DISCOVERING CAUSAL MODELS OF SELF-REGULATED LEARNING

by

David Brokenshire
B.Sc., Simon Fraser University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the School
of
Interactive Arts and Technology

© David Brokenshire 2007
SIMON FRASER UNIVERSITY
September 2007

All rights reserved. This work may not be
reproduced in whole or in part, by photocopy
or other means, without the permission of the author.

APPROVAL

Name: David Brokenshire

Degree: Master of Science

Title of thesis: Discovering Causal Models of Self-Regulated Learning

Examining Committee: Dr. John Bowes
Chair

Dr. Marek Hatala, Senior Supervisor

Dr. Vive Kumar, Supervisor

Dr. Phil Winne, External Examiner,
Professor, Education,
Simon Fraser University

Date Approved: _____

Abstract

New statistical methods allow discovery of causal models from observational data in some circumstances. These models permit both probabilistic inference and causal inference for models of reasonable size. Many domains, such as education, can benefit from such methods.

Educational research does not easily lend itself to experimental investigation. Research in laboratories is artificial and potentially affects measurement; research in authentic environments is extremely complex and difficult to control. In both environments, the variables are typically hidden and only change over the long term, making them challenging and expensive to investigate experimentally.

I present an analysis of causal discovery algorithms and their applicability to educational research, an engineered causal model of Self-Regulated Learning (SRL) theory based on the literature, and an evaluation of the potential for discovering such a theory from observational data using the new statistical methods and suggest possible benefits of such work.

To my wife

This too shall pass
Proverb

Acknowledgments

There are many people who have made it possible for me to complete this work. First and foremost I would like to thank my wonderful wife Brittney, without whom I would never have made it this far. Her care and understanding as I have worked on this thesis have kept me going and her brilliance and dedication has inspired me. Of course, I would like to think my parents for their ongoing support and understanding throughout my education, despite the dramatic ups and downs I have had. My supervisors Marek Hatala and Vive Kumar for their endless help, encouragement, and support throughout this long process, and their understanding about how long the process ended up being.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction & Motivation

People make constant use of causal relationships in our everyday lives. We recognize many different types of causes, be they mechanical, psychological, or historical. We understand mechanical causes, like pressing on a gas pedal causes the car to accelerate and move; we understand psychological causes, that being mean to someone causes them to become angry and defensive. We understand historical causes, that I caused myself to fail a class because I didn't go to class and study and that things would have been different if I'd only not gone to that last party.

Scientists also make use of causal relationships. Medical researchers evaluate whether a drug causes a patient to recover, economists study the causes of successful economies, and chemists study what chemicals and conditions cause a reaction.

However, the methods of regular people and scientists differ considerably. Scientists perform experiments to discover causes, whereas proper scientific experiments are not available to ordinary people in their day to day lives. People do 'experiment' in the colloquial sense, we try new things, change one thing at a time and see what happens, but we do not do experiments in the way that scientists do. Scientists do *randomized controlled experiments* and regard them as the only reliable way to discover causal relationships.

Scientists typically do not do randomized controlled experiments in their every day lives though, because they, like everyone else, do not have the resources to do randomized controlled experiments for all things, even if it were possible, or sensible. Doing a randomized controlled experiment to test the effect of a hot element burning your hand would be frowned upon by most. Somehow though, people learn many reliable causal relationships without doing randomized controlled experiments. They do a combination of 'observational studies'

i.e. looking at the world and thinking, and 'quasi-experimental studies' where they look at the world, change something, and then look at the world some more, and draw some conclusions.

We are often wrong about the causal relationships we assume in everyday life. Scientists are wrong about relationships too, but they have formal methods of evaluating the likelihood that they are right or wrong. The everyday behaviour of individuals lacks these formal mathematical methods.

The inability to always do randomized controlled experiments isn't just a problem for everyday life, but for scientists. We cannot always do randomized controlled experiments. We cannot do experiments to see if smoking 'really' causes cancer for ethical reasons, and we cannot do experiments on lots of things in the social sciences for practical reasons, but we still seek to discover such relationships.

People manage to learn causal relationships all the time, and it seems natural to try to formally characterize how to discover causal relationships from observational data. This would allow us to know with some certainty when we are right, when we are wrong, and how likely each is.

The classic response to such an attempt is to assert that correlation does not imply causation, and logically it does not. However, the impossibility of discovering causal relationships from observational data in any circumstance does not follow. Over the last two decades several groups of researchers, in particular Judea Pearl and colleagues at UCLA and Peter Spirtes group at CMU, as well as David Heckerman at Microsoft Research have begun to answer this question by making major advances in formally representing the notion of causality, providing mathematical manipulations of it, and algorithms for discovering causal relationships from data.

