Received: 7 December 2022 | Accepted: 16 March 2023

DOI: 10.1111/bjet.13321

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Using causal models to bridge the divide between big data and educational theory

Kirsty Kitto | Ben Hicks | Simon Buckingham Shum

Connected Intelligence Centre, University of Technology Sydney, Sydney, Australia

Correspondence

Kirsty Kitto, Connected Intelligence Centre, University of Technology Sydney, PO Box 123, Ultimo NSW 2007 Australia.
Email: kirsty.kitto@uts.edu.au

Abstract

An extraordinary amount of data is becoming available in educational settings, collected from a wide range of Educational Technology tools and services. This creates opportunities for using methods from Artificial Intelligence and Learning Analytics (LA) to improve learning and the environments in which it occurs. And yet, analytics results produced using these methods often fail to link to theoretical concepts from the learning sciences, making them difficult for educators to trust, interpret and act upon. At the same time, many of our educational theories are difficult to formalise into testable models that link to educational data. New methodologies are required to formalise the bridge between big data and educational theory. This paper demonstrates how causal modelling can help to close this gap. It introduces the apparatus of causal modelling, and shows how it can be applied to well-known problems in LA to yield new insights. We conclude with a consideration of what causal modelling adds to the theory-versus-data debate in education, and extend an invitation to other investigators to join this exciting programme of research.

KEYWORDS

big data, causal model, educational theory, learning sciences, learning analytics

Practitioner notes

What is already known about this topic

- ‘Correlation does not equal causation’ is a familiar claim in many fields of research but increasingly we see the need for a causal understanding of our educational systems.
- Big data bring many opportunities for analysis in education, but also a risk that results will fail to replicate in new contexts.
- Causal inference is a well-developed approach for extracting causal relationships from data, but is yet to become widely used in the learning sciences.

What this paper adds

- An overview of causal modelling to support educational data scientists interested in adopting this promising approach.
- A demonstration of how constructing causal models forces us to more explicitly specify the claims of educational theories.
- An understanding of how we can link educational datasets to theoretical constructs represented as causal models so formulating empirical tests of the educational theories that they represent.

Implications for practice and/or policy

- Causal models can help us to explicitly specify educational theories in a testable format.
- It is sometimes possible to make causal inferences from educational data if we understand our system well enough to construct a sufficiently explicit theoretical model.
- Learning Analysts should work to specify more causal models and test their predictions, as this would advance our theoretical understanding of many educational systems.

INTRODUCTION

As online learning becomes mainstream with the response to Covid-19, the shift to new models of learning has dramatically accelerated in a way that was unanticipated before 2020. In response, a wide range of Educational Technology (EdTech) tools which were previously patchily adopted have now become mainstream in both schools and universities (Rapanta et al., 2021). For example, while universities have made use of them for decades, most schools have now moved their resources into Learning Management Systems (LMSs), and web conferencing tools have become a standard technology for delivering lectures. Similarly, online proctoring for exams has become common, albeit with a fair amount of accompanying controversy (Selwyn et al., 2021). This explosion in technology has resulted in the generation of an unprecedented amount of educational data which practitioners are increasingly attempting to use to support student learning. Indeed, the shift to online and hybrid models of instruction frequently leaves instructors in the dark about student engagement, where they are struggling and how class dynamics are evolving, as the visual clues that they traditionally rely upon are lost. The resulting trend towards using methods from Artificial Intelligence (AI) and Learning Analytics (LA) to provide insights to instructors, institutions and even learners, has accordingly accelerated.

However, as the educational domain moves towards an era of big data, we might ask if the data being collected are fit for purpose (Kitto et al., 2020). Indeed, there is growing evidence to suggest that the data resulting from these many different learning environments are difficult to link to theoretically grounded educational concepts that decision makers can utilise to improve student outcomes (Knight & Buckingham Shum, 2017; Wise et al., 2021). This has led to a rise in calls to more tightly embed educational theory into these advanced analytics tools and methods (Guzmán-Valenzuela et al., 2021; Marzouk et al., 2016; Winne, 2020), and to claims that as our datasets grow, theory becomes more important than ever (Wise & Shaffer, 2015). But what precisely do we mean by this rather ambiguous concept? Almost every field of research has a different notion of what a theory is, what it is not (Sutton & Staw, 1995), along with the associated debates about how theory should inform our analytical methods. The learning sciences are no exception. These ongoing debates open up possibilities for us to learn from past efforts to conceptualise the interplay between theory and data in other domains. So what lessons can we learn from these past controversies? We will start by considering a similar debate that occurred in AI.

A historical example from Artificial Intelligence

In 2011 a very public debate exploded in the world of AI, when Peter Norvig[1] wrote a blistering response (Norvig, 2012) to Noam Chomsky[2] who had

> derided researchers in machine learning who use purely statistical methods to produce behaviour that mimics something in the world, but who don't try to understand the meaning of that behaviour.
>
> (Cass, 2011)

Chomsky insisted that we need to construct theories and interpretable models that explain the underlying patterns in a dataset, disparagingly referring to the statistical models being used in computational linguistics as ‘stamp collecting’. Norvig (2012) countered by pointing out that modern statistical models were performing so well that they were beating most theoretically informed models, and that more than 90% of researchers in the field were using them. In short, Norvig was claiming that Chomsky's insistence on a theoretical model was outdated, and that with enough data we would not need to develop an underlying understanding of the system being modelled. Instead, Norvig (2012) argued that we must accept that ‘nature's black box cannot necessarily be described by a simple model,’ (p 32), and that Chomsky's desire for an interpretable model was naive. While this debate centred upon linguistics, it can be understood as a much broader argument about what counts as a satisfactory model in a field exposed to big data. Norvig claimed that developing statistical models using large datasets and testing the results obtained using predictive accuracy was sufficient and that we did not need to understand the resulting predictions. In contrast, Chomsky claimed that more is required, and that we should be seeking explanatory principles that provide an understanding of the phenomena being studied. No clear consensus emerged from this debate about which approach was ‘best’. Indeed, the two sides of this debate centre around very different philosophical positions, which leads to very different interpretations of their claims. This argument between the ‘two cultures’ (Breiman, 2001) of science is hardly new, and indeed, with the rise of large language models and new tools like ChatGPT we see the power of statistical models coming to the fore again, albeit with an associated loss of interpretability (Bender et al., 2021). Arguments about the relative value of statistical-versus-theoretical modelling have arisen in many fields, under a wide range of guises, but fundamental to this debate is a difference in understanding about the purpose of our models. Are they for *predicting* system behaviour or for providing a causal understanding of the system being studied? More recently, when Judea Pearl referred to deep learning as mere 'curve fitting' (Hartnett, 2018) he can be seen as reinvigorating this data-versus-theory debate in AI. Interestingly, Pearl argues that a far richer understanding of *causality* is required for AI to reach a level of modelling equivalent to human general intelligence.

In this paper, we will argue that the modern apparatus of causal modelling (Pearl, 2009) offers materials to build a 'middle space' for *learning* and *analytics* to meet (Knight et al., 2014; Suthers & Verbert, 2013), by formalising the *causal* claims made by learning theory in a form that enables them to come into dialogue with the *statistical* machinery often used by analytical approaches.

Before commencing on this journey, we will consider how this debate plays out in the domain of the educational data sciences, pointing to a need in the field to make causal claims without resorting to controlled trials. This will provide us with motivation to more clearly articulate what we mean by a theory in education, before we move onto introducing the apparatus of causal modelling.

Education and big data

Education, at its heart, is about improving our society, our productivity, our culture, our knowledge, our understanding and our thinking (UNESCO, 2020). To do this, educational institutions need to be able to make effective interventions that improve learning outcomes and the environments in which people learn. The emphasis upon evidence-based practice has led to the increasing collection and presentation of data in education. The resulting wide ranging availability of data has driven the emergence of fields like LA and Educational Data Mining (EDM), which aim to provide actionable insights that would help to support this improvement (Jørnø & Gynther, 2018).

However, successfully intervening in a system as complex as education is no easy task (Davis & Sumara, 2010). A number of high-profile attempts to generate lasting change in educational settings have floundered. Consider for example the 2018 finding that a $575 million Bill & Melinda Gates Foundation initiative, which sought to measure and improve teacher effectiveness, failed to boost student achievement in any measurable sense (Stecher et al., 2018). The report notes that 'The [Intensive Partnerships] initiative might have failed to achieve its goals, because it succeeded more at measuring teaching effectiveness than at using the information to improve student outcomes' (Stecher et al., 2018, p. 564). Deciding what to measure, and then determining how to gather sufficient evidence to support an intervention, is a difficult problem. Even more problematic is assessing evidence of whether the intervention was successful. In the above case, it was found that measuring teacher effectiveness does not necessarily equate to improving student outcomes. It could be argued that the wrong data were being used to assess the effectiveness of the intervention that was supported by this program. But how are we to decide which data are the correct data to use?

Approaching this problem requires a deeper knowledge of the system than can be provided by the data alone (Pearl & Mackenzie, 2018). It is not enough to draw inferences about the associations between variables in an educational dataset. If we are to successfully intervene in a system then we must be able to understand *what* can be changed and *why* this should work. But this requires an understanding of the direction in which our associations (or correlations) flow, i.e. how one variable might influence another. In short, we require a *causal* understanding of the relationship between two variables.

However, many who work with data are all too familiar with the catchphrase that 'correlation does not equal causation', a claim that harkens back to the early days of statistics when Pearson equated causation with the statistically well-defined concepts of perfect correlation or anticorrelation (Pearl & Mackenzie, 2018, p. 66). This rather dramatic move resulted from the well-known difficulty of ensuring that a sequence of repeating events *must* continue to repeat in that pattern with absolute certainty:

> That a certain sequence has occurred and recurred in the past is a matter of experience to which we give expression in the concept causation. … Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated. (Pearson, 1911, p. 113)

Thus, Pearson followed in Hume's footsteps, declaring that we must be sceptical about the possibility of linking two events using any claims stronger than the declaration that a correlation exists between them. Indeed, Pearson and his students were able to demonstrate that many causal relationships between two events *A* and *B* were spurious, or at best due to an underlying *confounding* variable that could explain the correlation. For example, we might be interested in whether having access to a home library influences students' academic performance. While there may be a direct effect between these two variables (eg, a student with access to books at home might read more and consequently perform better in class) it is also possible that some other variable is the underlying cause of both of these outcomes. Thus, it may be that the parents of the student are highly educated, better able to help them with their studies and expose them to many stimulating life experiences, while also buying many books.

It took Fisher to recover the notion of causal claims, some 25 years later, when he formalised the notion of a randomised experiment. Jamison (2019) provides a comprehensive overview of the complex history of this concept, tracing its historic use, before demonstrating how it became embedded in the social sciences as a tool for demonstrating causal relationships across a number of different fields. In short, a randomised experiment enforces the *randomised assignment* of conditions, in a series of experimental observations, resulting in a set of outcomes. This removes the exogenous causes of the conditions as they are now governed by the randomisation process, which in turn removes any confounding from a common cause between the conditions and the outcomes. Performing this removal enables us to measure the causal effect of the treatment (i.e., the conditions) on an outcome, a form of experimentation now commonly referred to as a randomised controlled trial (RCT). RCTs are now commonly considered the gold standard for making causal claims. Thus, we see that the stances adopted by Pearson and Fisher gained particular prominence in the field of statistics, and from there have become influential in all the sciences, including the Learning Sciences, EDM and LA (although we acknowledge that not all researchers in these fields adhere to this stance).

