

In the student retention case study (Section 4.1) it was often noted often by the participants that the DAG provided a structure to 'think with', and that particular combinations of verbal and graphical moves recurred. In developing this idea further I have created a set of *Participatory Modelling prompts* (PM prompts) to help guide the conversation forward between a context expert and a data expert when co-constructing a causal model. A challenge with designing these prompts is that they have to take into account a rhetorical move and a graphical move, which are intrinsically connected. The rhetorical move must prompt and elicit, from the context expert, the justification for changes in the causal structure. The graphical move involves changing the structure of the DAG in some way, based on the dialogue between experts. To do this I have given the PM prompts four components each.

1. An initial sub-graph. This is the structure that you would see (in some part of the causal graph) that would be the focus for interrogation.
2. A question, or series of questions, for the context experts to prompt their thinking about the model and how it might need to be developed. This is in coloured *italics* in the examples below.
3. A corresponding graph, or series of graphs, that show how the GCM would change in response to the question. Dashed lines indicate a path that might be added, and dotted lines indicate a path that might be removed. (The colour of the new parts of the graph matches the text colour of the corresponding question).
4. Corresponding GCM language to help add detail for the data expert to consider when changing the graph. (This is in plain text, in parenthesis, below each question).

Components 1 and 3 articulate the graphical move but stipulating the starting structure of the DAG to be examined, and the possible changes. Component 2 provides questions to facilitate the knowledge elicitation from the context expert in order to adjust the graph. The final component is to help the data expert understand the statistical implications of the graphical move. I shall use X to denote the intervention of interest, and Y to denote the outcome of interest.

These prompts have been used in the subsequent PM sessions, outlined below in Section 4.3. I took the role of the data expert in these sessions, and used the list of the prompts to help guide my questions if I was unsure of what to ask the context expert next. This was not an effective test of how useful the prompts would be, in general, to a data expert, since I had developed them and knew each rhetorical and graphical move from memory. I will return to this point in Section 5.2 when I discuss my future work on this topic.

4.2.2.1 The simplest path

Don't know where to start? Start with the simplest possible path, connecting the *intervention*, X, with the *outcome*, Y.

What are you interested in measuring the effect / impact of?
(Understanding what X is.)
What is the outcome / result you are trying to change?
(Understanding what Y is.)

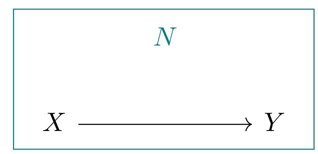
$$X \longrightarrow Y$$

4.2.2.2 Adding a node

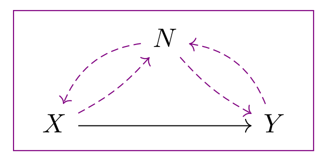
Beginning with any graph:

$$X \longrightarrow Y$$

What else might be important to include? Let's call it N
(N is the new node, or variable, or blob. It could be something in the data, or just a concept / construct that is important.)



What things might N influence? What things might influence N?
(This question can be stepped through each existing node in the graph.)



4.2.2.3 Adding causal chains

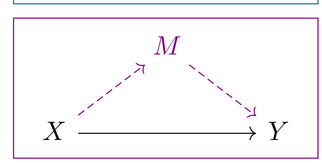
Beginning with any two causally connected nodes in the graph:

$$X \longrightarrow Y$$

Does X influence Y ***directly****, or is there some other variable (M) in between?*
(M is called a *mediator* which we are inserting into the path $X \to Y$)

$$X \longrightarrow M \longrightarrow Y$$

Are there other things (M) that X changes that in turn change Y?
(Adds a new path with a mediator in it)

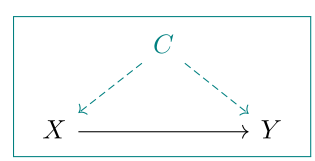


4.2.2.4 Adding a common cause

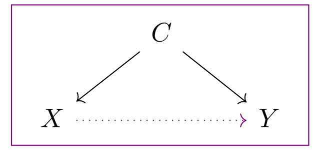
Beginning with any two causally connected nodes in the graph:

$$X \longrightarrow Y$$

Is there anything else (C) that we haven't added that might influence both X and Y?
(C is called a *confounder*.)



Could C explain all the association between X and Y?
(This is a common source of *spurious association*.)

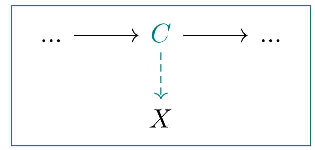


4.2.2.5 Distinguishing constructs and data

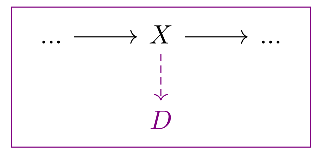
Beginning with any node in the graph:

$$\ldots \longrightarrow X \longrightarrow \ldots$$

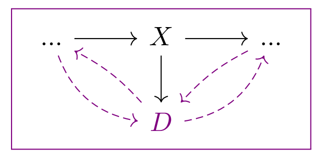
Is X the causal mechanism here, or is it just the data we have? Is X actually a proxy for some causal construct C?
(X is the data we have representing a construct, C.)



Do we have any data, D, that might represent X?
(In this case X is the construct, and D the data representing it.)



This new data D - does it influence (or is influenced by) anything else in the model?

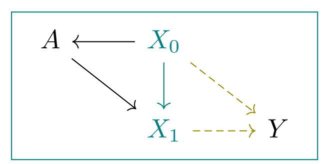


4.2.2.6 Adding an epoch

With any graph, but this works well to untangle loops:

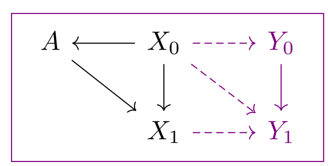
$$A \rightleftarrows X \longrightarrow Y$$

*Can we split X into **before** and **after** some event?*
(X_0 denotes before the event, at epoch $t = 0$, and X_1 after at $t = 1$)



When *do the other variables occur, relative to this event?*
(If Y is **before**, then use the path $X_0 \to Y$, but more likely it is **after** and you would use the path $X_1 \to Y$)

Do other variables also need to be split into before and after this event?
(Each of the paths connecting $X_{0,1}$ to $Y_{0,1}$ will need to be questioned.)



4.3 Case study: Belonging

Students' sense of "belonging" is becoming an increasingly relevant area of study in higher education. There is much recent literature and thought on the subject, but no formal model that encapsulates the theory [2, 19, 34]. This case study involves Co-Designing a causal model, based on the literature. It is being implemented as part of the Theory Construction Methodology (TCM) [13], which follows the cycle outlined in Figure 2.7 in Section 2.6.2.

The 'Abstraction' stage in the TCM will arise during the Co-Design process (to be outlined in the approach to RQ2, Section 5.2), developing the GCM, which will be used for the simulation.

4.3.1 Belonging model 1

Several sessions have been run with a doctoral researcher (Researcher 1) beginning to explore the literature on belonging where I aimed to prototype the construction of formal theory according to TCM (Section 2.6.2). The researcher was in the process of reading the literature on belonging, and the intent was for the causal model to synthesize this work.

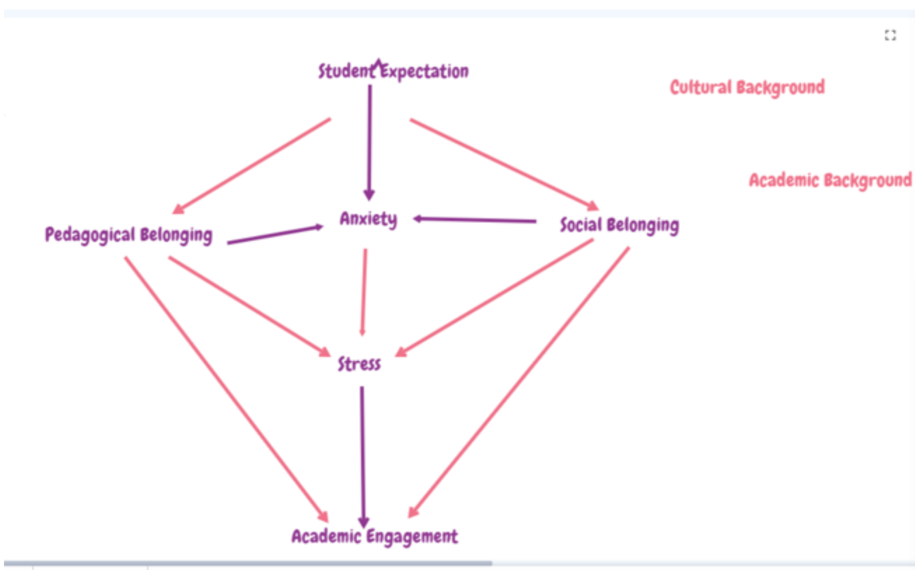


Figure 4.4: Initial DAG construction for belonging model.

This model sought to model the experience of students entering institutions from different cultural backgrounds to the cultural background of the academic institution they are studying at. Four participatory modelling sessions were run virtually via Microsoft Teams. Drawing, and sharing, the diagrams proved to be difficult in this environment. Pen and paper was trialled, with participants holding up the sketches to the camera, as well as sharing one participant's screen and utilising the web based version of DAGitty [166]. The best solution was found to be a collaborative drawing space available in Teams, which was slow but allowed for joint editing of the model.

Initially, the model started with the researcher's knowledge of the main drivers of academic engagement from social belonging and academic belonging, shown in Figure 4.4. Features such as 'cultural background' and 'academic background' of the student were deemed important by the context expert, and can be seen as unconnected nodes in the top right of Figure 4.4. As these were brought into the graphical structure by the data expert the context expert explored splitting the 'academic belonging' node into two parts, to better understand how the students sense of academic belonging at a classroom level relates to their sense of academic belonging at an institutional level. This is shown in Figure 4.5. These two new divisions of belonging then became the main focus of the DAG, with 'social belonging' getting removed. The latest DAG from this researcher is shown in Figure 4.6.