These advances now allow us to discover some causal relationships formally, given certain assumptions, from purely observational data. We can also characterize what relationships can be learned in this way, and what assumptions are necessary. They also allow us to evaluate when, how many, and which experiments are necessary to discover causal relationships we cannot derive from observational data alone.

In addition to their formal properties, these causal models are graphical, and are reasonably easy for people to understand. They allow the compact representation of the causal claims made in a field for easier understanding by researchers in that field.

These methods, though relatively new and still developing, have many possible areas

of application, and have begun to be applied in epidemiology, psychology, and the social sciences, as well as in Artificial Intelligence and its applications. In this thesis I evaluate how they may be applied to research in education, specifically Self-Regulated Learning.

1.1 Background

1.1.1 Self-Regulated Learning

Self-Regulated Learning (SRL) theory attempts to explain academic learning and achievement of learners in terms of various characteristics and processes individuals use to regulate their own behaviour. It emphasizes the student as an active participant in the learning process, as opposed to a passive recipient of information provided by a teacher.

There are several different theoretical perspectives in SRL which draw inspiration from different areas of scholarship and offer different explanations for the relationships between the key variables as well as their relative importance. Most perspectives agree that key factors include the learner's motivation, goals, self-monitoring, volitional strategies, and self-evaluation and reflection behaviours.

The popular models view SRL as a three or four stage cycle of broadly defined phases through which a learner passes repeatedly as they perform an academic task. Within each phase learners perform a variety of actions and employ skills to regulate and improve their learning behaviour. A wide variety of observational measures have been used to evaluate SRL and interventional studies have been conducted with positive results with some of these results being incorporated into mainstream educational practice [10].

SRL theory covers a large number of variables and situations which interact in a complex a difficult to control environment. The complexity of the environment and large number of variables, many of which are not directly observable, makes it difficult to conduct studies which provide causal relationships.

SRL research has accumulated a formidable body of literature over the last 30 years of active research. As is inevitable with research involving people in complex domains, much of the research seems to conflict, and it is difficult to draw systematic causal conclusions.

I propose using the new representations and methods of graphical causal models to improve both the ability to draw conclusions from complex observational studies, and to clearly and formally make sense of the existing results in the literature.

1.1.2 Representations in Education

Educational researchers have a variety of existing representations for theories and relationships. The basic form of the theoretical literature in SRL is narrative, presenting natural language descriptions of theories as well as of the empirical results which support them.

As is typical in the social sciences, SRL research analyzes data using statistical methods and the empirical results supporting the theories are presented in the language and representations of statistics and probability. Observational studies report descriptive statistics about the variables they measure as well as correlations between variables, often accompanied by intuitive attempts to propose causal relationships but warnings about doing so from observational data alone. When such studies do propose causal relationships and models, they often do so via multiple regression analysis, or through structural equation modelling and its specialization, path diagrams.

There is a large divide between the representation of the theoretical papers and the empirical papers in terms of formality and specificity. This makes it difficult to see exactly the relationships between the theories and the empirical data, and evaluate the support. The narrative nature of the reports also increases difficulty in combining results from multiple studies.

Meta-analysis provides one technique for evaluating the results of many studies, but it requires a researcher to parse the narrative results to discover the relationships and seek possibly incomplete or not fully described data in the papers. *Graphical causal models* provide one means of representing a model thoroughly, understandably, and formally, which can be useful in helping researchers come to a common understanding of the domain. *Structural Equation Models* (SEM), and their special cases path diagrams and factor diagrams, are the closest relatives of graphical causal models in the educational literature, and in fact functional graphical causal models can be regarded as an extension of SEM.

Structural Equation Models provide a graphical model of the causal relationships between variables of interest. However, the use of these models has generally been limited to conducting comparisons of a priori models proposed by a researcher, or evaluating the fit of a particular such model to a data set. They ignore the large class of other models which could be considered, many of which are equivalent to the models in question under observational data, thus limiting our ability to believe that the results of path diagrams are in some sense true.

1.1.3 Causal Models

There are several varieties of graphical causal models, which will be discussed in detail in the following chapters. All of these models have in common the ability to represent causal relationships between a set of measured variables. The representations differ in their ability to represent latent variables and cyclic relationships. They also differ in their amenability to algorithms which discover causal relationships from observational data and efficiently perform both causal and probabilistic inference. The different models rest on different assumptions about the structure of the domain being modelled.