There are numerous robust examples of RCT methodology in the learning sciences. For example, Brooks et al. (2018) performed a RCT on students enrolled in a Massive Open Online Course (MOOC). They assigned participants randomly to either a female condition (n = 23,365, male students = 18,482, where the instructor was a woman with two female 'data scientists' working in the background) or a male condition (n = 23,287, male students = 18,478, where a male instructor had two male 'data scientists' working in the background). The random assignment of participants to a condition enabled the authors to show that while no causal connection could be demonstrated between the condition and student persistence, the engagement of women assigned to the female condition was significantly improved. They were also able to show that a correlating strong but small negative effect occurred in men assigned to the female condition. Similar interesting results have been found by a number of other studies in MOOCs (Boaler et al., 2018; Hossain et al., 2015; Kizilcec et al., 2020; Vilkova, 2022), although they are far less common outside of the MOOC format. Interestingly, Motz et al. (2018) demonstrate that these claims need not be made exclusively in a laboratory setting with controlled experiments. They can also be embedded into a genuine learning setting, and often at scale.

The problem with experimental studies in education

However, it is far from true that every educational phenomenon that we would like to study can be meaningfully subjected to a RCT. Sometimes we are interested in an educational construct that cannot be directly measured, let alone directly changed by intervention (eg, motivation or self-regulation). In such cases we would need to create an experiment by changing the conditions of variables that are designed to serve as proxies for the latent variables thought to affect the outcome under study. At other times, it is impossible to split a set of students into two randomised groups (eg, there may only be one class of students, or a morning and evening class, which might be indicative of a confounding variable such as ‘employment status’). It might even be unethical to split a set of students into two groups. Performing an intervention on one group but not the other could lead to significant differences in life outcomes, and it can be difficult to justify an experiment in this context. It can also be difficult to properly blind participants in education, and students often become aware that they are under study (Sullivan, 2011). These complexities mean that many domains of education fail to meet the stringent requirements of a RCT. As a result, LA has tended to report upon correlations far more frequently than it makes causal claims. For example, Wong et al. (2019) conducted a systematic review of empirical LA papers that discussed student success, finding that prior to 2017 (when their search was performed) only two of 20 papers conducted a (quasi)experimental study, with the rest reporting correlations alone. Interestingly, 16 of these papers made use of a learning theory of some form, with self-regulation being the most commonly used theory ($n=6$).

Furthermore, while randomisation in experiments provides internal validity for a statistical result (ie, like is compared with like) it makes no guarantees about external validity (ie, the extent to which results can be generalised to other settings). This often leads to problems where previously reported results in LA fail to replicate. For example, Joksimovic et al. (2018) conducted a review of approaches to modelling learning in MOOCs, finding that many results in contemporary MOOC research lack generalisability. This finding was supported by the impressive study of Kizilcec et al. (2020), who tested a range of interventions with more than 250,000 students in 247 online courses, demonstrating that few established interventions actually resulted in beneficial outcomes for students. This failure to replicate frequently results from an over-reliance upon hypothesis testing, where a battery of tests is applied to a dataset with little theoretical motivation. As more tests are applied it becomes more and more likely that a false positive will be returned. Gelman and Loken (2014) have referred to the choices made in performing this type of analysis as a ‘garden of forking paths’. This term describes the way in which the many valid decisions that have to be made during a data analysis tend to add hidden variables which are not factored into hypothesis testing, making a result seem stronger than it actually is. One of the advantages of using an approach driven by theory is that it can limit this form of ad hoc data analysis. Rather than performing a battery of tests over every variable in a dataset, a theory provides the rationale for focusing on specific variables, and can make predictions about how a system should behave under changing conditions. Theory can also help a data analyst to define what data they should collect in the first place, a point recently made by Winne:

> This analysis reveals the essential and inescapable role of theory in deciding what trace data should be gathered and how trace data can contribute to recommendations for improving learning, one main goal for generating and using learning analytics. (Winne, 2020, p. 1)

Interestingly, in the same paper, Winne also points to a primitive notion of causality when he makes use of IF-THEN rules to describe how a system with states might change in time. We see here a desire to use theory to move beyond the standard statistical picture. But what precisely do *we* mean by a theory in this paper?

But what do we mean by theory in education? A lesson from the psychological sciences

LA often brings people with very differing perspectives together to construct a 'middle space' understanding of our educational systems, at the intersection of *learning* and *analytics* (Knight et al., 2014; Suthers & Verbert, 2013). This results in a wide array of understandings about what the term *theory* actually means (Lodge et al., forthcoming). Education has produced many different theories over three broad epochs, often falling into something approximating behaviourist, cognitive and contextual categories (see eg, Murphy & Knight, 2016). However, a recent paper by Kahlil et al. (2022), calls attention to the complexity of the concept of an educational theory, pointing to the many different claims about what a theory actually is in the learning sciences. The authors conclude that 'there is no single accepted definition of what constitutes a learning theory' (p. 5). We agree. It is often easier to define what theory is *not* than it is to define theory per se (Sutton & Staw, 1995). Indeed, Sutton and Staw mount a convincing argument from organisational theory to demonstrate that a number of rhetorical devices commonly deployed in papers as a representation of theory (specifically: references, data, variables, diagrams and hypotheses), cannot in fact be taken to be sufficient representations of this extremely complex concept.

Interestingly, Kahlil et al. (2022) perform a scoping review of LA publications (in the two Society for Learning Analytics Research venues, the *Learning Analytics and Knowledge* conferences and the *Journal of Learning Analytics*) from 2011 to 2020, demonstrating that Self-Regulated Learning (SRL) is by far the most dominant theory used in the field, with 26 papers using this concept in that date range. The next closest theories were Cognitive Load theory and Constructivism (at six papers each). One thing to note about each of these theories is that they are very different in flavour from the theories that have emerged in the physical sciences. For example, in physics Newton's laws of motion serve to: identify the relevant phenomena that lie at the heart of motion (forces and mass); explain the epiphenomena that result from their interaction (velocity and acceleration); and make testable predictions about how an object will move if we understand its mass and the forces that it is subjected to. While we would not claim that all theories must take this 'physical' form, we do think that this type of explanatory power is something for which the learning sciences should be striving.

One way of understanding this divergence between the various approaches to theory in different fields can be found in Borsboom et al. (2021), which focuses upon *explanatory theories* in the psychological sciences. These are the theories which help us to draw a relationship between measured data and the underlying phenomena that it represents. They help us to understand the world, and in particular to intervene in it by providing a 'thinking tool' that allows us to track the consequences of our theoretical principles. We believe that this approach has broad applicability across all of the learning sciences. Intriguingly, Borsboom et al. (2021) propose a Theory Construction Methodology (TCM) which consists of five stages:

1. *Identify empirical phenomena*: This step aims to find robust, stable and reproducible empirical generalisations that explain the phenomena of interest. These phenomena must be well established, and perhaps even self-evident in order to provide a solid foundation. This means that they have a solid empirical grounding and so can function as explanatory targets.
2. *Develop proto-theory*: At this stage of the TCM abductive reasoning is typically used to develop an explanatory model. This normally involves the identification of a small set of general principles which somehow explain the empirical phenomena identified in step 1 using hypotheses, models and theories. Proto-theories are *explanatory*, and often consist of conceptual diagrams and stories about how phenomena affect one another, and might 'borrow' explanatory principles from other fields. However, they are difficult to empirically test as a number of choices must still be made in formalising them which can lead to non-replicable results between different research groups (Gelman & Loken, 2014; Kitto et al., 2023).
3. *Develop a formal model*: A formal model captures the explanatory principles of the proto-theory in a set of rules or equations that can be used in a simulation or computer program. The formal model can be understood as an implementation of the proto-theory, although in some highly formalised fields (eg, physics) they will be almost the same. Crucially, Borsboom et al. (2021) point out that formal theoretical models should not be confused with data models. That is, while fitting parameters to data (eg, by performing a regression, correlation analysis, ANOVA, etc.) helps us to understand the data we have collected, it does not help us to understand the underlying phenomena that we are trying to explain. This suggests that the TCM aligns more closely with Chomsky's position than that of Norvig.
4. *Check explanatory adequacy*: Once an explanatory theory has been formalised into a testable model it becomes possible to test whether it is able to explain the empirical phenomena identified in step 1. These phenomena must also be formalised in the same language as the formal model, which then enables us to test whether the theory, as represented in the model, generates the phenomena.
5. *Evaluate the overall worth of the theory*: At this point it becomes possible to systematically evaluate the worth of the theory, using an established method. The hypothetico-deductive method is often used to evaluate the predictive success of the model, but other approaches are also possible. For example, Borsboom et al. (2021) prefer to evaluate theories in terms of their explanatory value.

Through the lens provided by the TCM, it becomes possible to claim that many of the theories we see in the learning sciences could be placed at the proto-theory stage. They have yet to formalise the connection between the phenomena that they seek to describe at a level of detail sufficient for checking the explanatory adequacy of the theory. Of most concern, the data that we collect for LA are often only loosely coupled to the empirical phenomena that we seek to model. How might we identify which empirical phenomena are primary to the behaviour that we seek to understand? And how can we become more sophisticated about the data that we collect?

Here we will argue that causal modelling provides a mechanism for moving from proto-theory into more formalised theoretical frameworks in the educational data sciences. In what follows we will proceed by gradually introducing the technical apparatus of causal modelling. This will enable us to explore some of its key advantages using examples from educational scenarios, and to demonstrate how causal modelling facilitates the move from an educational proto-theory into a model that can be tested for its explanatory adequacy. We will use two examples, one from SRL and another from Reflective Writing Analytics (RWA) to demonstrate how a causal approach offers new insights about the form that a formal model of these phenomena should take. This will then lead to a higher level discussion about how causal modelling could advance the learning sciences, before we issue a call for more work to be completed on this exciting topic.

CAUSAL MODELLING—A BRIDGE BETWEEN DATA AND EDUCATIONAL THEORIES

Causal modelling enables us to make well informed causal claims by imposing a theoretical structure upon a dataset (Pearl, 2009, p. 38). This approach to causal inference provides a mechanism for data science and AI to move beyond statistical models based primarily upon correlation by encouraging modellers to think about the processes (and hence the underlying phenomena) that generate a dataset (Lübke et al., 2020). Importantly, causal modelling provides a way to move beyond the 'correlation does not imply causation' mantra that is common in many fields using observational data, without the need to resort to the running of RCTs. Recent papers in LA have suggested that this opens up opportunities to involve stakeholders in the construction of educational models (Hicks et al., 2022) and to reduce potential bias in educational research that relies upon observational data (Weidlich et al., 2022).