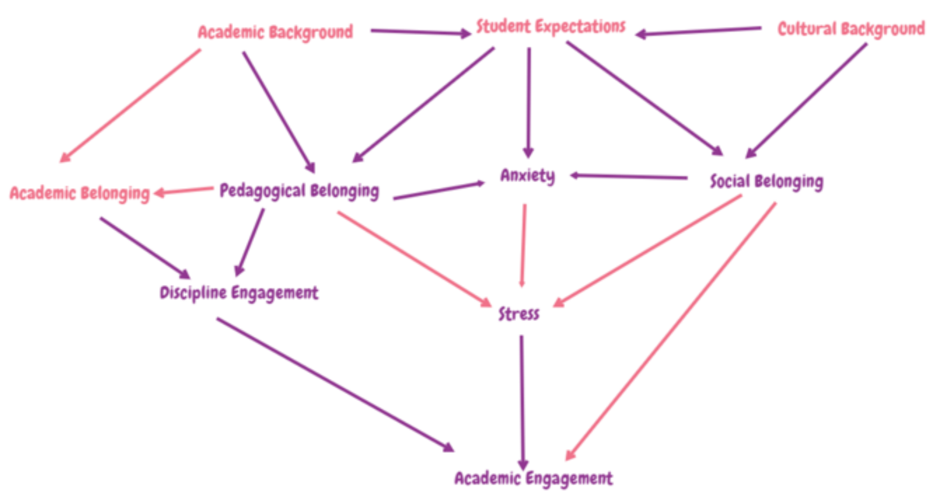


Figure 4.5: Second iteration DAG construction for belonging model.

There are two important insights, relevant to the data expert, that can be read from the DAG in Figure 4.6. Causal effects of the two kinds of belonging on engagement and academic achievement are likely to be confounded by the variables academic background, student expectations, and cultural background. This highlights the complex underlying causal structure, and why it may be difficult to understand how the student's sense of belonging interacts with the broader educational system. Secondly, an intervention that changes, say, academic belonging would be difficult to assess due to confounding. Data on all three variables in the top row would be required to get an unbiased estimate of the causal effect.

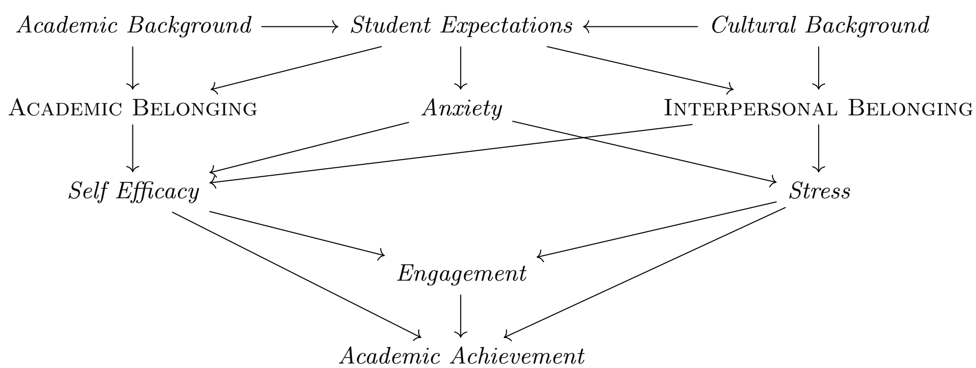


Figure 4.6: Researcher 1's model of belonging, attempting to synthesize literature.

4.3.2 Belonging model 2

To explore the potential transferrability of the resulting belong models a belonging model was then developed with a different academic (Researcher 2) who had a different lens to view belonging, this time from a more institutional perspective. This researcher was trying to understand the impact of the teaching and learning environment on belonging. A single PM session was run via Zoom, utilising the web app version of the DAGitty package to draw the causal model. The variables self efficacy and belonging (institutional level) were added by the researcher, and all initial causal connections drawn, in Figure 4.8.

At this point, the following dialogue occurred:

Modeller: *Ok, are there any extra dots (variables) that we should get onto here?*

Researcher 2: *Ah, motivation. And actually self-efficacy is an aspect of motivation.*

Modeller: *Okay, when you say self-efficacy is an aspect of motivation does that mean that the motivation helps improve self-efficacy or the other way around?*

Researcher 2: *The other way around. Because once they feel confident in their ability to perform it makes them more motivated to engage.*

This exchange is illustrative of the co-creation of a causal model process and how it prompts the participants to think more deeply and precisely about a system. Firstly, the prompt to think about 'what else' might be important is getting the researcher to think about the system. Then the modeller was able to add the extra extra variable motivation, and establish the direction of the causal connection with the variable self-efficacy, but not without further clarification. The word "aspect" needed to be unpacked, in order to correctly

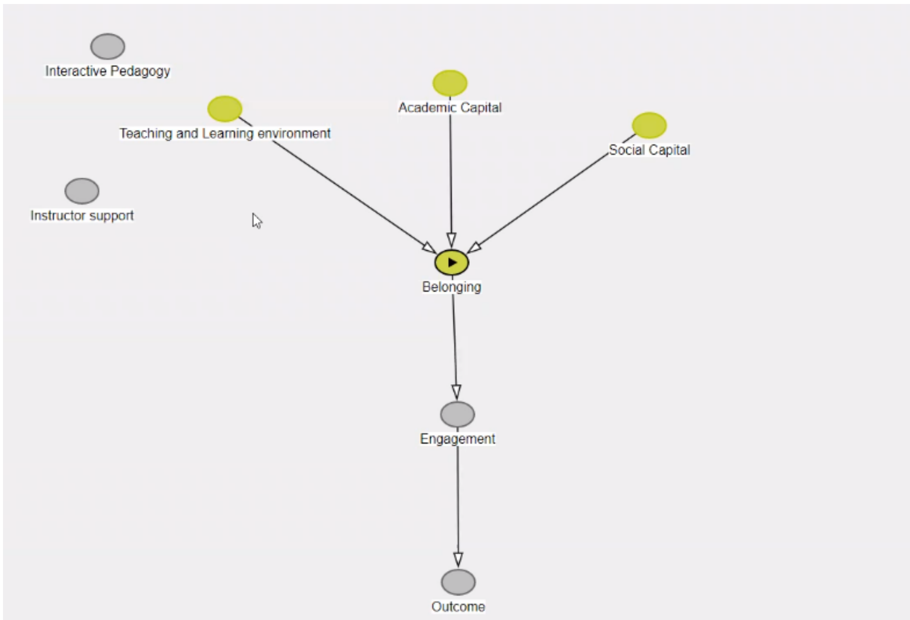


Figure 4.7: Researcher 2's starting DAG for modelling belonging.

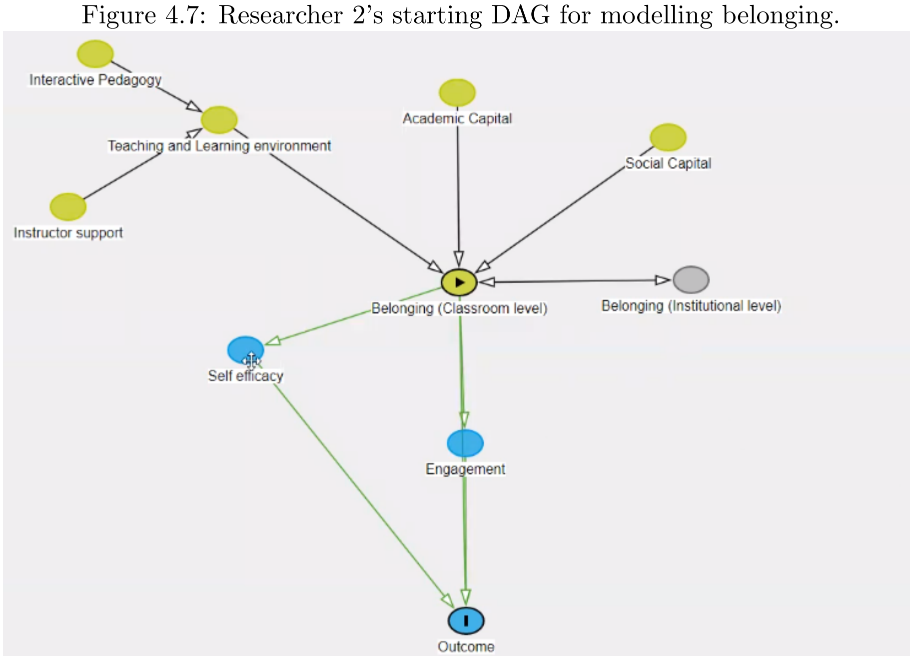


Figure 4.8: Researcher 2's DAG for modelling belonging, second stage.

establish the direction of the causal arrow. Without the goal of creating a semi-formal structure such as a DAG this clarification may not have happened, and the precision in the dialogue and the thinking occurring may have been less.

The new DAG, with motivation added and some further variables, is shown in Figure 4.9. The final model from Researcher 2 is shown in Figure 4.10.

An interesting relationship came up in Researcher 2's final model, one cannot currently be represented using a standard GCM [127]. The edge between belonging at a classroom level and belonging at a institutional level was bi-directional, with each variable influencing the other, a feature not permitted in causal DAGs. Sometimes this indicates the presence of an unmeasured confounding variable, however this was not the case, and the researcher could describe in more detail how the causal relationship worked both ways. The causal influence was described to flow 'downhill' only, depending on the level of the two variables, classroom belonging and institutional belonging. For instance, if classroom belonging was higher than institutional belonging, then the causal effect would flow from classroom belonging to institutional belonging, increasing the latter. If instead institutional belonging was higher, the process would work the other way around. Although this is a well articulated description of causal relationship between two variables it is not permitted in the existing GCM framework, which requires the flow of causation to be strictly one direction between any two variables [127].

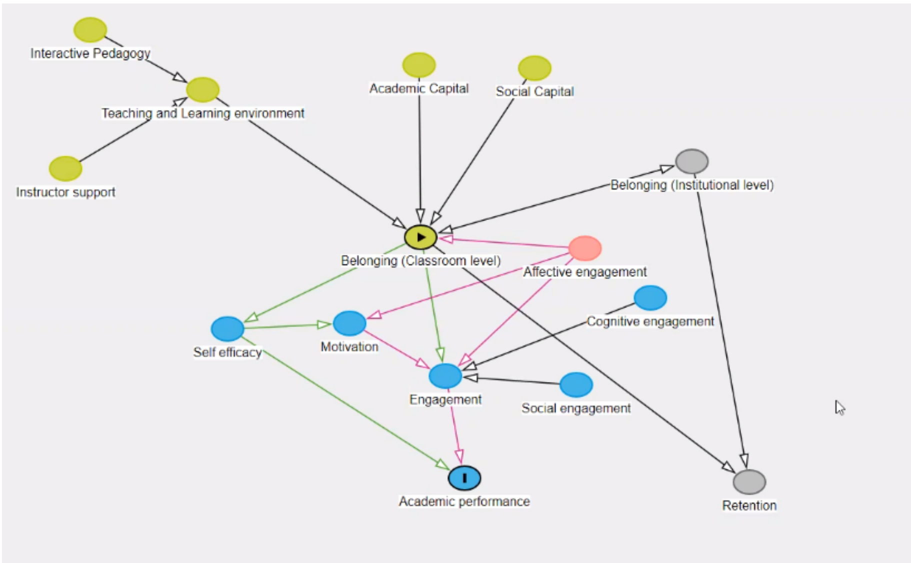


Figure 4.9: Researcher 2's DAG for modelling belonging, third stage.