By using a combination of techniques from the graphical causal model literature, we are able to discover some causal relationships from observational data, even allowing latent variables, and improve the model using both background knowledge and experimental results. We may also learn parametrisations of the models from data, and conduct causal and probabilistic inference over the models to answer questions of interest.

While graphical causal models can be proposed a priori by researchers as with SEM, the ability they provide to discover an equivalence class of models directly from data provides additional power to the researcher. The equivalence class indicates which potential relationships between variables require background knowledge or experimental results to determine. This knowledge allows a researcher to guide their investigation into productive areas, collecting observational data only where useful and experimental data only where necessary.

Researchers have designed algorithms for performing causal inference over graphical causal models, allowing computational answers to questions about the likelihood of world states given an intervention, and counterfactual questions about alternative states of the world. These algorithms extend the capabilities of both SEM and existing probabilistic representations, and should prove useful in educational research.

The two greatest limitations faced by these methods are the assumptions they require in order to connect statistical relationships with causal relationships, and the amount of data necessary for reliability. Detailed arguments have been made as to the when the assumptions can be expected to hold and these will be discussed in Chapter 3. The data requirements are dealt with at length in Chapter 5

1.2 Contributions

This thesis offers two primary contributions to the literature. The first contribution is a pair of engineered causal models derived from the theoretical and empirical literature of SRL. The second contribution is to demonstrate the ability to discover such casual models of SRL from observational data, and assess to what degree this can be done. To my knowledge, both of these contributions are novel within the educational literature.

The first contribution is two engineered causal models of SRL, one based on the theoretical literature and one based on a composite of empirical results from the literature. These models were created by conducting a review of the SRL literature to identify variables and relationships between those variables that have been established via experiment or observation, as well as the theoretical relationships expected by SRL researchers.

Such models have several uses. Graphical causal models clarify the state of the theory in a clear and formal manner. Due to the narrative form of the literature it may not be clear exactly what claims have been supported and which have not, what different theoretical perspectives predict and how they relate. Represented as graphical causal models, they are clearly and formally defined, including what variables are implicit, explicit, covert, measurement variables, and the relationships between them all.

Modelling the theory this way allows us to attempt to draw conclusions about appropriate interventions to take based on the theory. For example, if having goals (self-set or otherwise) *causes* a student to perform better, we can intervene to set goals for the student. However if having goals is only *correlated* with performance as a result of a common cause like goal orientation, then we should attempt to intervene to cause goal orientation. They also provide the ability to incorporate additional observations and experiments to improve the model over time. If we are monitoring the behaviour of students with a system like a Intelligent Tutoring System (ITS), we can continue to add collected data to our model to improve our confidence in the values.

The second primary contribution of my thesis is to demonstrate to what degree such a model can be discovered from observational data using structure learning algorithms and how much observational data would be required to reach the point of statistical indistinguishability. Two models are statistically indistinguishable when they cannot be differentiated using a given set of assumptions and statistical data. Additionally this indicates which relationships would require additional background information or experimental results to

fully determine the causal relationships.

Knowledge of how much of the causal relationships we can learn from observational data is useful because it is generally easier to collect observational data than experimental data. Hence we would like to learn as much as possible from the observational data. We would also like to know what assertions about causal relationships we can and cannot establish with observational data, and approximately how much data is required because it governs our distribution of effort, and our confidence in our conclusions.

As a minor contribution I also analyze the models to determine the which experimental investigations would be necessary, and how many such investigations are theoretically required. It is my hope that the methods outlined in this thesis will motivate the use of these representations in SRL research and educational research in general.

1.2.1 Methodology

Creating Engineered Models

Two separate methods were used to construct the engineered models. In order to create the theoretical model a literature review was conducted of SRL papers which described the theory and the relationships it predicts, as well as review articles which summarized the body of empirical work in terms of the theories. This resulted in a collection of papers which were then read closely for variables' definitions and any correlational or causal relationships which were proposed between the variables. These relationships were then composed to form a complete model.

To create the empirical model a correlation matrix was derived from existing meta-analysis in SRL. The correlational relationships suggested by these results was then employed using the FCI algorithm (Algorithm 3.3.2) with background knowledge to create the empirical network.

Causal Structure Learning

In order to demonstrate the ability to discover such causal models from observational/statistical data I took several steps.

The Fast Causal Inference (FCI) algorithm, (described in Algorithm 3.3.2) is one algorithm for discovering causal models from conditional independence relationships, which can be determined from statistical data. When the statistical data is perfectly accurate it

represents the true dependencies and independencies which exist in the population. When supplied with the conditional independence relationships, either directly or through accurate statistical data, the FCI algorithm produces a representation of the equivalence class. Causal models represent the conditional independence relationships between the variables. Using the TETRAD IV software package [52] I employed the FCI algorithm on the conditional independence relationships represented by the engineered theoretical model to discover the model up to the point of statistical indistinguishability. By doing this from the conditional independence relationships represented by the graph we avoid any artifacts due to sampling variation. I then incorporate temporal background information to determine what if any additional links can be established.