Pearl and Mackenzie (2018) sketch out three sources of association between variables that we might see when performing a data analysis: causation, confounding and endogenous selection. Causal *mediation* is usually what we are most interested in, and is the result of an association that arises because one variable *causes* changes in the other variable. Confounding can occur when the association between the two variables arises from a third variable that is a *common cause* of the other two. For instance, an association between *time spent in the laboratory* and *achievement in science* might be due to a confounding variable, *subject enrolment*, and not a causal link between merely occupying a particular space and learning. Finally, endogenous selection can induce associations in the data, and occurs when we (sometimes unwittingly) select data based on a third variable that is a *common effect* of our variables of interest. For example, suppose we analyse the behaviour of students at a prestigious institution. To be accepted they would need to be *hard working*, or *talented*, or possibly both. We might then observe that *hard working* students tend to be less *talented* amongst the population under study. However, this association might be due to the students being selected into an elite institution (a common effect of *talent* and *hard work*) and not due to a causal link between *talent* and *hard work*. These three sources of association form the basis of all causal models, and the way in which causal modelling enables us to formalise them will be discussed in the next section.

Pearl and Mackenzie (2018) argue that a three-layer *ladder of causation* (see Table 1) can be used to classify the causal claims made by the different types of questions asked in an analysis:

1. *Association*: This basic level of causality uses statistical relationships, correlations, curve fits, etc. to infer relationships between variables in a dataset. No causal information can be extracted, leading to the frequently cited adage that 'correlation does not equal causation'. Much of the current work in LA falls into this category.
2. *Intervention*: This level involves not just measuring an effect, but changing an input and then recording an outcome. In performing this form of analysis we learn much more about how we might *change* outcomes, rather than simply observing them (which leads to a higher classification in the ladder). Work in LA that relies upon RCTs can be seen to fall into this category.
3. *Counterfactual*: This highest level of the ladder involves considering *what would have happened* had we done things differently. Beginning with Pearl (2009), an advanced

mathematical apparatus has been constructed which enables us to construct counterfactual models in a mathematically precise way. However, very little work in LA currently falls into this category.

Counterfactuals are placed at the top of the ladder by Pearl and Mackenzie (2018) because they subsume interventional and associational questions. That is, it is possible to answer questions about interventions and associations if we can answer counterfactual questions, and similarly, we can often answer questions about associations if we have performed an intervention. For example, Albacete et al. (2019) subjected control and experimental groups to different treatments in an Intelligent Tutoring System (ITS) to test two different initialisation strategies. This is a clear intervention type study according to our understanding of the causal ladder illustrated in Table 1. The data collected in this study can be used to answer associational type questions at lower levels of the ladder, such as: what is the relationship between a student's pre-test performance (which is used to construct the learner model used in the control group) and the rate at which they correctly answer questions in the ITS. However, these data cannot answer questions that are further up the ladder. In essence, data from measuring associations cannot provide us with information about RCTs, and we cannot re-run an A/B test to find out what *would have happened* had the treatment groups been reversed.

Importantly, developing models at higher levels of the causal ladder requires substantial theoretical grounding. Instead of simply performing a correlation analysis where patterns between variables are discovered, both interventions and counterfactual claims require a model to be well enough developed that sensible hypotheses can be proposed ahead of time and then tested. Pearl and Mackenzie (2018) explore many methods for recovering causality in an analysis, demonstrating that it is possible to do this without relying solely upon RCTs. This flexibility makes causal inference an ideal modelling tool for the observational studies that frequently occur in the educational data sciences. A suite of tools including Bayesian networks, do-calculus, and causal diagrams have been developed over the last two decades, and can now be used to make far stronger causal claims about a system of interest than is generally realised.

How does causal inference work?

To make causal inferences we must first find and ameliorate for potential sources of confounding and endogenous selection to unmask the causal effect of interest. This is often done by using a theoretical understanding of the system to specify a causal model, and then encoding that structure into relevant guidelines for a statistical model (Pearl, 2009).

One approach for achieving this translation uses causal diagrams to represent our theoretical understanding of a system (Pearl, 2009). These diagrams have formal constraints on how they are drawn, which enables us to utilise a graphical calculus of interventions, the *do*-calculus, to build a statistical model that makes causal inferences by minimising non-causal sources of association. The diagrams used in this method are Directed Acyclic Graphs (DAGs), where the nodes of the graph represent our variables of interest and directed arrows between the nodes represent the flow of causation. DAGs are required to be *acyclic* in order for the diagram to be understood by the statistical model. This means that there are no closed-paths (ie, loops) allowed if we follow the arrows around the diagram. Figure 1 shows an example of how we might use a DAG to represent a theoretical hypothesis about the relationship between *Available Time, Prior Learning, Effort* and *Performance*.

Used like this a DAG is a simple causal model. DAGs are sometimes also referred to as Graphical Causal Models or Structural Causal Models, albeit with some technical differences

TABLE 1 The three layers of Pearl and Mackenzie's (2018) causal ladder, with the types of questions asked at each layer, and indicative examples from education, along with typical formalism that is provided by Pearl (2009) for modelling such systems.

Causal ladder level	Types of question	Examples from education	Mathematical structure	Example models
1. Association	What is…?	• *What is* the probability of a student being 'at risk' of *X*? • *What is* the relationship between *Y* and student success? • *What* clusters of student activity exist in a logfile for a LMS? • *What* curriculum object most closely fits an identified knowledge gap?	Association type models tend to construct statements like $P(y\|x)=p$ (ie, the probability of event $Y=y$ given that we observed event $X=x$ is equal to p)	Jovanovic et al. (2019), Tempelaar et al. (2018)
2. Intervention	What if…? A/B questions	• *What if* I try replacing the lectures for this subject with flipped lessons? • *What effect* does an open learner model have on the performance of a sub-cohort? • *How does* the performance of a group of students using an ITS compare to a group exposed to a standard lecture and tutorial structure	Intervention type models are more concerned with statements like $P(y\|do(x),z)$ (ie, the probability of event $Y=y$ given that we intervene and do something such as setting the value of X to x and then subsequently observe event $Z=z$)	Albacete et al. (2019), Brooks et al. (2018), Kizilcec et al. (2020)
3. Counterfactual	Why? Did X cause Y? What if I had…? (subjunctive)	• *Why* do my students get better grades when I do *X*? • *Did* this open learner model *cause* higher student self-efficacy? • *What if* a student who dropped out had been provided with support?	Counterfactual models go one step further again, talking about $P(y_x\|x', y')$ (ie, the probability that event $Y=y$ would have been observed had X been x, given that we actually observed X to be x' and Y to be y')	Brokenshire (2007), Brokenshire and Kumar (2009), Deho et al. (2022)

Note: The last column lists example papers at this level in the ladder.

(Peters et al., 2017, p. 84). For this paper we shall use the term DAG to refer to a graphical and acyclic representation of a causal model.

From conditional independence to causal effects: Chains, forks and colliders

Once created, the causal structure of a DAG can be analysed graphically to yield two key insights. Firstly, the DAG implies a number of conditional independence relationships between the variables that would be expected in the data if the model is a good approximation of the system under study. Secondly, the DAG can help make causal inferences. DAGs provide a way to identify sets of variables to include as controls in a statistical model in order to build an estimator for a causal effect. This process of identifying variable sets from the causal structure in order to estimate a causal effect is known as *identification*. Being able to focus on *identification* separately to the *estimation* of causal effects is a distinguishing feature of the DAG based approach (Weidlich et al., 2022). While most work with causal models has focused on identification and estimation of causal effects, or the discovery of causal structures from data, this paper will explore the affordances of the implied conditional independence relationships of a causal DAG for the development of a stronger (ie, more formalised) theory.

It is worth taking a moment to be precise about language, as many different fields, each with their own lexicon, have contributed to the study of causal models. Key to understanding causal modelling is a notion of *conditional independence*, which implies that two variables do not influence one another under certain conditions. This is a property of the causal structure, or probability space, of the model. When translated to actual data, this property of conditional independence between two variables implies that under certain *conditions* (often termed *controlling* for a variable) these variables exhibit *independence* (ie, are not *associated* or *correlated*). Depending upon the field, *controlling for* a variable can sometimes be called *conditioning on* that variable, or *adjusting for* it. However, all of these phrases mean the same thing; that we are including knowledge of the controlling variable into our model. Controlling for a variable can be achieved using various statistical techniques, such as using *stratification* to split the analysis by the different levels of the control variable and then pool the results, or including the variables as *predictors* or *regressors* in a regression model (Lübke et al., 2020). Here we will use the term most common to education, *controlling for*, to indicate the inclusion of information about a variable into a causal model.

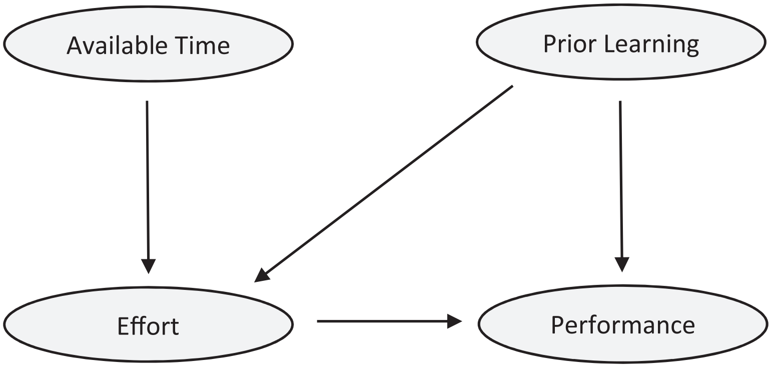

FIGURE 1 A Directed Acyclic Graph representing the causal relationships between *Prior Learning, Available Time, Effort* and *Performance*.

Having clarified this point of terminology, we are now ready to demonstrate how it becomes possible to move from a graphical causal structure to practical applications. Put simply, the graphical analysis of any DAG, no matter how complex, can be broken down into three elementary patterns with their own implications for conditional independence: the chain; the fork; and the collider, which correspond to mediation, confounding and endogenous selection respectively.

The chain (causal mediation)

A chain describes the pattern $A \rightarrow B \rightarrow C$. If we have a chain in our DAG then we would expect to see an association between the variables *A* and *C* due to the influence of *A* on *B* which flows through to *C*. However, we would expect to see the influence along this path 'broken' if we control for *B* (indicated graphically by placing a box around the relevant variable: $A \rightarrow \cdot\cdot\boxed{B} \rightarrow$ C), which serves to block the causal path. This can be articulated as $A \perp\!\!\!\perp C \mid B$ (*A* and *C* are conditionally independent on *B*).

In Figure 1, the pattern *Prior Learning → Effort → Performance* is an example of a chain. We would expect to see an association (ie, a dependence) in the data between *Prior Learning* and *Performance* due to the proposed influence that our model suggests flows along this path. Note that Figure 1 also proposes a direct influence of *Prior Learning* on *Performance*. However, if we control for *Effort* this would block the causal influence of *Prior Learning* on *Performance* along that mediating path. Performing this form of control can therefore be used to isolate the direct causal effect of *Prior Learning* on *Performance* and so to create experiments that can be used to test the causal claim.