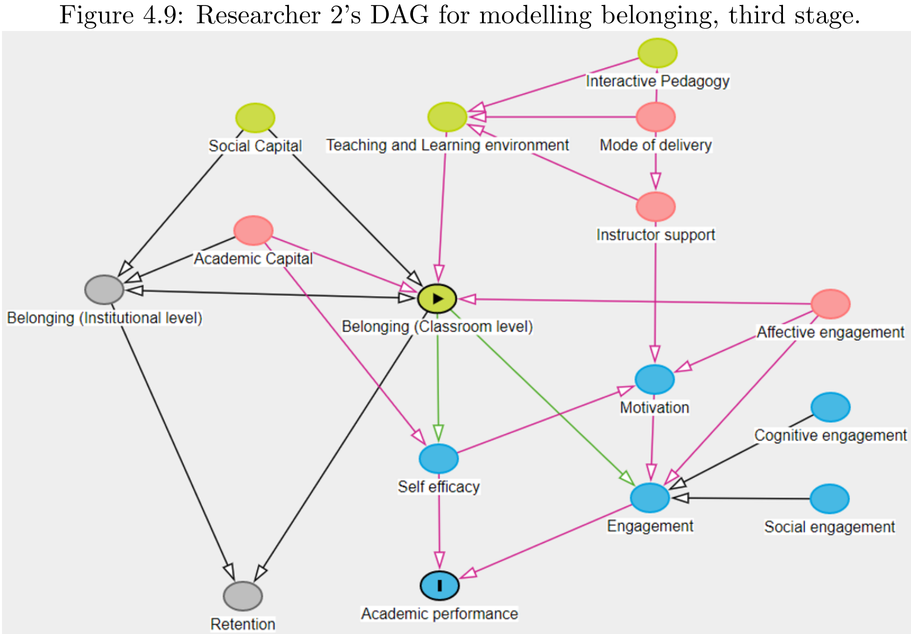


Figure 4.10: Researcher 2's final model of *Belonging*, focused on the effect of the teaching and learning environment and belonging.

5 Research plan

5.1 Approach to research question 1

How can Learning Analytics make rigorous causal claims?

A key goal of my thesis is to understand the valid use cases for causal models in learning analytics. I will explore this initially through the two case studies on student retention in Section 4.1 and belonging in Section 4.3. The two cases studies are very different in their application of the GCM. The retention model is being utilised to evaluate the causal effect of an intervention. The belonging model is being used to construct and evaluate a theory, which will use the GCM for simulation rather than estimation.

5.1.1 Evaluating an intervention

This stream of work aims to estimate the causal effect of an intervention on an outcome. Initially I will continue the work outlined in Section 4.1, the student retention case study. In this case the intervention is the outreach phone call to the at-risk student, and the outcome is their academic performance in the session.

Ethics will need to be procured for the data set to be used. Because the causal model is designed around a general student intervention scenario, it need not be the Charles Sturt University program. Any program with a similar causal structure can be used as a test case, if there are any ethical concerns with using original case study setting.

Two causal models have been proposed, and a logistic regression model will be fitted to each. This would provide causal estimates for the effect of the intervention on the academic outcome, conditioned upon the proposed causal model and the data. This effect will measure the change in probability of a successful academic outcome when the intervention occurs. Before proceeding the construct *Fixed* will need to be estimated (see Figure 4.3). As it stands *Fixed* represents characteristics that influence *Mutable* and *Intervention* in the model, and is not measured directly in the data. Other variables that serve as good proxies for *Fixed* will need to be included, and their causal relationships with the existing variables in the DAG will need to be assessed.

Covariate choice will then be informed by the graphical structure of the model and the *do*-Calculus, which will formulate a plan to estimate the causal effect [127]. Effects for the overall population as well as subgroups will be calculated, with the subgroups of interest being informed by the needs of the context experts. Modelling and analysis will be done using the R [133] statistical software language. The DAGitty [166] will be utilised to include the graphical structure in the models, and the DoWhy [150] package may also be leveraged packages to causal modelling and inference.

To evaluate the effectiveness of using the causal model, the estimates of the effects of the intervention will be compared to common approaches when no causal model is used. These will be:

1. Including no covariates in the regression model.
2. Including all seemingly relevant covariates in the regression model.

Lastly, the results will be presented back to the Charles Sturt Retention Team, or other relevant stakeholders. Of particular interest will be the comparison in overall effectiveness of the program with the effectiveness of the various subgroups, and if the model results cohere with the anecdotal evidence. Any actions that the team take, based on this evidence, will be reported.

5.1.2 Theory evaluation via simulation

This will continue the work with the belonging case study, outlined in Section 4.3. I will follow the TCM process [13] introduced in Section 2.6.2, utilising the GCMs developed in conjunction with the belonging context experts as tools for simulation. Each step of the TCM will be executed as follows:

1. *Developing a prototheory* has been conducted as part of the PM sessions with the context experts. This has been supported by the general prompts.
2. Constructing a *formal theory* through the process of abstraction was also be part of the PM sessions, and supported by the development of the PM prompts outlined in Section 4.2.2. This formal theory aims to meet the requirements of a Causal DAG. As the sessions with different academics have been run asychronously, I will develop methods of comparing and contrasting the GCMs created in each session. These will be used to support consensus building sessions with both academics to construct a final set of causal models to evaluate.
3. Simulations will be run on the model(s) to *check the explanatory adequacy* of the theory. These will be done using the R statistical software language [133], supported by the GCM similation tools from the DAGitty package [166]. This will need to be guided by the context experts in two areas. Firstly, relevant phenomena that the simulation should be able to replicate need to be identified. These phenomena then need to be formalised so that the model can parse them. One such phenomena has already been

identified as "switching to an online only learning environment". In this case the formalisation is simply switching the binary variable *Mode of Delivery* between online and in-person and comparing the results. Secondly, model parameters, i.e. the relative strength of the causal paths in the DAG, will need to be estimated. This will involve eliciting the knowledge from the experts, and possibly direct from the literature, for each causal link in the diagram. Johnson et al. [72] outline a potential method for eliciting priors that would be applicable here. Estimates do not need to be exact, and in fact the uncertainty may be just as informative as any particular point estimate.

4. *Evaluating the theory* will follow the approach suggested by Borsboom et al. [13], assessing the theory for explanatory coherence. This will include checking for explanatory breadth, analogy and simplicity. Explanatory breadth refers to the number of relevant phenomena that the theory can explain. Analogy refers to how well the theory can be described as being 'like' some other theory. For example, in the mutualism model of intelligence presented by Van Der Maas et al. [174] uses an ecological example as an analogy for their model of intelligence. Lastly, simplicity refers to how economic the explanations of phenomena by this theory, usually in comparison to other theories. This will need to be done in conjunction with the context expert. A list of test phenomena will be made and the explanation of the model then compared to any competing theories. Simplicity will be judged on the number of causes that need to invoked in order to explain the phenomena [101].

5.1.3 Self education in modelling and simulation

There are components of the methodology above that I will need to research further. To enable this I will complete self education in:

- Modelling methods appropriate for causal models in LA. For this I will be continuing the Statistical Rethinking course [116] as it has strong causal focus. The DoWhy package [150] (R / Python) has implementation of several causal inference approaches, and this may be a quicker path to applying more straight-forward techniques, where appropriate.

- Methods and coding approaches (in R) for running simulations on causal models. The DAGitty package [166] can simulate directly from a causal graph, and may be a suitable starting point.

5.1.4 Contributions

5.1.4.1 Rigorous causal claims in Learning Analytics

Making an explicit, rigorous causal claim in LA is rare, and the use of GCMs even more so (Section 2.2). Any strong claims made by the case studies will be a contribution to the field of LA. In particular, the retention causal model outlined in Section 4.1 would be a valuable contribution to LA and the student retention literature once conditioned on the data to estimate a causal effect. The student retention causal model is also generic enough to potentially be adaptable to other student intervention scenarios, providing a transferable theoretical basis for making causal claims in similar LA systems.

The DAG itself is also a readily interpretable way of representing the assumptions required to make the causal claim, allowing for a more transparent LA product that allows for criticism.

5.1.4.2 Testable formal theories

Any formal theories developed, such as the belonging model outlined in Section 4.3, will be a contribution to its field. Depending on the field, they may also be a new way to view the current state of theory in the field. A GCM of Belonging, for instance, is a new way to conceptualise the existing literature on Belonging — a publishable 'visual' literature review. The elicitation of parameters from the context experts required to run the simulations on the causal model may also identify parts of the existing theory lacking strong evidence. Even if they are shown to be 'wrong' or incongruous with the data at a later point, GCMs are reasonably easy to interpret and as such can be contested and interrogated by others, possibly leading to improvements and towards better causal models for LA.

Leveraging the relatively recent TCM framework (Section 2.6.2) in LA is new, and considering the numerous calls for stronger theory in LA (Section 2.6.1) possibly a fertile path forward for how LA develops and evaluates theory. Even if TCM is not adopted in its entirety for LA theoreticians, there are some key components that could be useful on their own, such as the use of simulation to replicate known phenomena or hypothesise new phenomena.

5.2 Approach to research question 2

How can we inclusively build Learning Analytics that embeds contextual knowledge?

Research question 2 will be approached by investigating the potential for viewing the use of GCMs as a generative tool in a Co-Design process, where stakeholders take part in the *Participatory Modelling* process introduced in Section 5.2.1.

A challenge noted in Section 2.5 to the current Co-Design generative toolkits was the imbalance in the learning required by the context experts and data experts. While data experts may be experienced in the use of a standard regression model, or something similar, constructing a GCM requires them to move from a directionless associational framing of their problem to a structured understanding of the system, where direction is important. They will also have to adjust their focus from modelling the data to modelling the causal structure. This can involve working with, possibly unmeasurable, variables that are more abstract or conceptual than what a data expert is familiar with. The context expert has to make a similar number of conceptual jumps. The first is in the reduction of their tacit knowledge of the system, in the form of their experience and their everyday understanding of the causal mechanisms, into a more explicit form as a collection of variables represented by nodes on the GCM DAG. The second is in the joining of these nodes in the diagram with 'causal' links. At this point a notion such as X 'influences' Y may be more instructive than X 'causes' Y to help the context expert think about causation as probabilistic instead of deterministic.

5.2.1 Participatory modelling methodology

To support both experts in making these conceptual jumps I will develop a framework for the Participatory Modelling sessions, based on analysis of the sessions conducted so far and future sessions. The framework developed here will include:

1. Minimal instruction materials to help non-technical experts know enough to take part in the design process.
2. Participatory modelling prompts (see existing work in Section 4.2.2) that provide questions to assist knowledge elicitation matched with graphical DAG transformations.
3. Evidence through qualitative analysis of the utility of this process in enhancing thinking about the system.

This will build upon the work already completed for the case studies (Section 4.1 and Section 4.3), as well as the Participatory Modelling prompts (Section 4.2.2). The PM sessions will be conducted in a semi-structured way and recorded in order to develop the process. Recording via Zoom will allow for auto-transcription, and the transcription will be cleaned and loaded into NVivo. Qualitative analysis of the case study working sessions will be performed in NVivo, looking for themes around a change in thinking, trust and confidence.

Ethics approval will need to be obtained for the Participatory Modelling Co-Design sessions, in order to record and analyse the transcripts. Prior to setting up the protocols for the sessions, and analysing the data, I will study the following:

- Co-Design methods from LA that may be applicable to the participatory modelling methodology.
- Qualitative analysis methods for coding the participatory modelling sessions, including working with NVivo to analyse the transcripts.

The Participatory Modelling sessions will hopefully show, through the qualitative analysis of transcripts, a deepening understanding of the system that the context expert is trying to understand, showing more nuanced and thoughtful questions as the session progresses. To assess this I will need to develop an evaluation framework to judge this. I will also attempt to transfer and apply the GCMs developed in the PM sessions to other contexts. If the modelling process was a success then the causal structures developed should be generalisable to some degree.

5.2.2 Contributions

5.2.2.1 A participatory modelling methodology

Although there has been work in the Co-Design and Participatory Design space in LA, this is the first to my knowledge that is aimed at building theories of the world; that is mapping out the causal structure of how the experts think the target system works. This is a more abstract generative tool than other Co-Design tools, but this may also allow the GCMs to be more transferable.

The PM methodology that builds the causal model to make these claims will form part of a methodology for moving from theory, to a diagram representing the theory, to the statistical machinery to make a causal claim. As such it will hopefully be a new and accessible way to embed theory in LA.

Additionally I will be developing an evaluation framework to assess the improvement in thinking and understanding that occurs in the PM sessions. This will be a contribution as well, particularly as it will need to judge different types of simultaneous reasoning at once; verbal, graphical, and causal.

5.2.2.2 Analysis of opportunities and limits of Participatory Modelling

The qualitative analysis of the PM sessions should provide some useful insights. I have hypothesized that their will be 'mutual learning' from both the context experts and the data experts as they attempt to meet somewhere in a conceptual middle ground with the use of DAGs. Analysis of the PM transcripts may show how this occurs, and offer insights into what are the limits to the shared understanding developed by the two types of experts, and if they share similar challenges in finding an epistemic common ground. Understanding what is of value the PM methodology adds to the collaborative process in order for its components to be refined, adopted or discarded in the future will be an interesting contribution.

5.3 Publicity

I plan to turn each of the case studies, if appropriate, into a paper. The LAK conference or Journal of Learning Analytics would be my initial targets. To help with the application of causal modelling I intend to run workshops for the LA community on introductory graphical causal modelling. Initially I will target LAK and ALASI conferences for this.

I believe the LA field needs to address causality head on, so as I (hopefully) become more confident in the value of these methods for the field I will publish and blog materials to help promote and use graphical causal models in LA.

5.4 Timeline

My planned schedule for complete this research is outlined in Figure 5.1.

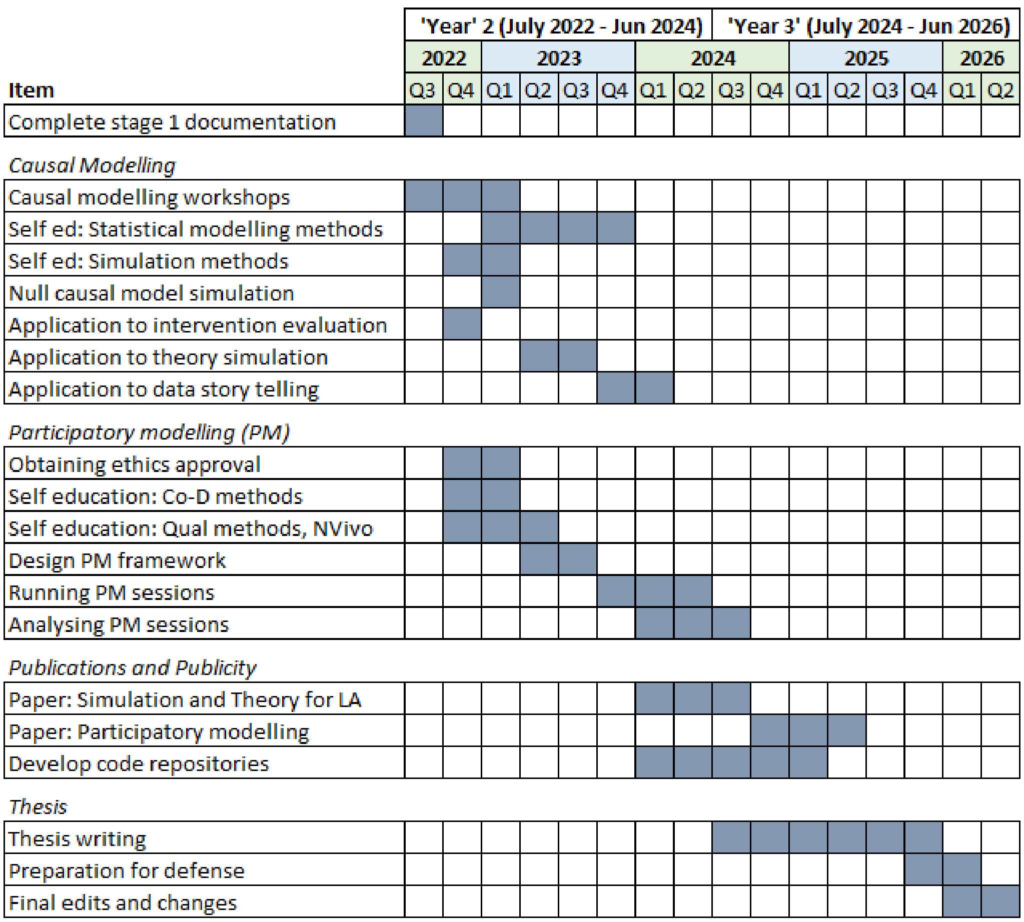


Figure 5.1: Final 'two' years schedule of work (over 4 years, as PhD is part-time).

Bibliography

[1] Carlos Prieto Alvarez, Roberto Martinez-Maldonado, and Buckingham Shum. La-deck: A card-based learning analytics co-design tool. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 63–72, 2020.

[2] Chi Baik, Ryan Naylor, Sophie Arkoudis, and Anna Dabrowski. Examining the experiences of first-year students with low tertiary admission scores in australian universities. *Studies in Higher Education*, 44(3):526–538, 2019.

[3] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

[4] Seyyed Kazem Banihashem, Khadijeh Aliabadi, Saeid Pourroostaei Ardakani, Ali Delaver, and Mohammadreza Nili Ahmadabadi. Learning analytics: A systematic literature review. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 9(2), 2018.

[5] Sanam Shirazi Beheshitha, Marek Hatala, Dragan Gašević, and Srećko Joksimović. The role of achievement goal orientations when studying effect of learning analytics visualizations. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 54–63, 2016.

[6] Louis-Philippe Beland and Dongwoo Kim. The effect of high school shootings on schools and student performance. *Educational Evaluation and Policy Analysis*, 38(1):113–126, 2016.

[7] Madhumita Bhattacharya and Steven Coombs. How to enable student formative learning and assessment: Formalytics for learning enhancement. In *Proceedings of Smart Learning Conference, Innovation Arabia*, volume 10, pages 58–66, 2017.

[8] Robert Bodily and Katrien Verbert. Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, 10(4):405–418, 2017.

[9] Robert Bodily and Katrien Verbert. Review of Research on Student-Facing Learning Analytics Dashboards and Educational Recommender Systems. *IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES*, 10(4):405–418, OCT-DEC 2017. ISSN 1939-1382. doi: {10.1109/TLT.2017.2740172}.

[10] Benjamin CM Boerebach, Kiki MJMH Lombarts, Albert JJ Scherpbier, and Onyebuchi A Arah. The teacher, the physician and the person: exploring causal connections between teaching performance and role model types using directed acyclic graphs. *PloS one*, 8(7):e69449, 2013.

[11] Kenneth A Bollen and Judea Pearl. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer, 2013.

[12] Denny Borsboom. Theoretical amnesia. *Open Science Collaboration Blog*, 2013.

[13] Denny Borsboom, Han LJ van der Maas, Jonas Dalege, Rogier A Kievit, and Brian D Haig. Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, 16(4):756–766, 2021.

[14] Nicholas A Bowman, Julie J Park, and Nida Denson. Student involvement in ethnic student organizations: Examining civic outcomes 6 years after graduation. *Research in Higher Education*, 56(2):127–145, 2015.

[15] Guy A Boysen, Timothy J Kelly, Holly N Raesly, and Robert W Casner. The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education*, 39(6):641–656, 2014.

[16] Robyn Brunton, Jasmine MacDonald, Nicole Sugden, and Ben Hicks. Discussion forums: A misnomer? examining lurkers, engagement and academic achievement. *Australasian Journal of Educational Technology*, pages 37–54, 2022.

[17] Simon Buckingham Shum, Rebecca Ferguson, and Roberto Martinez-Maldonado. Human-centred learning analytics. *Journal of Learning Analytics*, 6(2):1–9, 2019.

[18] Paulo F Carvalho, David W Braithwaite, Joshua R de Leeuw, Benjamin A Motz, and Robert L Goldstone. An in vivo study of self-regulated study sequencing in introductory psychology courses. *PloS one*, 11(3):e0152115, 2016.

[19] Darnell Cole, Christopher B Newman, and Liane I Hypolite. Sense of belonging and mattering among two cohorts of first-year students participating in a comprehensive college transition program. *American Behavioral Scientist*, 64(3):276–297, 2020.