The fork (confounding)

A fork describes the pattern $A \leftarrow B \rightarrow C$. Here we would expect to see an association between the variables *A* and *C* due to the common influence of *B* on both *A* and *C*, not due to a direct causal flow between *A* and *C*. However, we would expect to see the influence along this path 'broken' if we condition on *B* as we would now be looking at the relationship between *A* and *C* within specific levels of *B* (eg, that could have been created by a stratification process). This can be articulated as $A \perp\!\!\!\perp C \mid B$ (*A* and *C* are conditionally independent on *B*).

In Figure 1 we can see an example of a fork pattern in the *Effort ← Prior Learning → Performance* component of the diagram. We would expect to see an association (ie, a dependence) in the data between *Prior Learning* and *Performance* due to the influence along this path, as well as the direct influence of *Effort* on *Performance*. As a result, we would expect to see an association between *Effort* and *Performance*, but this would not be due to a direct causal influence, rather it would be due to the confounding *Prior Learning* variable. However, if we control for *Prior Learning* then this would block the influence of *Effort* on *Performance* along that alternative path, which can be used to isolate the direct causal effect of *Effort* on *Performance*.

The collider (endogenous selection)

A collider, also recently called an 'inverted fork' by Weidlich et al. (2022), describes the pattern $A \rightarrow B \leftarrow C$. We would not expect to see an association arise between the variables *A* and *C* if we had theorised a pattern of this form. However, in this case we would expect to see an association arise if we were to control for *B*. To understand how this occurs imagine a scenario where *B* is the result of adding together *A* and *C*, where *A* and *C* are independent of each other. Suppose further that our control enabled us to only analyse high values of *B*. Under the condition of *B* being high a low value of *A* would imply that *C* must be high. Introducing knowledge of *B* (controlling for *B*) created a dependence between *A* and *C* that was previously absent. This can be articulated as $A \perp\!\!\!\perp C \mid B$ (*A* and *C* are conditionally *dependent* on *B*).

In Figure 1 *Available Time → Effort ← Prior Learning* is an example of a collider pattern. We would not expect *Available Time* and *Prior Learning* to be associated in the data. However, if we control for *Effort* this opens up the path between *Available Time* and *Prior Learning*. If we examine only students exhibiting low *Effort*, then knowing a student has low *Prior Learning* implies they likely do not have much *Available Time*, because they should hopefully be putting in more *Effort* if they could. Because we have conditioned on the collider *Effort* it has opened up the flow of information (and association) between the variables *Prior Learning* and *Available Time*, where previously they were independent.

From DAG to statistical tests

Having developed a graphical representation of our theoretical understanding of performance, we can then translate our theory into practical information for a statistical modeller using the *do*-calculus. For instance, the DAG in Figure 1 lets the statistical modeller know that if they want to estimate the direct *causal* effect of *Effort* on *Performance* they should control for *Prior Learning* (and not *Available Time*), and inspect the resulting coefficient between *Effort* and *Prior Learning*. These kinds of *causal* models lie somewhere between statistical models, which attempt to identify associations between phenomena without the notion of the *direction* of influence, and the physical models of the 'hard' sciences, which explicitly state the dynamics of the system (Peters et al., 2017).

A causal DAG is a good theoretical representation of the system under study if the claims that it makes about conditional independence are in fact observed in the data that are collected for the variables described. That is, when we control for certain variables we should see that other pairs of variables have very low measures of association (ie, are independent). In Figure 1 we would expect *Available Time* and *Performance* to be independent, when controlling for both *Effort* and *Prior Learning*. We would also expect *Available Time* and *Prior Learning* to be unconditionally independent. Each causal DAG produces a set of these implications, in the form of conditional independence relationships, that can be then tested using a dataset. This could be done by using a test for statistical independence, such as Pearson's chi-squared test, on each pair of variables stratified by the given conditional variables. Numerous software packages, such as *bnlearn* (Scutari, 2010), and *DoWhy* (Sharma & Kiciman, 2019) provide functions to perform this type of test automatically.

The link between causal models and SEM

A Structural Equation Model (SEM) uses a functional representation of the relationships between variables, rather than the (non-parametric) graphical representation in a DAG. The causal relationship $A \rightarrow B$ in a DAG would be represented as $B = \beta A + \varepsilon$ in an SEM, introducing the parameters β and ϵ, and placing the cause(s) on the right-hand side of the equation and the effect on the left. We now show how the causal DAG in Figure 1 can also be represented, this time with parameterisation, using a SEM. Setting *Prior Learning* to L, *Available*

Time to A, *Effort* to E and *Performance* to P, with coefficients β_0, β_1, β_2, β_3 and errors ε_0, ε_1 the model structure is defined as follows:

$$E = \beta_0 L + \beta_1 A + \varepsilon_0$$

$$P = \beta_2 E + \beta_3 L + \varepsilon_1$$

This pair of equations encodes the causal paths of the DAG by placing the *immediate* causes of each variable on the right-hand side of the equation. So the path *Available Time* → *Effort* is represented by the presence of *A* on the right-hand side of the equation with *E* on the left-hand side. Both representations address a key ingredient not captured by statistical correlations alone; the *direction* of the flow of influence between variables. In DAGs this is shown with the direction of the arrows, and in SEMs with the side of the equation the variables are on, where the flow is from the right-hand side to the left-hand side. Crucially, each model highlights the variables that are *not* directly causally related through the omission of a variable in the right-hand side of a structural equation or the absence of an arrow between two variables. As such both models contain a series of implications about the conditional independence in the joint probability distribution of the variables. If the variables are measurable this can be tested in the data. SEMs are becoming more widely used in educational data science, and so are a useful bridge into causal thinking. In particular, SEM path analysis is highly reminiscent of the approach of Pearl, once the causal DAG is parameterised. This leads to an important question: what is the difference between the two approaches? It is common to see a number of different claims that SEMs cannot handle nonlinear relationships, or are essentially equivalent to regression, but Bollen and Pearl (2013) provide a definitive argument that dispels many of these myths, arguing that the two modelling approaches are in fact equivalent, even though the foundations of causal inference are often considered more rigorous. While there has also been a reluctance from some to imbue SEM parameters with causal interpretations, perhaps made permissible by the symmetric interpretation of the "=" sign, this is not a necessity. Indeed, as discussed convincingly by Pearl (2012), the founders of SEM very much thought of them as ways to demonstrate causal relationships between variables under study.

Causal discovery

So far we have examined how modelling the causal structure of a system can provide testable implications for a dataset in the form of conditional independence relationships, which can then inform strategies for making causal inferences. It is also possible to move in the other direction, beginning with the conditional independence relationships found in a dataset. There are several algorithms for moving from observed conditional independence relationships to a causal model, using a process known as causal discovery (Pearl, 2009). Uncertainty in the inferred causal structure is handled by generating a class of equivalent causal models, all of which could potentially produce data with the same conditional independence relationships. This exploratory, rather than confirmatory, approach to causal modelling can scope the space of potential causal models that explain the data. The space of potential models can also help to inform future work into improving our theoretical understanding, as the uncertain edges or sections of the DAG indicate where better data or more experiments are needed. The subsequent results can then be fed back into the body of knowledge, with new data refining and improving the model.

Interestingly, a careful search of past learning science literature reveals an ongoing set of attempts to argue that causal models could help us to make better use of educational data, particularly in this causal discovery context. Perhaps the earliest example is a Masters thesis by Brokenshire (2007) that dates back to before the advent of both LA *and* Pearl's seminal book (Pearl, 2009). This work attempted to formalise our models of SRL through the use of discovery-based methods which extract a causal relationship from input data and even appeared in an AI in Education paper (Brokenshire & Kumar, 2009) but failed to find traction in the field. A later paper arguing that causal models would help LA practitioners to intervene in educational systems was written by one of Brokenshire's supervisors (Kumar et al., 2015), and explored models of writing analytics (WA) and metacognitive reasoning using this same causal discovery-based approach. However, this line of research never made it into the standard learning sciences literature, and appears to have halted. It is time to build on these early results. This paper aims to renew interest in this promising early work.

EXAMPLE 1: TURNING AN EDUCATIONAL PROTO-THEORY INTO A TESTABLE CAUSAL MODEL

Rather than following the causal discovery model of Brokenshire and Kumar, here we will explore a method more closely related to the human centred approach that first appeared in Hicks et al. (2022). In that paper we argued that the process of constructing a DAG can help data analysts to communicate with educational experts, meeting in the middle space of a field that all too often sees people from different backgrounds talking past one another. Thus, causal models can help us to *think more clearly* about educational data. In this first example, we will demonstrate how the attempt to construct a causal DAG can help an analyst to turn an educational proto-theory into a testable formal model. This approach works because the TCM implies that the move from a proto-theory to a formal theory requires an imposition of new constraints. We show here how the constraints of causal DAGs can help provide a scaffold for incrementally formalising a proto-theory.

To demonstrate our argument, we will make use of the Zimmerman (1989) model of SRL, which was also considered by Brokenshire (2007) as a part of his causal discovery method. We acknowledge that this is a simple SRL model, and that a number of more recent and detailed competing models have been proposed (Panadero, 2017). We believe that causal modelling could help to make these various hypothesised models explicit, taking them from a proto-theory stage to testable formal models. We have chosen the Zimmerman (1989) model merely for its simplicity in this example demonstration, and reserve a similar process of formalising the other models for future work.

Figure 2 has a representation of the proto-theory developed by Zimmerman. We define this as a proto-theory because of two characteristics that it possesses. First, it is explanatory, defining causal relationships between three constructs to explain the phenomena of students using various strategies to self-regulate their learning. However, it provides no precise model of how they will impact upon one another. Second, it is difficult to empirically test the hypothesised relationships between phenomena depicted by the nodes of the model, as a number of choices must be made in formalising them and the associated model of their behaviour such that they can be computationally represented and tested. In particular, a number of decisions need to be made about how the various feedback loops in Figure 2 should be represented in an empirically testable computational model.

How might DAGs be used to move towards a more formal theory that can be tested for explanatory adequacy as per the TCM? Rather than focussing upon the variables themselves, here we will focus upon the structure of the model proposed by Zimmerman. We start by noting that Figure 2 contains a causal loop in the form of the feedback cycle depicted from *Self* → *Behaviour* → *Environment* → *Self*. A causal DAG requires that these loops be untangled by sequencing the feedback process into two separate epochs where nodes update from one epoch to the next (see Figure 3). The first epoch ($t = 0$) includes the nodes

S_0 (Self at $t=0$), B_0 (Behaviour at $t=0$) and E_0 (Environment at $t=0$), and the second epoch ($t=1$) likewise includes the nodes S_1, B_1 and E_1. However, formalising this model requires a number of specific choices to be made, and the point at which we initiate a causal path can have a demonstrable impact upon the resulting model. In this case, reproducing the model of Figure 2 over two different epochs results in three different DAGs, depending upon which node we use to initiate the causal path. We now explore the implications of these possible formalisations, by drawing three different causal paths that initiate from S_0, E_0 or B_0 (see Figure 3a–c respectively). This will enable us to point to a possible battery of testable predictions that are made by each model, which in turn creates possibilities for empirically choosing between them as data are acquired. These tests will also help us to determine *what data we should collect*, to move forward in formalising the theory of SRL.