[20] P Connolly. The trials of evidence-based practice in education. *Research Intelligence*, 2015.

[21] Andrew RA Conway, Kristof Kovacs, Han Hao, Sara A Goring, and Christopher Schmank. The struggle is real: Challenges and solutions in theory building. *Psychological Inquiry*, 31(4):302–309, 2020.

[22] José M Cordero, Victor Cristobal, and Daniel Santín. Causal inference on education policies: A survey of empirical studies using pisa, timss and pirls. *Journal of Economic Surveys*, 32(3):878–915, 2018.

[23] Ilja Cornelisz, Rolf van der Velden, Inge de Wolf, and Chris van Klaveren. The consequences of academic dismissal for academic success. *Studies in Higher Education*, 45(11):2175–2189, 2020.

[24] Peter Craig, Cyrus Cooper, David Gunnell, Sally Haw, Kenny Lawson, Sally Macintyre, David Ogilvie, Mark Petticrew, Barney Reeves, Matt Sutton, et al. Using natural experiments to evaluate population health interventions: new medical research council guidance. *J Epidemiol Community Health*, 66(12):1182–1186, 2012.

[25] Brent Davis and Dennis Sumara. 'if things were simple...': complexity in education. *Journal of Evaluation in Clinical Practice*, 16(4):856–860, 2010.

[26] Dan Davis, René F Kizilcec, Claudia Hauff, and Geert-Jan Houben. The half-life of mooc knowledge: a randomized trial evaluating knowledge retention and retrieval practice in moocs. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 1–10, 2018.

[27] Shane Dawson, Jelena Jovanovic, Dragan Gašević, and Abelardo Pardo. From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 474–478, 2017.

[28] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science Medicine*, 210:2–21, 2018. ISSN 0277-9536. doi: https://doi.org/10.1016/j.socscimed.2017.12.005. URL https://www.sciencedirect.com/science/article/pii/S0277953617307359. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.

[29] Mollie Dollinger and Jason M Lodge. Co-creation strategies for learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 97–101, 2018.

[30] Andrew C Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. Regression discontinuity designs based on population thresholds:

Pitfalls and solutions. *American Journal of Political Science*, 62(1):210–229, 2018.

[31] Niklas Elert, Fredrik W Andersson, and Karl Wennberg. The impact of entrepreneurship education in high school on long-term entrepreneurial performance. *Journal of Economic Behavior & Organization*, 111:209–223, 2015.

[32] Markus I Eronen and Laura F Bringmann. The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4):779–788, 2021.

[33] Fulya Ersoy. Returns to effort: experimental evidence from an online language platform. *Experimental Economics*, 24(3):1047–1073, 2021.

[34] Peter Felten and Leo M Lambert. *Relationship-rich education: How human connections drive success in college.* JHU Press, 2020.

[35] Rebecca Ferguson and Doug Clow. Where is the evidence? a call to action for learning analytics. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 56–65, 2017.

[36] Ed Fincham, Alexander Whitelock-Wainwright, Vitomir Kovanovic, Srecko Joksimovic, Jan-Paul van Staalduinen, and Dragan Gasevic. Counting Clicks is Not Enough: Validating a Theorized Model of Engagement in Learning Analytics. In Azcona, D and Chung, R, editor, *PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS & KNOWLEDGE (LAK'19)*, pages 501–510. Soc Learning Analyt Res; ACM SIGCHI; ACM SIGWEB, 2019. ISBN 978-1-4503-6256-6. doi: {10.1145/3303772.3303775}. 9th International Conference on Learning Analytics and Knowledge (LAK), Arizona State Univ, Tempe, AZ, MAR 04-08, 2019.

[37] Christian Fischer, Rachel Baker, Qiujie Li, Gabe Avakian Orona, and Mark Warschauer. Increasing success in higher education: The relationships of online course taking with college completion and time-to-degree. *Educational Evaluation and Policy Analysis*, 44(3):355–379, 2022.

[38] Anders Fredriksson and Gustavo Magalhães de Oliveira. Impact evaluation using difference-in-differences. *RAUSP Management Journal*, 54: 519–532, 2019.

[39] Eiko I Fried. Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4):271–288, 2020.

[40] Eiko I Fried. Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry*, 31(4):336–344, 2020.

[41] J. Dmitri Gallow. The Metaphysics of Causation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[42] Dragan Gašević, Negin Mirriahi, and Shane Dawson. Analytics of the effects of video use and instruction to support reflective learning. In *Proceedings of the fourth international conference on learning analytics and Knowledge*, pages 123–132, 2014.

[43] Dragan Gašević, Shane Dawson, and George Siemens. Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71, 2015.

[44] Dragan Gašević, Shane Dawson, Tim Rogers, and Danijela Gasevic. 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, 2016.

[45] Dragan Gasevic, Yi-Shan Tsai, Shane Dawson, and Abelardo Pardo. How do we start? an approach to learning analytics adoption in higher education. *The International Journal of Information and Learning Technology*, 36(4):342–353, 2019.

[46] Andrew P Gibson. *Reflective writing analytics and transepistemic abduction*. PhD thesis, Queensland University of Technology, 2017. URL https://eprints.qut.edu.au/106952/.

[47] Lisa Gitelman. *Raw data is an oxymoron*. MIT press, 2013.

[48] M Maria Glymour. Natural experiments and instrumental variable analyses in social epidemiology. *Methods in social epidemiology*, 1:429, 2006.

[49] Linda S Gottfredson. The general intelligence factor, 1998.

[50] Michael A Gottfried and Jay Stratte Plasman. Linking the timing of career and technical education coursetaking with high school dropout and college-going behavior. *American Educational Research Journal*, 55(2): 325–361, 2018.

[51] Anne-Britt Gran, Peter Booth, and Taina Bucher. To be or not to be algorithm aware: a question of a new digital divide? *Information, Communication & Society*, 24(12):1779–1796, 2021.

[52] Jeff Grann and Deborah Bushway. Competency map: Visualizing student learning to promote student success. In *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 168–172, 2014.

[53] Jim Greer and Mary Mark. Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education*, 26(1):387–392, 2016.

[54] Michael P Grosz, Julia M Rohrer, and Felix Thoemmes. The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5):1243–1255, 2020.

[55] Maureen A. Guarcello, Richard A. Levine, Joshua Beemer, James P. Frazee, Mark A. Laumakis, and Stephen A. Schellenberg. Balancing Student Success: Assessing Supplemental Instruction Through Coarsened Exact Matching. *TECHNOLOGY KNOWLEDGE AND LEARNING*, 22(3, SI):335–352, OCT 2017. ISSN 2211-1662. doi: {10.1007/s10758-017-9317-0}.

[56] Shenyang Guo, Mark Fraser, and Qi Chen. Propensity score analysis: recent debate and discussion. *Journal of the Society for Social Work and Research*, 11(3):463–482, 2020.

[57] Reid Hastie. Causal thinking in judgments. *The Wiley Blackwell handbook of judgment and decision making*, 2:590–628, 2015.

[58] Miguel A Hernán. Causal diagrams: Draw your assumptions before your conclusions. https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your. Accessed: 2021-03-28.

[59] Miguel A Hernán. The c-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5):616–619, 2018.

[60] Miguel A Hernán and James M Robins. Causal inference: what if, 2020.

[61] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

[62] Ben Hicks, Kelly Linden, and Neil Van Der Ploeg. Opportunities to improve learning analytics for student support when using online assessment tools. In *Back to the Future–ASCILITE '21*, pages 60–64. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021.

[63] Ben Hicks, Kirsty Kitto, Leonie Payne, and Simon Buckingham Shum. Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 250–259, 2022.

[64] R Hinton and M Robinson. Assessing the strength of evidence in the education sector. *London, England: United Kingdom Department for International Development*, 2015.

[65] Christopher Hitchcock. Three concepts of causation. *Philosophy Compass*, 2(3):508–516, 2007.

[66] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Co-designing a real-time classroom orchestration tool to support teacher–ai complementarity. *Journal of Learning Analytics*, 6(2), 2019.

[67] Jordana Hunter. Improving student learning in australia. 2022.

[68] Dirk Ifenthaler and Jane Yin-Kim Yau. Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, 68(4):1961–1990, 2020.

[69] Juliane Jarke and Andreas Breiter. The datafication of education. *Learning, Media and Technology*, 44(1):1–6, 2019.

[70] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[71] Ioana Jivet, Maren Scheffel, Hendrik Drachsler, and Marcus Specht. Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In *European conference on technology enhanced learning*, pages 82–96. Springer, 2017.

[72] Sindhu R Johnson, George A Tomlinson, Gillian A Hawker, John T Granton, Haddas A Grosbein, and Brian M Feldman. A valid and reliable belief elicitation method for bayesian priors. *Journal of clinical epidemiology*, 63(4):370–383, 2010.

[73] Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Ines Friss De Kereki. Translating network position into performance: importance of centrality in different network configurations. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 314–323, 2016.

[74] Rasmus Leth Jørnø and Karsten Gynther. What constitutes an 'actionable insight'in learning analytics? *Journal of Learning Analytics*, 5(3):198–221, 2018.

[75] Marzieh Karimi-Haghighi, Carlos Castillo, and Davinia Hernández-Leo. A causal inference study on the effects of first year workload on the dropout rate of undergraduates. In *International Conference on Artificial Intelligence in Education*, pages 15–27. Springer, 2022.

[76] Christian Kayembe and Danielle Nel. Challenges and opportunities for education in the fourth industrial revolution. *African Journal of Public Affairs*, 11(3):79–94, 2019.

[77] Mickey Keenan and Karola Dillenburger. When all you have is a hammer ...: Rcts and hegemony in science. *Research in Autism Spectrum Disorders*, 5(1):1–13, 2011. ISSN 1750-9467. doi: https://doi.org/10.1016/j.rasd.2010.02.003. URL https://www.sciencedirect.com/science/article/pii/S1750946710000218.

[78] David Kember. *Open learning courses for adults: A model of student progress*. Educational Technology, 1995.

[79] David Kember, David Murphy, Irene Siaw, and KS Yuen. Towards a causal model of student progress in distance education: Research in hong kong. *American Journal of Distance Education*, 5(2):3–15, 1991.

[80] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. Ripple: a crowd-sourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, 6(3):91–105, 2019.

[81] Rogier A Kievit. Sensitive periods in cognitive development: a mutualistic perspective. *Current Opinion in Behavioral Sciences*, 36:144–149, 2020. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2020.10.007. URL https://www.sciencedirect.com/science/article/pii/S2352154620301522. Sensitive and critical periods.