In formalising the model in Figure 2 we are required to be more specific about which variable influences which, and how those influences propagate over time. In each case there are five edges to draw in order to reproduce the structural relationships of the proto-theory: $S \rightarrow S$, $S \rightarrow B$, $B \rightarrow S$, $B \rightarrow E$ and $E \rightarrow S$. There are three ways in which this formalisation can be carried out:

- Figure 3a: Using "Self" (S_0) as the starting point (depicted here with a thicker boarder), we can draw the path $S_0 \rightarrow B_0 \rightarrow E_0$ within epoch $t = 0$ without creating any loops. From there, the paths $S_0 \rightarrow S_1$, $E_0 \rightarrow S_1$ and $B_0 \rightarrow S_1$ must be drawn from $t = 0$ to $t = 1$.
- Figure 3b: Using "Environment" (E_0) as the starting point lets us draw the path $E_0 \rightarrow S_0 \rightarrow B_0$ within epoch $t = 0$ without creating any loops. From there the paths $S_0 \rightarrow S_1$, $B_0 \rightarrow S_1$, and $B_0 \rightarrow E_1$ must be drawn from $t = 0$ to $t = 1$.
- Figure 3c: Using "Behaviour" (B_0) as the starting point lets us draw the paths $B_0 \rightarrow E_0 \rightarrow S_0$ and $B_0 \rightarrow S_0$ within epoch $t = 0$. From there the paths $S_0 \rightarrow S_1$ and $S_0 \rightarrow B_1$ must be drawn from $t = 0$ to $t = 1$.

Each formalised model generates its own set of testable implications, which are statements about the conditional independence of variables in the DAG. The three different approaches produce different testable implications, generated using the software tool DAGitty (Textor et al., 2016) which are listed in Table 2. No statement about conditional independence was shared amongst all three DAGs. This could be used to help test these theories by seeing

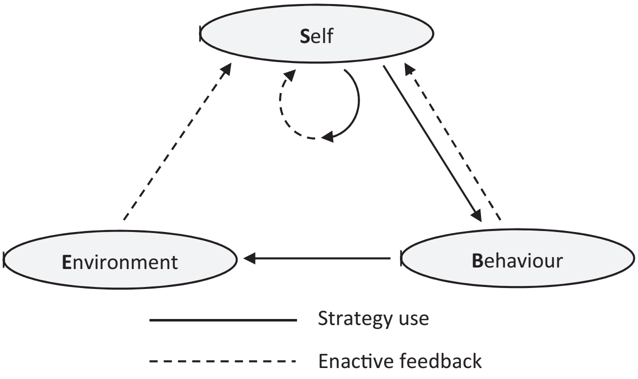

FIGURE 2 A triadic model of SRL, adapted from Zimmerman (1989).

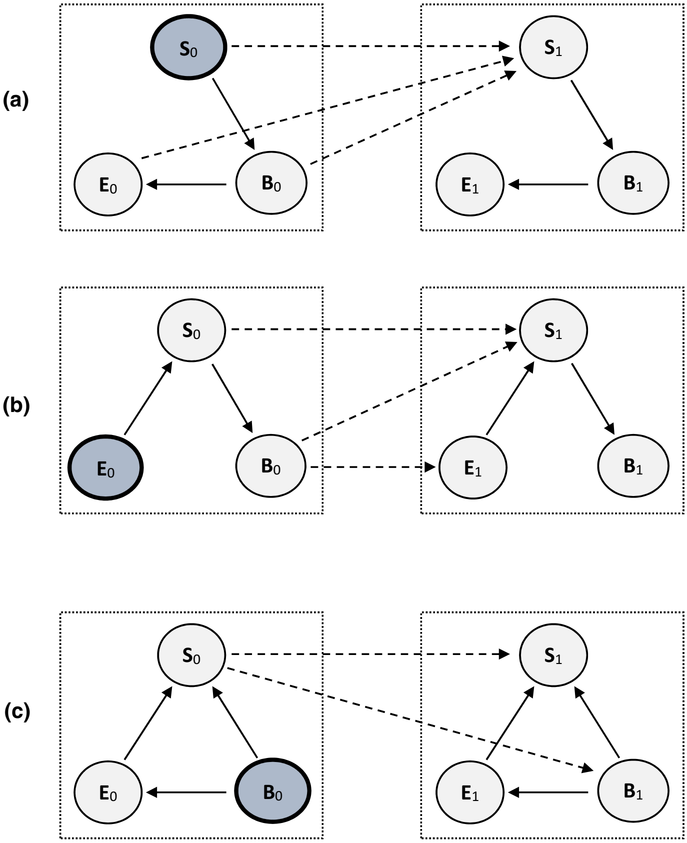

FIGURE 3 Formalised Causal DAGs of Zimmerman's triadic SRL model, using (a) ‘Self’, (b) ‘Environment’ or (c) ‘Behaviour’ as the initial node. Dashed lines correspond to the dashed lines in Figure 2.

how closely the implications of each DAG align with a dataset. Each statement of the form $A \perp\!\!\!\perp B \mid C$ (A and B are independent, given the value of C) can be tested by regressing both A and B on C and then examining the correlation between the residuals.

Framing the phenomena of SRL in the language of a formalised model helps to check the theory's capacity to explain the phenomena it hypothesises, using the implied conditional independence relationships. With data, we could start to test the implications of this model to see which of the proposed formalised models best fits the data. The S_0 initiating model matches Zimmerman's description of the ‘enactive feedback’ paths (the dashed paths in Figure 2), as these are the paths that traverse from one epoch to the next in Figure 3a. The E_0 initiating model makes intuitive sense, with the ‘environment’ preceding the other variables temporally which would indicate a causal influence of environment upon the other variables. For a similar reason the B_0 initiating model seems least intuitive; we must ask how a student's behaviour could influence the environment in which they are learning, a feat that is possible, but unlikely for this type of SRL model.

While this discussion centred around a toy example, we believe that the approach demonstrated shows promise for informing future work on SRL. Molenaar et al. (2022) have recently published an overview of how SRL has been measured over the past 5 years, tracking the emergence of Multimodal Learning Analytics as the source of a wide variety of data that could be used to model SRL. With the emergence of these new and more comprehen-

TABLE 2 List of testable implications from each DAG. Implications in bold are unique to that particular DAG.

Initiating at 'Person'	Initiating at 'Environment'	Initiating at 'Behaviour'
$\boldsymbol{S_0 \perp\!\!\!\perp E_0 \mid B_0}$	$B_0 \perp\!\!\!\perp B_1 \mid S_1$	$\boldsymbol{B_0 \perp\!\!\!\perp B_1 \mid S_0}$
$S_0 \perp\!\!\!\perp E_1 \mid B_1$	$\boldsymbol{B_0 \perp\!\!\!\perp E_0 \mid S_0}$	$B_0 \perp\!\!\!\perp E_1 \mid B_1$
$\boldsymbol{S_0 \perp\!\!\!\perp E_1 \mid S_1}$	$B_1 \perp\!\!\!\perp E_0 \mid S_0$	$\boldsymbol{B_0 \perp\!\!\!\perp E_1 \mid S_0}$
$S_0 \perp\!\!\!\perp B_1 \mid S_1$	$B_1 \perp\!\!\!\perp E_0 \mid S_1$	$\boldsymbol{B_0 \perp\!\!\!\perp S_1 \mid S_0}$
$\boldsymbol{S_1 \perp\!\!\!\perp E_1 \mid B_1}$	$\boldsymbol{B_1 \perp\!\!\!\perp E_1 \mid S_1}$	$B_1 \perp\!\!\!\perp E_0 \mid S_0$
$E_0 \perp\!\!\!\perp E_1 \mid B_1$	$B_1 \perp\!\!\!\perp S_0 \mid S_1$	$E_0 \perp\!\!\!\perp E_1 \mid B_1$
$\boldsymbol{E_0 \perp\!\!\!\perp E_1 \mid P_1}$	$\boldsymbol{E_0 \perp\!\!\!\perp E_1 \mid B_0}$	$E_0 \perp\!\!\!\perp E_1 \mid S_0$
$E_0 \perp\!\!\!\perp B_1 \mid S_1$	$E_0 \perp\!\!\!\perp E_1 \mid S_0$	$E_0 \perp\!\!\!\perp S_1 \mid S_0$
$B_0 \perp\!\!\!\perp E_1 \mid B_1$	$E_0 \perp\!\!\!\perp S_1 \mid S_0$	$E_1 \perp\!\!\!\perp S_0 \mid B_1$
$\boldsymbol{B_0 \perp\!\!\!\perp E_1 \mid S_1}$	$\boldsymbol{E_1 \perp\!\!\!\perp S_0 \mid B_0}$	
$B_0 \perp\!\!\!\perp B_1 \mid S_1$		

sive multimodal data streams, research on SRL is entering the age of big data. We believe that causal modelling has the potential to support theory generation by both informing this process of data collection, and then by supporting teams as they test between available theories of SRL based upon their collected data. We reserve this exciting avenue for future work.

In summary, the example in this section has demonstrated how moving towards a more formal model places us in a position where it becomes possible to ask more sophisticated questions, and to then direct our analysis by working to answer the resulting avenues that seem most promising.

EXAMPLE 2: TOWARDS A CAUSAL MODEL OF REFLECTIVE WRITING PERFORMANCE

How might we use causal models to enhance and extend an existing LA model? The previous section's analysis helps us to identify places where a proto-theory could be made more precise, providing testable predictions along the way. In this section we shall make use of a previous model of reflective writing performance, demonstrating how causal models can help us to formulate a set of new more precise hypotheses and predictions about what future data collection might demonstrate.

Within LA, the subfields of Automated Writing Evaluation (AWE) (Shermis & Burstein, 2013) and Writing Analytics (WA) (Gibson & Shibani, 2022) focus on the automated analysis of written texts for the purpose of generating automated feedback to support personal learning. A stream of activity within this body of work centres around Reflective Writing Analytics (RWA) which seeks to identify reflective elements in student generated texts (Gibson et al., 2016; Buckingham Shum et al., 2017; Gibson et al., 2017; Kovanović et al., 2018; Ullmann, 2019; Barthakur et al., 2022)

Recent work has sought to develop a learning progression scale that can be used to measure the depth of reflection in writing, tracking individual changes and differences between learners. Liu et al. (2021) hypothesised a formal model of reflective writing capability consisting of four sub-variables: *Context*; *Feelings*; *Challenges* and *Changes* that was based upon an extension of the model proposed by Gibson et al. (2017). A SEM, in the form of a Confirmatory Factor Analysis (CFA) was constructed to quantify the relative contributions that (i) textual features make to reflection factors and (ii) reflection factors make to the overall depth of a student's written reflection. The validity of the model was evaluated using reflections from two sets of masters-level reflective writing assessments in two different fields (Pharmacy and Data Science). The results of this analysis are depicted in Figure 4, and demonstrated that a number of the automatically extracted textual features in students' reflective writing were robustly present across the two different educational contexts. These features were shown to correlate to the latent sub-variables, and thus to contribute to the latent *Capability for Written Reflection* variable, which was itself shown to correlate with the grades which were awarded to the students. However, a few of the low-level traces extracted from the student reflections were specific to the learning context: *LIWC.percept*, *LIWC.focusfuture* and *LIWC.Analytic*. Note also that the *Feelings* variable was considered non-significant in the Data Science context. Thus, the two resulting SEMs provide a number of insights about *Capability for Written Reflection*, and how it might be measured in the reflective text that a student generates.