[82] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

[83] Kirsty Kitto, Simon Buckingham Shum, and Andrew Gibson. Embracing imperfection in learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 451–460, 2018.

[84] Kirsty Kitto, Kalervo Gulson, Greg Thompson, and Leonie Payne. Technical democracy in educational decision making. In *Fairness, Accountability, and Transparency in Educational Data (FATED): A 1-day workshop co-located with the EDM 2020 conference*, 2020.

[85] René F Kizilcec, Justin Reich, Michael Yeomans, Christoph Dann, Emma Brunskill, Glenn Lopez, Selen Turkay, Joseph Jay Williams, and Dustin Tingley. Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, 117(26):14900–14905, 2020.

[86] Simon Knight and Simon Buckingham Shum. Theory and learning analytics. *Handbook of learning analytics*, pages 17–22, 2017.

[87] Simon Knight, Simon Buckingham Shum, and Karen Littleton. Epistemology, assessment, pedagogy: where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2):23–47, 2014.

[88] Simon Knight, Andrew Gibson, and Antonette Shibani. Implementing learning analytics for learning impact: Taking tools to task. *The Internet and Higher Education*, 45:100729, 2020.

[89] Magia Krause. Undergraduates in the archives: Using an assessment rubric to measure learning. *The American Archivist*, 73(2):507–534, 2010.

[90] Anders Larrabee Sønderlund, Emily Hughes, and Joanne Smith. The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, 50(5):2594–2618, 2019.

[91] Walter L Leite, Zeyuan Jing, Huan Kuang, Dongho Kim, and A Corinne Huggins-Manley. Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (6):964–982, 2021.

[92] Lisa Lim, Shane Dawson, Srecko Joksimovic, and Dragan Gašević. Exploring students' sensemaking of learning analytics dashboards: Does frame of reference make a difference? In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 250–259, 2019.

[93] Lisa-Angelique Lim, Sheridan Gentili, Abelardo Pardo, Vitomir Kovanović, Alexander Whitelock-Wainwright, Dragan Gašević, and Shane Dawson. 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, 2021.

[94] Kelly Linden. Improving student retention by providing targeted support to university students who do not submit an early assessment item. *Student Success*, 12(3), Nov. 2021. doi: 10.5204/ssj.2152. URL https://studentsuccessjournal.org/article/view/2152.

[95] Kelly Linden, Neil van der Ploeg, Ben Hicks, and Prue Gonzalez. Peering into the crystal ball of the disengaged: What happens to students that do not submit an early assessment item? In *37th International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education: ASCILITE 2020*, pages 48–53. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2020.

[96] Kelly Linden, Neil Van Der Ploeg, and Ben Hicks. Ghostbusters: Using learning analytics and early assessment design to identify and support ghost students. In *Back to the Future–ASCILITE '21*, pages 54–59. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021.

[97] Kelly Linden, Neil Van Der Ploeg, Ben Hicks, and Kirsten Locke. Increasing student grades in large online subjects: Combining tutorial support with technology. In *Back to the Future–ASCILITE '21*, pages 208–212. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2021.

[98] Kelly Linden, Ben Hicks, and Sarah Teakel. Weathering the storm: Targeted and timely analytics to support disengaged students. *EDUCAUSE Review*, 2022.

[99] Ren Liu and Kenneth Koedinger. Going Beyond Better Data Prediction to Create Explanatory Models of Educational Data. In Charles Lang, George Siemens, Alyssa Friend Wise, and Dragan Gaševic, editors, *The Handbook of Learning Analytics*, pages 69–76. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition, 2017. ISBN 978-0-9952408-0-3. URL http://solaresearch.org/hla-17/hla17-chapter1.

[100] Xu Liu. Primary science curriculum student acceptance of blended learning: structural equation modeling and visual analytics. *Journal of Computers in Education*, 9(3):351–377, 2022.

[101] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257, 2007.

[102] Daniel Loton, Cameron Stein, Philip Parker, and Mary Weaven. Introducing block mode to first-year university students: a natural experiment on satisfaction and performance. *Studies in Higher Education*, 47(6):1097–1120, 2022.

[103] A.H. Lourie. *More Than Life Itself: A Synthetic Continuation in Relation Biology*. De Gruyter, 2009. ISBN 9783110321616.

[104] Karsten Lübke, Matthias Gehrke, Jörg Horst, and Gero Szepannek. Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, 28(2):133–139, 2020.

[105] Leah P Macfadyen and Shane Dawson. Numbers are not enough. why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, 15(3):149–163, 2012.

[106] Joyce B Main and Amanda L Griffith. From signals to success? the effects of an online advising system on course grades. *Education Economics*, 27 (6):608–623, 2019.

[107] Edwin P Martens, Wiebe R Pestman, Anthonius de Boer, Svetlana V Belitser, and Olaf H Klungel. Instrumental variables: application and limitations. *Epidemiology*, pages 260–267, 2006.

[108] Roberto Martinez-Maldonado, Doug Elliott, Carmen Axisa, Tamara Power, Vanessa Echeverria, and Simon Buckingham Shum. Designing translucent learning analytics with teachers: An elicitation process. *Interactive Learning Environments*, 30(6):1077–1091, 2022.

[109] Zahia Marzouk, Mladen Rakovic, Amna Liaqat, Jovita Vytasek, Donya Samadi, Jason Stewart-Alonso, Ilana Ram, Sonya Woloshen, Philip H Winne, and John C Nesbit. What if learning analytics were based on learning science? *Australasian Journal of Educational Technology*, 32(6), 2016.

[110] Ellicott C Matthay and M Maria Glymour. A graphical catalog of threats to validity: linking social science with epidemiology. *Epidemiology (Cambridge, Mass.)*, 31(3):376, 2020.

[111] Manolis Mavrikis and Wayne Holmes. Intelligent learning environments: Design, usage and analytics for future schools. *Shaping future schools with digital technology*, pages 57–73, 2019.

[112] Louise Maynard-Atem and Ben Ludford. The rise of the data translator. *Impact*, 2020(1):12–14, 2020.

[113] Tony McAleavy, Anna Riggall, and Ruth Naylor. The role of evidence in the improvement of school systems. *Education Development Trust*, 2021.

[114] Brian P McCall and Rob M Bielby. Regression discontinuity design: Recent developments and a guide to practice for researchers in higher education. *Higher education: Handbook of theory and research*, pages 249–290, 2012.

[115] Richard McElreath. Statistical rethinking 2022 lecture 20 - horoscopes. https://www.youtube.com/watch?v=qV3eos7maGs&t=1192s&ab_channel=RichardMcElreath. Accessed: 2022-08-31.

[116] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[117] Richard McElreath. Regression, fire, and dangerous things (1/3), Jun 2021. URL https://elevanth.org/blog/2021/06/15/regression-fire-and-dangerous-things-1-3/.

[118] Gerald Midgley. Systemic intervention. In *Systemic intervention*, pages 113–133. Springer, 2000.

[119] Shirin Mojarad, Alfred Essa, Shahin Mojarad, and Ryan S Baker. Studying adaptive learning efficacy using propensity score matching. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18)*, pages 5–9, 2018.

[120] Benjamin A Motz, Paulo F Carvalho, Joshua R de Leeuw, and Robert L Goldstone. Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, 5(2):47–59, 2018.

[121] Benjamin A. Motz, Paulo F. Carvalho, Joshua R. de Leeuw, and Robert L. Goldstone. Embedding Experiments: Staking Causal Inference in Authentic Educational Contexts. *JOURNAL OF LEARNING ANALYTICS*, 5 (2):47–59, 2018. doi: {10.18608/jla.2018.52.4}.

[122] Tommy Mullaney and Justin Reich. Staggered versus all-at-once content release in massive open online courses: Evaluating a natural experiment. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 185–194, 2015.

[123] Klaus Oberauer and Stephan Lewandowsky. Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, 26(5):1596–1618, 2019.

[124] Diego Olaya, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, and Wouter Verbeke. Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, 134:113320, 2020.

[125] Ryan D Padgett, Jennifer R Keup, and Ernest T Pascarella. The impact of first-year seminars on college students' life-long learning orientations. *Journal of Student Affairs Research and Practice*, 50(2):133–151, 2013.

[126] Frank Pasquale. *The black box society*. Harvard University Press, 2015.

[127] Judea Pearl. *Causality*. Cambridge University press, 2009.

[128] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[129] Judea Pearl. The causal foundations of structural equation modeling. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE, 2012.

[130] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[131] Matthew D Pistilli, Kimberly Arnold, and Matt Bethune. Signals: Using academic analytics to promote student success. *EDUCAUSE Review Online*, 10(8), 2012.

[132] Ben Pogodzinski, Peter Youngs, Kenneth A Frank, and Dale Belman. Administrative climate and novices' intent to remain teaching. *The Elementary School Journal*, 113(2):252–275, 2012.

[133] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

[134] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

[135] Teodoro F Revano and Manuel B Garcia. Designing human-centered learning analytics dashboard for higher education using a participatory design approach. In *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5. IEEE, 2021.

[136] Bart Rienties, Simon Cross, and Zdenek Zdrahal. Implementing a learning analytics intervention and evaluation framework: What works? In *Big data and learning analytics in higher education*, pages 147–166. Springer, 2017.

[137] Donald J Robinaugh, Jonas MB Haslbeck, Oisín Ryan, Eiko I Fried, and Lourens J Waldorp. Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, 16(4):725–743, 2021.

[138] James M Robins and Larry Wasserman. On the impossibility of inferring causation from association without background knowledge. *Computation, causation, and discovery*, 1999:305–21, 1999.

[139] Henry L Roediger III, Pooja K Agarwal, Mark A McDaniel, and Kathleen B McDermott. Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17(4):382, 2011.

[140] Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42, 2018.

[141] Julia M Rohrer and Ruben C Arslan. Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, 4(2):25152459211007368, 2021.

[142] Carolyn Penstein Rosé. Monolith, multiplicity, or multivocality: What do we stand for and where do we go from here?. *Journal of Learning Analytics*, 6(3):31–34, 2019.

[143] Roderick A Rose. Frameworks for credible causal inference in observational studies of family violence. *Journal of family violence*, 34(8):697–710, 2019.

[144] Robert Rosen. *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. Columbia University Press, 1991. ISBN 978-0-231-07565-7.

[145] Elizabeth B-N Sanders and Pieter Jan Stappers. Probes, toolkits and prototypes: three approaches to making in codesigning. *CoDesign*, 10(1): 5–14, 2014.

[146] Juan Pablo Sarmiento and Alyssa Friend Wise. Participatory and co-design of learning analytics: An initial review of the literature. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 535–541, 2022.

[147] Juan Pablo Sarmiento, Fabio Campos, and Alyssa Wise. Engaging students as co-designers of learning analytics. In *Companion proceedings of the 10th international learning analytics and knowledge conference*, pages 29–32. SoLAR Frankfurt, 2020.

[148] Nigel R Shadbolt, Paul R Smart, J Wilson, and S Sharples. Knowledge elicitation. *Evaluation of human work*, pages 163–200, 2015.

[149] William R Shadish. Randomized controlled studies and alternative designs in outcome studies: Challenges and opportunities. *Research on Social Work Practice*, 21(6):636–643, 2011.

[150] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[151] Frank M Shipman III and Raymond McCall. Supporting knowledge-base evolution with incremental formalization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 285–291, 1994.

[152] Ginger V Shultz, Amy C Gottfried, and Grace A Winschel. Impact of general chemistry on student achievement and progression to subsequent chemistry courses: A regression discontinuity analysis. *Journal of Chemical Education*, 92(9):1449–1455, 2015.

[153] Simon Buckingham Shum. Embedding learning analytics in a university: Boardroom, staff room, server room, classroom.

[154] Einat Shusterman, Hyunsoo Gloria Kim, Marc Facciotti, Michele Igo, Kamali Sripathi, David Karger, Avi Segal, and Kobi Gal. Seeding course forums using the teacher-in-the-loop. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 22–31, 2021.

[155] Joanna Sikora, MDR Evans, and Jonathan Kelley. Scholarly culture: How books in adolescence enhance adult literacy, numeracy and technology skills in 31 societies. *Social science research*, 77:1–15, 2019.

[156] Gordon CS Smith and Jill P Pell. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *Bmj*, 327(7429):1459–1461, 2003.

[157] Peter M Steiner, Thomas D Cook, William R Shadish, and Margaret H Clark. The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*, 15(3):250, 2010.

[158] Peter M Steiner, Yongnam Kim, Courtney E Hall, and Dan Su. Graphical models for quasi-experimental designs. *Sociological methods & research*, 46(2):155–188, 2017.

[159] Courtney Stewart. Learning analytics: Shifting from theory to practice. *Journal on Empowering Teaching Excellence*, 1(1):10, 2017.

[160] Nicole Sugden, Robyn Brunton, Jasmine McDonald, Ben Hicks, Cassandra Colvin, and Michelle Yeo. Delving deeper into student engagement: Conversations between academics and analysts. In *Australian College of Applied Psychology Conference: ACAP 2019*, 2019.

[161] Nicole Sugden, Robyn Brunton, Jasmine MacDonald, Michelle Yeo, and Ben Hicks. Evaluating student engagement and deep learning in interactive online psychology learning activities. *Australasian Journal of Educational Technology*, 37(2):45–65, 2021.

[162] Gail M Sullivan. Getting off the "gold standard": randomized controlled trials and education research. *Journal of graduate medical education*, 3 (3):285–289, 2011.

[163] Dan Suthers and Katrien Verbert. Learning analytics as a" middle space". In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 1–4, 2013.

[164] Robert I Sutton and Barry M Staw. What theory is not. *Administrative science quarterly*, pages 371–384, 1995.

[165] Dirk Tempelaar, Bart Rienties, and Quan Nguyen. Subjective data, objective data and the role of bias in predictive modelling: lessons from a dispositional learning analytics application. *PloS one*, 15(6):e0233977, 2020.

[166] Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016. doi: 10.1093/ije/dyw341.

[167] Deependra K Thapa, Denis C Visentin, Glenn E Hunt, Roger Watson, and Michelle L Cleary. Being honest with causal language in writing for publication. 2020.

[168] Felix J Thoemmes and Eun Sook Kim. A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46 (1):90–118, 2011.

[169] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125, 1975.

[170] Yi-Shan Tsai and Dragan Gasevic. Learning analytics in higher education—challenges and policies: a review of eight learning analytics policies. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 233–242, 2017.

[171] Yi-Shan Tsai, Oleksandra Poquet, Dragan Gašević, Shane Dawson, and Abelardo Pardo. Complexity leadership in learning analytics: Drivers, challenges and opportunities. *British Journal of Educational Technology*, 50(6):2839–2854, 2019. doi: https://doi.org/10.1111/bjet.12846. URL https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12846.

[172] Yi-Shan Tsai, Diego Rates, Pedro Manuel Moreno-Marcos, Pedro J Muñoz-Merino, Ioana Jivet, Maren Scheffel, Hendrik Drachsler, Carlos Delgado Kloos, and Dragan Gašević. Learning analytics in european higher education—trends and barriers. *Computers & Education*, 155:103933, 2020.

[173] Unesco. Education for sustainable development: a roadmap, 2020.

[174] Han LJ Van Der Maas, Conor V Dolan, Raoul PPP Grasman, Jelte M Wicherts, Hilde M Huizenga, and Maartje EJ Raijmakers. A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review*, 113(4):842, 2006.

[175] Neil van der Ploeg, Kelly Linden, Ben Hicks, and Prue Gonzales. Widening the net to reduce the debt: Reducing student debt by increasing identification of completely disengaged students. In *Proc. ASCILITE 2020 in Armidale*, ASCILITE's First Virtual Conference, pages 54–59. Australasian Society for Computers in Learning in Tertiary Education-ASCILITE, 2020.

[176] Neil van der Ploeg, Kelly Linden, and Ben Hicks. Job ready graduates package: The difficulty in identifying 'genuine students'. In *HERDSA 2022*. HERDSA, 2022.

[177] Sam Van Horne, Maura Curran, Anna Smith, John VanBuren, David Zahrieh, Russell Larsen, and Ross Miller. Facilitating student success in introductory chemistry with feedback in an online platform. *Technology, Knowledge and Learning*, 23(1):21–40, 2018.

[178] Yvonne Vezzoli, Manolis Mavrikis, and Asimina Vasalou. Inspiration cards workshops with primary teachers in the early co-design stages of learning analytics. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 73–82, 2020.

[179] Olga Viberg, Mathias Hatakka, Olof Bälter, and Anna Mavroudi. The current landscape of learning analytics in higher education. *Computers in human behavior*, 89:98–110, 2018.

[180] Toby Walsh. The ai revolution. *NSW Department of Education Education: Future Frontiers*, 2017.

[181] Xu Wang, Miaomiao Wen, and Carolyn P. Rose. Towards triggering higher-order thinking behaviors in MOOCs. In *LAK '16 CONFERENCE PROCEEDINGS: THE SIXTH INTERNATIONAL LEARNING ANALYTICS & KNOWLEDGE CONFERENCE,*, pages 398–407. Soc Learning Analyt Res; ACM, 2016. ISBN 978-1-4503-4190-5. doi: {10.1145/2883851.2883964}. 6th International Conference on Learning Analytics and Knowledge (LAK), Univ Edinburgh, Edinburgh, SCOTLAND, APR 25-29, 2016.

[182] Barbara Wasson, Cecilie Hansen, and Grete Netteland. Data literacy and use for learning when using learning analytics for learners. 2016.

[183] Sarah J Whatmore. Mapping knowledge controversies: science, democracy and the redistribution of expertise. *Progress in Human Geography*, 33(5): 587–598, 2009.

[184] Alexander Whitelock-Wainwright, Dragan Gasevic, Ricardo Tejeiro, Yi-Shan Tsai, and Kate Bennett. The Student Expectations of Learning Analytics Questionnaire. *JOURNAL OF COMPUTER ASSISTED LEARNING*, 35(5):633–666, OCT 2019. ISSN 0266-4909. doi: {10.1111/jcal.12366}.

[185] Jeff Whittle, Kathlyn E Fletcher, Jeffrey Morzinski, Kristyn Ertl, Leslie Patterson, Wayne Jensen, and Marilyn M Schapira. Ethical challenges in a randomized controlled trial of peer education among veterans service organizations. *Journal of Empirical Research on Human Research Ethics*, 5(4):43–51, 2010.

[186] Alyssa Friend Wise and David Williamson Shaffer. Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2 (2):5–13, 2015.

[187] Alyssa Friend Wise, Juan Pablo Sarmiento, and Maurice Boothe Jr. *Subversive Learning Analytics*, page 639–645. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450389358. URL https://doi.org/10.1145/3448139.3448210.

[188] Alan Woodley, Paul de Lange, and George Tanewski. Student progress in distance education: Kember's model re-visited. *Open Learning: The Journal of Open, Distance and e-Learning*, 16(2):113–131, 2001.

[189] Jim Woodward. Data and phenomena. *Synthese*, pages 393–472, 1989.

[190] Chi Zhang, Carlos Cinelli, Bryant Chen, and Judea Pearl. Exploiting equality constraints in causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR, 2021.

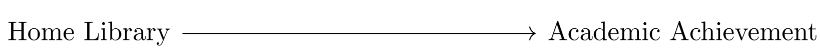
[191] John Zilvinskis, Victor Borden, and Ronald E Severtis Jr. Employing a quasi-experimental technique to measure the effects of conditional admissions. *Journal of College Student Retention: Research, Theory & Practice*, 22(1):141–154, 2020.

Appendices

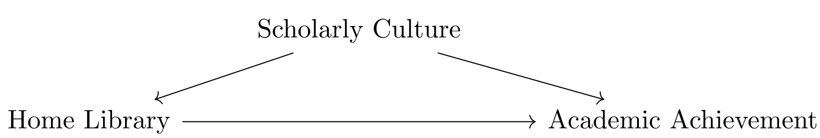
A A brief introduction to graphical causal models

In this section we provide a very brief introduction to graphical causal modelling (GCM) with DAGs. This is an extract from my paper Hicks et al. [63]. Our discussion will privilege concepts that support transdisciplinary communication over statistical rigour. For a deeper exploration of causal models see Pearl and Mackenzie [130] for an approachable introduction, or Pearl [127] for the full mathematical apparatus.

Let us start by supposing we want to model the effect of having an extensive home library on later academic achievement. We might start by assuming that children with access to a *Home Library* demonstrate improved *Academic Achievement*, as they are able to make use of those books to practice reading and learn about the world beyond the family home. This relationship is introduced by the simple DAG shown in Figure A.1(a).



(a)



(b)

Figure A.1: Two simple alternative causal models that describe the effect of a home library on a student's academic achievement.