The CFA models in Figure 4 are already quite formal, connecting theory (reflective writing constructs) to data (automatically extracted textual features in student writing) in a way that quantifies the relative weights of association. They make clear statements about what influences the quality of a student's reflection, and how these can be computationally extracted from the text a student generates using natural language processing. However, we have two sets of results obtained from fitting variables to our data, and our model currently makes no claims about *why* the differences between the two different educational scenarios occurred. To improve this situation, it would help to develop a causal theoretical model that could be applied across both educational contexts, checked for its explanatory adequacy across new datasets, and evaluated against a set of well-defined criteria for its predictive power (Borsboom et al., 2021).

To move towards this more formal causal model we can start by generating the DAG depicted in Figure 5. Here, we see that the arrows from the intermediate latent variables to the *Capability For Written Reflection* variable have been reversed (when compared to Figure 4), representing our belief that they causally determine it (and not the other way around). We have also added a measurable variable, *Reflection Quality*, which explicitly links to the measurable grade that a student obtains for their reflection. To join the two models from Figure 4, we have taken account of the two different assessment models that were present in the task itself. The Data Science assessment prompt asked students to write a ‘performance review’ and had no focus upon feelings at all, but in contrast the Pharmacy task included a specific prompt to explain how students felt during their placement (Liu et al., 2021 provide more details about the assessments of these two tasks). We note that both models in Figure 4 link the *Feelings* variable only weakly to the quality of a reflection (and in the Data Science context this link is not significant), suggesting that a causal chain through this variable is unlikely to have a strong effect upon *Capability for Written Reflection*. Thus, the difference in reflection quality was not well captured by the *Feelings* variable, but rather in the textual features associated with the *Challenges* and *Changes* variables. Our new extended model hypothesises that it is a new latent variable, *Educators intent*, which influences an observable *Assessment document*, and so frames students' understanding of the task, and hence the quality of their written reflections through the *Challenges* and *Changes* pathways, as depicted in Figure 5. If a new reflective writing dataset was generated, with a different assessment design, this new model would help us to make predictions about what features in a student's reflective writing might change, and how we would expect these to affect the quality of the reflection so produced.

To summarise, using DAGs to generate Figure 5 has enabled us to take an additional step in formalising our theory, introducing a construct and associated measurable from a broader learning context than just the student's reflective capabilities and performance (via

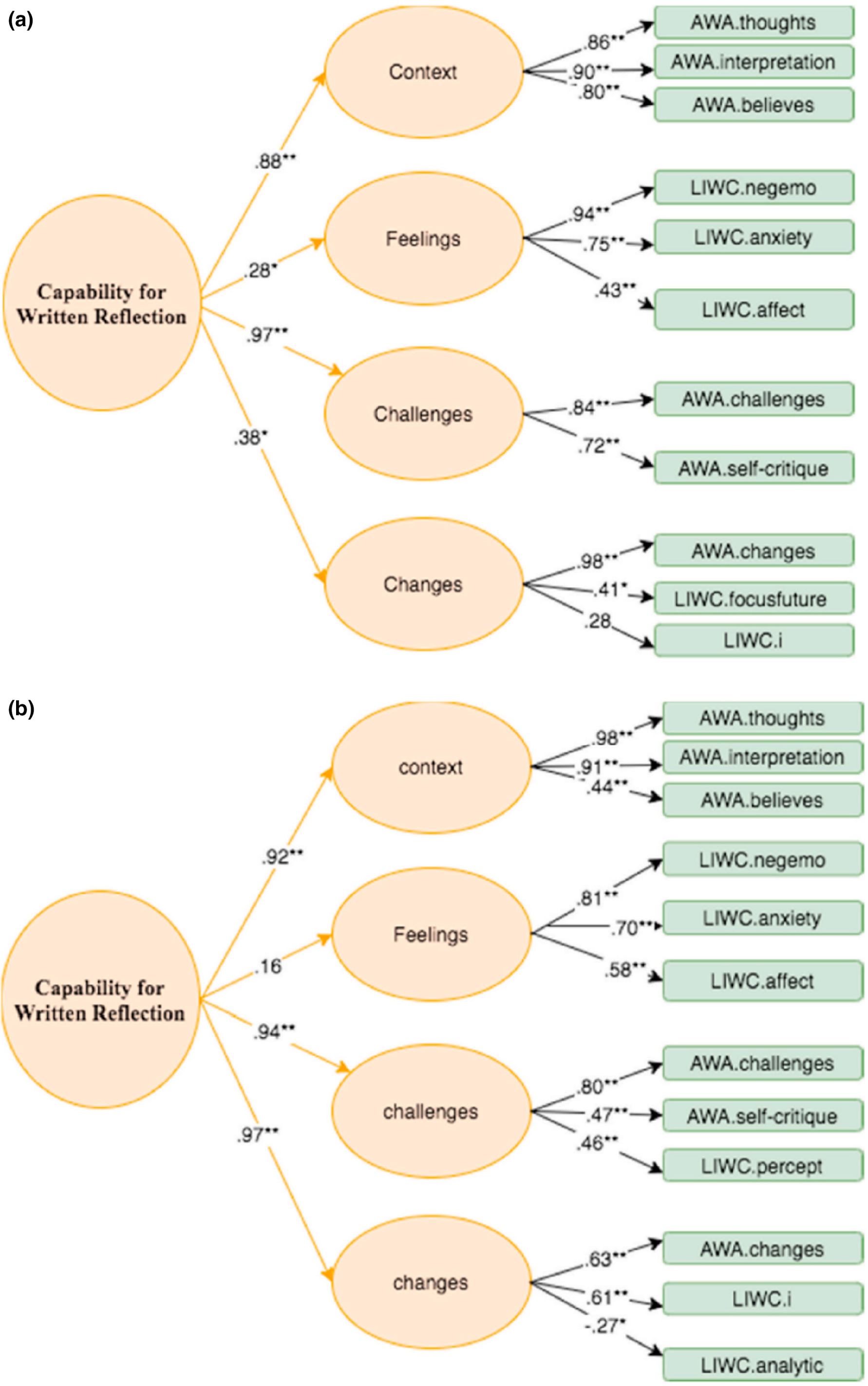

FIGURE 4 The Confirmatory factor analysis models of capability for written reflection that emerged from the (a) pharmacy dataset and (b) data science dataset. These quantify the relative contributions that textual features make to reflection factors, and to the overall depth of written reflection (reproduced from Liu et al., 2021).

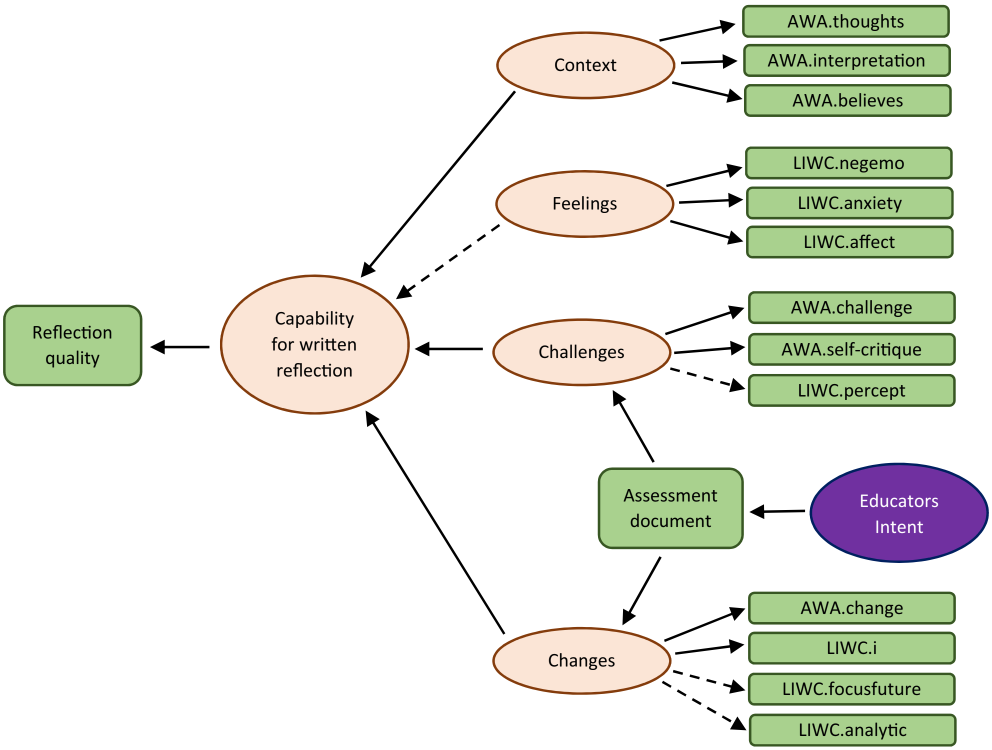

FIGURE 5 Our proposed new Directed Acyclic Graph for written reflection, adding a measurable node for *reflection quality* (operationalised as its grade), and for *Assessment document* to account for the two contexts. Arrows are redrawn to depict the flow of causation, and dashed paths indicate that an effect was significant in one CFA but not the other.

Educator's intent which shapes the *Assessment document*). Our existing data enable us to hypothesise that this pathway has impacted the student reflections, and in particular the textual features that they contain, in addition to the quality of the resulting reflections. We now have the start of a hybrid sociotechnical systems model, incorporating a theoretically grounded, formal model of reflective writing that makes explicit a 'theory of change' behind an intervention around assessment design. This theory makes testable predictions about how student writing will change in response to assessment design. The machinery of causal modelling will then enable more in-depth probing of this extended theory through the implied conditional independence relationships of the DAG. For instance, controlling for the new *Assessment Document* variable should result in weaker association between the latent constructs *Challenges* and *Changes* (due to the *fork* pattern emanating from *Assessment Document*), but not between *Feelings* or *Context* (due to the collider at *Capability for written reflection*). This can be tested in the data, or through experiment, to validate or refute the model.

DISCUSSION: CAUSAL MODELS AS A CONCEPTUAL FRAMEWORK TO ASSIST WITH THEORY DEVELOPMENT

The previous two examples illustrate the utility of striving for a causal understanding of our systems, with DAGs providing a mechanism for making explicit claims about how causality flows through a model. Causal models rendered as DAGs can assist us, not only conceptually but also quantitatively, to map the educational constructs developed by theorists to the data that we can now collect. Our claim is that causal models offer the learning sciences a bridge that will help us to develop more rigorous and testable educational theories. We now turn to a discussion of the implications that we consider most interesting about this approach.

Theory versus data in the domain of education: The clicks to constructs problem

We commenced this paper by discussing the large amount of data now available in education. However, big data does not necessarily equate with useful data. Historically, the clickstream data collected from educational environments were often first instrumented to aid developers in debugging software. This means that it is rarely well structured for describing educationally relevant phenomena. All too frequently, educational institutions have expended significant effort on storing large volumes of data, only to discover that many of the variables required for modelling educational phenomena are missing when the time finally comes to analyse it for actionable insights (Kitto et al., 2020). Applying data science methods to this type of data can be seen as following in the big data tradition advocated by Norvig. However, there is no guarantee that the specific features of machine-perceptible human behaviour being logged are plausible proxies for a *learning* phenomenon, as represented by an educational theory. While this approach can sometimes yield interesting insights in education, it more frequently leads to results that are either obvious to educators and already well understood, or impossible to link to concepts that the field can interpret and test. A mapping is often necessary between these two representational layers, but this can be challenging to generate after the application of a brute force approach. With no guarantee that the *right* data have been collected, the analyst is in danger of fixating upon variables that have no relevance to the field of education. Instead these variables are quite likely to be artefacts of the EdTech system in which the data were collected.

In LA, this challenge of mapping the low-level ‘data exhaust’ left on platforms to the higher order constructs that are the discourse of theory and pedagogy is increasingly referred to as the ‘clicks to constructs’ problem (Buckingham Shum & Crick, 2016). Wise et al. (2021) have provided a schematic that summarises the concept visually (see Figure 6). It is widely recognised that these types of models are necessary to assist the field in *designing* for the capture of useful data. However, despite a general recognition of its importance, modelling approaches that facilitate the mapping of educationally relevant constructs to low-level digital traces (and vice versa) are still difficult to find.

In this paper we have demonstrated how the apparatus of causal reasoning and DAGs forces us to be explicit about the relationships that we hypothesise exist between educational constructs and their measurable proxies, so providing a mechanism that helps us link clicks to constructs. Diagrams like Figure 6 are naturally generated in the construction of a DAG. Thus, causal models show promise for helping us to ensure that we only collect the data we need, and can clearly articulate *why* we need it if challenged.

Thinking explicitly about complex educational constructs to formalise theoretical representations

The rigorous nature of the causal modelling apparatus forces us to be very precise in formulating a model. We must think explicitly about the complex educational constructs that we are constructing in our models, working carefully to specify the decisions that we are making and their statistical implications. As such this approach can assist with removing some of

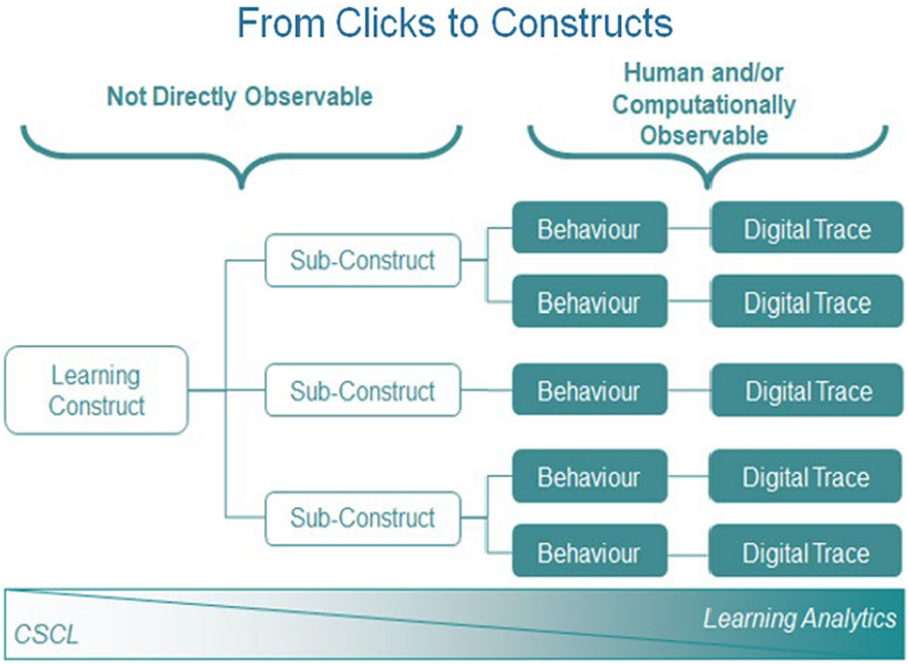

FIGURE 6 Schematic diagram of the 'clicks to constructs' challenge in the context of collaborative learning analytics (reproduced from Wise et al., 2021).

the ambiguity inherent in the proto-theories that often arise in the learning sciences. Causal modelling enables us to move towards more formal theories that can be experimentally tested (Borsboom et al., 2021). This increase in rigour comes with two significant advantages. First, more rigorously formulated models are easier to communicate. Their explicit nature means that the specific assumptions and hypotheses that they make can be explored by other researchers and challenged if considered inappropriate. Second, the use of well-defined theoretical constructs makes the resulting data analysis less prone to a potential failure to replicate. This is because an analysis that starts with the construction of a DAG is explicitly linked to empirical constructs that are derived from our theoretical understanding of a system. The resulting models are less brittle, as they are less likely to result from overfitting to the data that are available. Thus, thinking explicitly about the constructs in our model and how they are causally related often forces us to reject data which do not fit into that model, and can even tell us that data of a certain type might be missing.

It is particularly interesting to consider the lessons we have learned from the 'no loop' requirement of DAGs in Example 1. The models that resulted from applying this requirement (in Figure 3) could be considered more complex than that of the original Zimmerman model as that proto-theory consisted of only three nodes. A critic might challenge the DAG-based approach on the basis of the complexity it is likely to induce in more sophisticated educational proto-theories. The learning sciences contain many loops and complex feedback cycles, and so it could be claimed that the acylic requirement of DAGs is too strict, and that this will limit their potential domain of application. We acknowledge that this is a potentially legitimate concern. However, there are two ways in which we would like to respond to this critique. First, it is important to recognise that this complexity is *already* present in a model with cycles. It is abstracted into a diagram, but the presence of a feedback cycle necessarily adds to the model complexity. In converting this model to a DAG we have explicitly recognised this complexity, which enables us to control for it in the modelling process. Second, we would like to emphasise that we do not by any means suggest that causal modelling should be the *only* approach adopted by the learning sciences for theory development. Rather, that it can be a *useful* method to apply that can assist with theory development by forcing us to be more explicit in our modelling. Sometimes however, this very explicitness may be detrimental, and other modelling approaches may be more appropriate. Other methods are available, and indeed, Peters (2017, p. 28) suggests that if feedback behaviour is the primary area of interest then a dynamic systems approach may be more appropriate. Nonetheless, in attempting to more formally model our educational proto-theory we have learned that it is still possible to make use of the causal apparatus in a situation with a loop. Further, working to explicitly specify how a causal influence flows from one epoch to the next has helped us to make a number of claims about conditional independence in our system that are potentially testable. While the resulting formal model is structurally more complex than the proto-theory, in this case we consider it more informative. It is also more testable using computational models and data analysis.

It would be beneficial if tools were developed that enabled both views of a system. That is, we can envisage a scenario where a model with loops constructed by subject matter experts could be further refined in a tool that asked for clarification about the source of a causal flow over epochs that disentangled this loop structure. Toggling between both views of the system (ie, the loop and the causal DAG) would enable various users to explore the implications of their hypotheses, and facilitate theory building while minimising the need to view the resulting models in their full complexity. We consider this an avenue likely to be fruitful in future tool development.

Causal modelling as an aid to interdisciplinary theory building in the learning sciences

One of the key challenges facing interdisciplinary fields such as the learning sciences is the extreme diversity that arises, in disciplinary assumptions, languages, methodologies and symbolic representations, as people from a wide range of fields interact and collaborate. When researchers from the social sciences interact with those from data science and AI there is much room for misunderstandings to occur. We must learn to build common ground across these highly diverse fields, but the language of data modelling and statistics often feels incompatible with many of the research methods that are used in the learning sciences. Moving from proto-theories to more formal theories requires symbolic representations that can assist in the joint sensemaking that has to occur between all stakeholders, and DAGs provide an intuitively graspable artefact that supports this form of communication between disciplinary perspectives. Elsewhere, we have argued in more detail that DAGs, from informal sketches to more rigorous models, offer visual affordances that scaffold conversations between data and educational experts, comparable to the way that concept maps and other diagramming schemes relieve cognitive load by providing a form of shared, persistent, but malleable, external memory aid (Hicks et al., 2022). Similarly, the recent work by Weidlich et al. (2022) demonstrates how causal models can help modellers and educators to work together to identify sources of bias in various models, and ways in which they might be ameliorated. In attempting to construct a DAG we will sometimes construct competing models that are empirically testable, and these can be interrogated and challenged by educational experts. Indeed, if multiple models emerge from an attempt to construct a DAG, then this can be seen as a measure of our underlying uncertainty about the educational system under study (Boerebach et al., 2013), which in turn expresses a need for further research and data collection. The emergence of an artefact that can be interrogated by researchers without expertise in the data sciences therefore offers promise for providing a more equitable conversation between experts from all domains. Indeed, we consider the potential benefits of causal modelling for interdisciplinary communication to be one of its least discussed but most compelling advantages.

Why has causal inference failed to gain traction in the learning sciences?

It is interesting to note that causal modelling was first used to analyse educational data and make causal inferences more than 15 years ago, before LA had even become a field of study (Brokenshire, 2007; Brokenshire & Kumar, 2009). And yet this approach failed to gain traction at the time—why so? We consider it likely that it was the sheer innovativeness of this approach when first introduced to the field which made it difficult for this method to thrive. Three problems are likely to have hampered its broader adoption.

1. *A lack of expertise*: As a very new and quite mathematically sophisticated method, this was a difficult approach for non-experts to apply. The classic book on the topic (Pearl, 2009) had not even been published, and few people were aware of causal modelling as an alternative method for making causal claims.
2. *A lack of data*: Similarly, far less educational data were available at the time of this early work, which made it far more difficult to implement these models in authentic contexts. As such, the complexity required to formulate DAGs and then implement statistical tests of their predictions was difficult to justify.
3. *A lack of tools*: As a highly novel method, many of the tools now available for performing causal analysis had yet to be developed, making it much harder to implement this approach. Recent advances, such as DAGitty for turning DAGs into statistical models (Textor et al., 2016), and the *DoWhy* python library (Sharma & Kiciman, 2019), which is built to support end to end causal modelling, will continue to accelerate the adoption of these methods across data science.

Each of these issues is far less problematic today, and for this reason we believe that the time has come to develop a stream of research applying causal inference methods, especially causal modelling, to the learning sciences. With more resources available for learning about causal inference, far more educational data, and new more intuitive tools, this method is a promising candidate for helping us to bridge between theory and data, one which we consider ripe for further development in the toolkit of the learning sciences.

CONCLUSION

Theory is a difficult concept, one that has many different meanings across the wide range of fields that have fed into the learning sciences. Here we have argued that causal modelling can help education with theory construction, by providing a rigorous approach for turning a number of educational proto-theories into more formalised models that are ready for testing with data. Thus, causal modelling provides a well-defined avenue for linking between theory and data, one that enables us to give prominence to both, moving beyond a largely futile debate between the two cultures of statistical thinking in the process. This approach helps us to escape the common assumption that RCTs are the only way in which to demonstrate causal influences between variables. It also helps us to determine what data we should collect, and why, through the formulation of explicit predictions in our educational models. In short, the time has come for the learning sciences to take causal modelling very seriously indeed, as we are now in a position where causal models can be used to work towards a more rigorous, reproducible, and communicable notion of theory in the field. We encourage others to pursue this promising line of work.