However, on presenting such a diagram to an educator they would be able to raise an immediate (rather obvious to them) challenge to the model it represents, by pointing out that a child with an extensive home library is likely to have parents with a more scholarly background (after all, someone bought the books in the library!) But, says the educator, would such parents not have an impact upon the *Academic Achievement* of the student as well? At this point the modeller could respond that this is actually a concept that is well understood statistically: it is a possible confounding effect upon our model [155]. The modeller could construct the DAG shown in figure A.1(b) to incorporate this new causal claim, and the educator would be able to interrogate and approve (or challenge) the new representation. While simple, this vignette demonstrates the utility of a simple pictorial representation which has a well defined modelling apparatus to support it. The educator is facilitated in understanding and contributing to the development of the models shown in Figure A.1, without needing to understand the full complexity of the statistical model it represents.

Unpacking the statistical apparatus introduced in Figure A.1 a little more, the blobs (nodes) of the DAG represent variables that we have decided to include in the model. The arrows (edges) in the DAG represent our assumptions about how those variables influence one another. A useful way of analysing a DAG is to pick two variables that you want to explore the causal relationship between, (in this case from *Home Library* to *Academic Achievement*), and classify them as (i) open or closed, and (ii) causal or non-causal.

A.1 Causal Paths

Put your finger on any node of a graph and then trace your finger along the edges to reach another node and you have described a *path*. For instance starting at *Home Library* in Figure A.1(b) then moving up to *Scholarly Culture* and across to *Academic Achievement* traces the path *Home Library* $\leftarrow$ *Scholarly Culture* $\rightarrow$ *Academic Achievement*. A path like this shows a connection between the variables *Scholarly Culture* and *Academic Achievement*, but what kind of connection?

A.1.1 Causal or non-causal paths.

A *causal* path from X to Y means that if X is forced to change then Y will change also. This is not an easy phenomena to directly observe in the data alone. Data excels at showing how the variables are associated, not how influence flows from one variable to the next — this is why the direction of the arrows is an essential ingredient of a causal model. However, it is simple in a DAG to discriminate between causal and non-causal paths: a causal path is one where all the arrows go in the same direction.

A.1.2 Open or closed paths.

An *open* path allows information to flow along it. The way this manifests in the data is that the variables along an open path will be associated with one another — they will *not* change independently. A *closed* path will have a feature blocking the flow of association so that the two variables at each end of the path are independent of each other. To decide if a path is open or closed it is enough to examine a DAG to find one of the three possible patterns that can arise in them; the chain, the fork and the collider, depicted in Figure A.2.

Chain	$\mathrm{X} \longrightarrow \mathrm{C} \longrightarrow \mathrm{Y}$	*open*
Fork	$\mathrm{X} \longleftarrow \mathrm{C} \longrightarrow \mathrm{Y}$	*open*
Collider	$\mathrm{X} \longrightarrow \mathrm{C} \longleftarrow \mathrm{Y}$	*closed*

Figure A.2: The (unadjusted) three elemental DAG patterns; the Chain, Fork and Collider.

Why is the fork open? C is a common cause of X and Y, so when C changes we see changes in X and Y *together.* As they move together they will appear associated in the data. Why is the collider different? X and Y are common causes of C, so whilst C depends on the values of X and Y there is no reason that X and Y share information with each other. Plants need water and sunlight to grow, $Water \rightarrow Grow \leftarrow Sunlight$. But knowing how much *Water* a plant is getting does not tell us anything about much *Sunlight* it is getting; information does not flow along this path automatically, it is blocked by the collider *Grow*. This all changes if we learn something about the collider, and for this we will need to discuss *adjustment*.

A.1.3 Adjustment, opening and closing paths.

Adjusting[1] for a variable means that we include our knowledge about the variable in the analysis. Take the plant growth example, which begins with the variables *Water* and *Sunlight* independent of one another. Let us say we separate our study into groups; we look at plants that are growing well compared to plants that are not, that is we add knowledge of the variable *Grow* to our model. In this case, knowing something about the *Water* variable within the context of, for instance, dead plants, does say something about the status of *Sunlight*. Indeed, if you are looking a plant that is not growing, knowing that it has plenty of *Water* tells you something about the likelihood it is getting enough *Sunlight* — information now flows between *Water* and *Sunlight*, but only if we know the value of *Grow* as well. To express this again for any DAG, adjusting for the common cause in a collider pattern *opens* the path and allows the flow of association between variables. Visually, an adjusted $\boxed{\text{variable}}$ is placed in a box.

Why do the other patterns in Figure A.3, the chain and the fork, close when adjusting for the central variable? Remember that adjusting for a variable can

[1]Adjustment is also called *controlling*, *conditioning on* or *stratifying*.

Chain	$\mathrm{X} \longrightarrow \boxed{\mathrm{C}} \longrightarrow \mathrm{Y}$	*closed*
Fork	$\mathrm{X} \longleftarrow \boxed{\mathrm{C}} \longrightarrow \mathrm{Y}$	*closed*
Collider	$\mathrm{X} \longrightarrow \boxed{\mathrm{C}} \longleftarrow \mathrm{Y}$	*open*

Figure A.3: The three elemental DAG patterns; with central variable adjustment.

be conceptualised as *learning about* that variable. To illustrate this for a chain let us extend the plant growth model by adding the path $Grow \rightarrow Flower$:

$$\begin{tikzcd} & \text{Sunlight} \arrow[d] & \\ \text{Water} \arrow[r] & \text{Grow} \arrow[r] & \text{Flower} \end{tikzcd}$$

(a)

$$\begin{tikzcd} & \text{Sunlight} \arrow[d] & \\ \text{Water} \arrow[r] & \boxed{\text{Grow}} \arrow[r] & \text{Flower} \end{tikzcd}$$

(b)

Figure A.4: DAGs with a chains (e.g. *Water* → *Grow* → *Flower*) and a collider (*Water* → *Grow* ← *Sunlight*). In (a) association flows between *Water* and *Flower* but not between *Water* and *Sunlight*. When we adjust for *Grow* in (b) this opens the path between *Water* and *Sunlight* but closes the path between *Water* and *Flower*.

Knowing if a plant has had enough water certainly tells us about the chance of it flowering (*Water* and *Flower* are associated). But if we then learn that the plant is growing well, our knowledge of *Water* (and *Sunlight*) loses relevance. To see why, imagine we are looking at a verdant specimen brimming with life; its high level of *Grow* tells us all we need to know about the plants prospect of flowering — the importance of knowing if it is getting enough *Water* is diminished because the effect of *Water* on *Flower* is mediated through *Grow*, which we now have knowledge of. As such, within groups of plants with a similar level of growth (adjusting for *Grow*) there is no association between *Water* and *Flower*, the path is now blocked.

A.2 Confounding

Confounding occurs when we have association flowing along non-causal paths between the two variables we are trying to compare, which is what is happening when we try to measure the effect of *Home Library* on *Academic Achievement*. Returning to our original example, the path $Home\ Library \rightarrow Academic\ Achievement$ is an open causal path. It is a causal path because all the arrows flow in the direction we are thinking about, and it is an open path because nothing is 'blocking' the flow of information. As the path is *open* we would expect to see association between the variables *Home Library* and *Academic Achievement* in the data. As the path is *causal* we would expect that manipulating *Home Library* would change *Academic Achievement*. This is the path we are interested in measuring the strength of.

In contrast, the path $Home\ Library \leftarrow Scholarly\ Culture \rightarrow Academic\ Achievement$ is an open *non-causal* path. As the path is non-causal, manipulating *Home Library* will not influence *Academic Achievement* along this path. It represents the causal claim that adding books to the home library *does not* change the scholarly culture and in turn change academic achievement. As the path between the variables *Home Library* and *Academic Achievement* is *open* we would expect to see an association emerge in a dataset with these variables — even if there was no direct effect between *Home Library* and *Academic Achievement*! This is due to the common cause *Scholarly Culture*; as it increases we expect an (on average) increase in size of the *Home Library* and also in a student's *Academic Achievement*. While this results in an association between *Home Library* and *Academic Achievement* the cause of the association is not the *Home Library*. It is the harder to measure, or latent, variable *Scholarly Culture*. This is why *Scholarly Culture* is called a confounder, because the total association we see in the data between *Home Library* and *Academic Achievement* comes from two sources, $Home\ Library \rightarrow Academic\ Achievement$ (causal) and $Home\ Library \leftarrow Scholarly\ Culture \rightarrow Academic\ Achievement$ (non-causal); the direct causal effect of *Home Library* on *Academic Achievement* is confounded by the existence of other open non-causal paths. If we want to measure the direct effect of *Home Library* on *Academic Achievement* we need to block the flow of association along the non-causal path, isolating the causal path that we are interested in. There are two main methods that can achieve this goal; *randomisation* or *back-door adjustment*.

A.3 Removing confounding through randomisation

Randomisation involves manipulating how the influencing variable (*Home Library*) is generated, and forcing its value (the number of books) to be randomly assigned. This is what a Randomised Control Trial does, and in terms of the DAG it effectively removes all arrows *into* the influencing variable. This becomes governed purely by a random process and is now no longer influenced by any other variable in our model (See Figure A.5). However randomisation is not generally available as a tool in observational studies such as this, so another way to block the path is needed.

$$\begin{tikzcd}
\text{Random Number} \arrow[d] & \text{Scholarly Culture} \arrow[dr] & \\
\text{Home Library} \arrow[rr] & & \text{Academic Achievement}
\end{tikzcd}$$

Figure A.5: A randomised control trial DAG.

A.4 Removing confounding with the back-door adjustment

The non-causal path *Home Library* ← *Scholarly Culture* → *Academic Achievement* is known as a *back-door path* since it begins flowing against the direction of the arrow. The back-door adjustment provides a method of closing back-door paths by *adjusting* for key variables to block the flow of association along the path.

In terms of our example adjusting for *Scholarly Culture* blocks the flow of information along the path *Home Library* ← *Scholarly Culture* → *Academic Achievement*. This is because knowing about the level of *Scholarly Culture* tells us everything we want to know about the strength of the association between *Home Library* and *Academic Achievement* due to this common cause, information no longer flows along this path.

$$\begin{tikzcd}
& \boxed{\text{Scholarly Culture}} \arrow[dl] \arrow[dr] & \\
\text{Home Library} \arrow[rr] & & \text{Academic Achievement}
\end{tikzcd}$$

Figure A.6: Blocking the non-causal path by adjusting for Scholarly Culture.

So if we can measure the association between *Home Library* and *Academic Achievement* within levels of *Scholarly Culture* and then pool the results we obtain an unbiased estimate of the causal effect of *Home Library* on *Academic Achievement*, assuming the DAG is sufficient. We could do this by surveying all families in our system, to measure their scholarly culture, and then segmenting our sample accordingly.