ACKNOWLEDGMENTS

The authors have no acknowledgements to make.

Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

None.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest associated with this work.

DATA AVAILABILITY STATEMENT

This paper is not based upon empirical data and so no data are available.

ETHICS STATEMENT

This paper is theoretical, and so was not conducted under ethical clearance.

ORCID

Kirsty Kitto https://orcid.org/0000-0001-7642-7121
Ben Hicks https://orcid.org/0000-0003-4062-6035
Simon Buckingham Shum https://orcid.org/0000-0002-6334-7429

ENDNOTES

[1] Google's Director of Research and co-author of a very popular AI textbook that is now in its 4th edition (Russell & Norvig, 2022).

[2] One of the most influential linguists in the world.

REFERENCES

Albacete, P., Jordan, P., Katz, S., Chounta, I. A., & McLaren, B. M. (2019). The impact of student model updates on contingent scaffolding in a natural-language tutoring system. In *International Conference on Artificial Intelligence in Education* (pp. 37–47). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_4

Barthakur, A., Joksimovic, S., Kovanovic, V., Mello, R. F., Taylor, M., Richey, M., & Pardo, A. (2022). Understanding depth of reflective writing in workplace learning assessments using machine learning classification. *IEEE Transactions on Learning Technologies*, *15*(5), 567–578. https://doi.org/10.1109/TLT.2022.3162546

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3442188.3445922

Boaler, J., Dieckmann, J. A., Pérez-Núñez, G., Sun, K. L., & Williams, C. (2018). Changing students minds and achievement in mathematics: The impact of a free online student course. *Frontiers in Education*, *3*, 1–7. https://doi.org/10.3389/feduc.2018.00026

Boerebach, B. C., Lombarts, K. M., Scherpbier, A. J., & Arah, O. A. (2013). 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. https://doi.org/10.1371/journal.pone.0069449

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research* (pp. 301–328). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6094-3_15

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brokenshire, D. (2007). *Discovering causal models of self-regulated learning* (Masters thesis). Simon Fraser University. https://central.bac-lac.gc.ca/.item?id=MR41042&op=pdf&app=Library&oclc_number=667804408

Brokenshire, D., & Kumar, V. (2009). Discovering causal models of self-regulated learning. In *Artificial Intelligence in education: Building learning systems that care: From knowledge representation to affective modelling* (Vol. 200, pp. 257–264). https://doi.org/10.3233/978-1-60750-028-5-257

Brooks, C., Gardner, J., & Chen, K. (2018). How gender cues in educational video impact participation and retention. In J. Kay & R. Luckin (Eds.), *Proceedings of the International Conference of the Learning Sciences (ICLS)* (pp. 1835–1842). ISLS.

Buckingham Shum, S., & Crick, R. D. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics*, *3*(2), 6–21. https://doi.org/10.18608/jla.2016.32.2

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: Rationale, methodology and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. https://doi.org/10.18608/jla.2017.41.5

Cass, S. (2011). *Unthinking machines*. MIT Technology Review. https://www.technologyreview.com/s/423917/unthinking-machines/

Davis, B., & Sumara, D. (2010). 'If things were simple…': Complexity in education. *Journal of Evaluation in Clinical Practice*, *16*(4), 856–860. https://doi.org/10.1111/j.1365-2753.2010.01499.x

Deho, O. B., Liu, L., Joksimovic, S., Li, J., Zhan, C., & Liu, J. (2022). Assessing the causal impact of online instruction due to COVID-19 on Students' grades and its aftermath on grade prediction models. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (pp. 32–38). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3524458.3547232

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—A "garden of forking paths"—Explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460. https://doi.org/10.1511/2014.111.460

Gibson, A., & Shibani, A. (2022). Natural language processing-writing analytics. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *The Handbook of Learning Analytics*, Society for Learning Analytics Research (SoLAR) (2nd ed., 96–104). Vancouver, Canada.

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 153–162). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3027385.3027436

Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics*, *3*(2), 22–36. https://doi.org/10.18608/jla.2016.32.3

Guzmán-Valenzuela, C., Gómez-González, C., Tagle, A. R. M., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: A preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education*, *18*(1), 1–19. https://doi.org/10.1186/s41239-021-00258-x

Hartnett, K. (2018). To build truly intelligent machines, teach them cause and effect. *Quanta Magazine*. https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

Hicks, B., Kitto, K., Payne, L., & Buckingham Shum, S. (2022). Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 250–259). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3506860.3506899

Hossain, M. S., Islam, M. S., Glinsky, J. V., Lowe, R., Lowe, T., & Harvey, L. A. (2015). A massive open online course (MOOC) can be used to teach physiotherapy students about spinal cord injuries: A randomised trial. *Journal of Physiotherapy*, *61*(1), 21–27. https://doi.org/10.1016/j.jphys.2014.09.008

Jamison, J. (2019). The entry of randomized assignment into the social sciences. *Journal of Causal Inference*, *7*(1), 20170025. https://doi.org/10.1515/jci-2017-0025

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

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

Jovanovic, J., Mirriahi, N., Gašević, D., Dawson, S., & Pardo, A. (2019). Predictive power of regularity of pre-class activities in a flipped classroom. *Computers & Education*, *134*, 156–168. https://doi.org/10.1016/j.compedu.2019.02.011

Khalil, M., Prinsloo, P., & Slade, S. (2022). The use and application of learning theory in learning analytics: A scoping review. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-022-09340-3

Kitto, K., Manly, K., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 303–314), March 13–17. New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3576050.3576096

Kitto, K., Whitmer, J., Silvers, A., & Webb, M. (2020). *Creating data for learning analytics ecosystems*. Position Paper, Society for Learning Analytics Research. https://www.solaresearch.org/core/creating-data-for-learning-analytics-ecosystems/

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkey, S., Williams, J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900–14905. https://doi.org/10.1073/pnas.1921417117

Knight, S., & Buckingham Shum, S. (2017). Theory and learning analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gaevic (Eds.), *The handbook of learning analytics* (1st ed., pp. 17–22). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.001

Knight, S., Buckingham Shum, S., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, *1*(2), 23–47. https://doi.org/10.18608/jla.2014.12.3

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge—LAK '18* (pp. 389–398). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3170358.3170374

Kumar, V. S., Clemens, C., & Harris, S. (2015). Causal models and big data learning analytics. In *Ubiquitous learning environments and technologies* (pp. 31–53). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44659-1_3

Liu, M., Kitto, K., & Buckingham Shum, S. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior*, *120*(106), 733. https://doi.org/10.1016/j.chb.2021.106733

Lodge, J. M., Knight, S., & Kitto, K. (forthcoming). Theory and learning analytics, a historical perspective. In K. Bartimote, S. K. Howard, & D. Gašević (Eds.), *Theory informing and arising from learning analytics*. New York: Springer.

Lübke, K., Gehrke, M., Horst, J., & Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, *28*(2), 133–139. https://doi.org/10.1080/10691898.2020.1752859

Marzouk, Z., Rakovic, M., Liaqat, A., Vytasek, J., Samadi, D., Stewart-Alonso, J., Ram, I., Woloshen, S., Winnie, P., & Nesbit, J. C. (2016). What if learning analytics were based on learning science? *Australasian Journal of Educational Technology*, *32*(6), 1–18. https://doi.org/10.14742/ajet.3058

Molenaar, I., de Mooij, S., Azevedo, R., Bannertd, M., Järveläe, S., & Gaševićf, D. (2022). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, *107*, 540. https://doi.org/10.1016/j.chb.2022.107540

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

Murphy, P. K., & Knight, S. L. (2016). Exploring a century of advancements in the science of learning. *Review of Research in Education*, *40*(1), 402–456. https://doi.org/10.3102/0091732x16677020

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, *9*(4), 30–33. https://doi.org/10.1111/j.1740-9713.2012.00590.x

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Pearl, J. (2009). *Causality*. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. Guilford Press. https://doi.org/10.21236/ada557445

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pearson, K. (1911). *The grammar of science* (3rd ed.). Adam and Charles Black.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms* (288). The MIT Press.

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2021). Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education. *Postdigital Science and Education*, *3*(3), 715–742. https://doi.org/10.1007/s42438-021-00249-1

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th Global ed.). Pearson.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*(3), 1–22. https://doi.org/10.18637/jss.v035.i03

Selwyn, N., O'Neill, C., Smith, G., Andrejevic, M., & Gu, X. (2021). A necessary evil? The rise of online exam proctoring in Australian universities. *Media International Australia*, *186*, 149–164. https://doi.org/10.1177/1329878X211005862

Sharma, A., & Kiciman, E. (2019). *DoWhy: A Python package for causal inference*. https://github.com/microsoft/dowhy

Shermis, M. D., & J. Burstein (Eds.) (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge. https://doi.org/10.4324/9780203122761

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Robyn, A., Baird, M. D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G. P., & Chambers, J. (2018). Improving teaching effectiveness. In *Final report: The intensive partnerships for effective teaching through 2015–2016*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2242.html. https://doi.org/10.7249/rr2242

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

Suthers, D., & Verbert, K. (2013). Learning analytics as a “middle space”. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)* (pp. 1–4). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/2460296.2460298

Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, *40*(3), 371–384. https://doi.org/10.2307/2393788

Tempelaar, D., Rienties, B., Mittelmeier, J., & Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior*, *78*, 408–420. https://doi.org/10.1016/j.chb.2017.08.010

Textor, J., Van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package ‘dagitty’. *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*, 217–257. https://doi.org/10.1007/s40593-019-00174-2

UNESCO. (2020). *Education for sustainable development: a roadmap*. ISBN: 978-92-3-100394-3. https://unesdoc.unesco.org/ark:/48223/pf0000374802.locale=en

Vilkova, K. (2022). The promises and pitfalls of self-regulated learning interventions in MOOCs. *Technology, Knowledge and Learning*, *27*(3), 689–705. https://doi.org/10.1007/s10758-021-09580-9

Weidlich, J., Gašević, D., & Drachsler, H. (2022). Causal inference and bias in learning analytics: A primer on pitfalls using directed acyclic graphs. *Journal of Learning Analytics*, *9*(3), 183–199. https://doi.org/10.18608/jla.2022.7577

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*(106), 457. https://doi.org/10.1016/j.chb.2020.106457

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

Wise, A., Knight, S., & Buckingham Shum, S. (2021). Collaborative learning analytics. In U. Cress, C. Rosé, A. Wise, & J. Oshima (Eds.), *International Handbook of Computer-Supported Collaborative Learning*. Springer. https://doi.org/10.1007/978-3-030-65291-3

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

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, *81*, 329–339. https://doi.org/10.1037/0022-0663.81.3.329

How to cite this article: Kitto, K., Hicks, B., & Buckingham Shum, S. (2023). Using causal models to bridge the divide between big data and educational theory. *British Journal of Educational Technology*, *54*, 1095–1124. https://doi.org/10.1111/bjet.13321