The equivalence class of the theoretical model establishes the upper limit of our ability to discover causal relationships from observational data alone, assuming that data is perfect. Unfortunately infinite perfect data is not available. In order to determine the amount of observational data required to accurately discover the equivalence class I next conducted a simulation study repeatedly using the engineered models to generate statistical data samples and then applying the causal discovery algorithms to that data, increasing the amount of data generated and evaluating the match of the discovered model against the model discovered directly from the conditional independence relationships, thus establishing the viability and limitations of discovering causal relationships from observational studies.

Causal Analysis

Finally I use theoretical results about the number of experiments necessary to evaluate which and how many experiments would be required to fully orient the causal network. I then compare this to the number of experiments required if such methods are not possible.

1.3 Summary

Causality is an essential concept to people in our everyday lives and to scientists and philosophers attempting to understand the world. The formalization of causal representation and reasoning over the last two decades offers an opportunity to bring more powerful analysis of cause and effect in the statistical sciences, and educational research provides an ideally challenging domain. I offer a demonstration of the ability of these formal models of causality to represent SRL theory clearly and concisely, and the new possibilities they offer for

educational research.

1.3.1 Overview

I present a brief review of self-regulated learning in Chapter 2. In Chapter 3 I provide a detailed account of the casual modelling and structure learning methods used in the thesis and their foundations. I detail the methods used in model creation and analysis in Chapter 4. Chapter 5 covers the results of the model engineering and analysis and their meaning, and in Chapter 6 I conclude and present pointers to further work which could be done in this area. In the appendix I cover background material and definitions in graph theory, probability, and statistics.

Chapter 2

Self Regulated Learning

2.1 Overview

Self-regulation of behaviour is a general concept in psychology and is studied in many different domains. SRL research focuses on self-regulation of learning, primarily in academic or classroom environments. SRL theory places the learner at the centre of the learning process as an engaged, proactive agent [76, 10, 42]. Students are considered self-regulated if they actively manage their own learning behaviour, setting and evaluating goals, monitoring the progress, using strategies and tactics for learning and accomplishing their goals, and conducting self-evaluation in order to improve their goals and strategies. This is in contrast to theories which have influenced American educational reform over the past 50 years, which regard the learner as reactive instead of proactive, as a recipient of taught information [76].

SRL is a complex concept with multiple theories which differ in inspiration and emphasis on different learning elements. Boekaerts and Corno [10] suggest several assumptions common to most if not all SRL theoreticians:

> "...students who self-regulate their learning are engaged actively and constructively in a process of meaning generation and that they adapt their thoughts, feelings, and actions as needed to affect their learning and motivation."
>
> "...biological, developmental, contextual, and individual difference constraints may all interfere with or support efforts at regulation."
>
> "...students have the capability to make use of standards to direct their learning, to set their own goals and sub-goals."

> "...there are no direct linkages between achievement and personal or contextual characteristics; achievement effects are mediated by the self-regulatory activities that students engage to reach learning and performance goals."

In essence the researchers believe that actively engaged students take a proactive role in their learning employing learnable skills, strategies, and tactics in all their different domains of action, and that while their personal or environmental characteristics may affect their performance, those effects take place in the context of a students SRL skills.

In particular SRL researchers consider self-awareness of cognitive control strategies and learning strategies. Metacognition, the process of thinking about thinking, is a key aspect in the ability of learners to regulate their own behaviour. The process of learning in a self-regulated fashion is conceived of as a cycle. The student repeatedly passes through a series of phases as they learn, taking relevant actions and using appropriate cognitive and meta-cognitive strategies and tactics as they progress through the cycles.

SRL researchers have proposed, used, and evaluated multiple models of the SRL. The differing models emerge from different theoretical orientations, differentiate and organize the phases differently, and focus on different actions and behaviours within each [46]. Some focus more on overtly visible behaviour, others on more covert behaviour. Zimmerman identifies seven major theoretical traditions in SRL: operant, phenomenological, social cognitive, information processing, volitional, Vygotskian, and constructivist [81]. While the major traditions generally agree on the few assumptions noted above, their stance on what is significant and included in SRL research varies with their source of inspiration. Two major models of SRL are the social cognitive model of Zimmerman [82]. I discuss these two models briefly below, and the model of Winne et al. based on information processing [66].

2.1.1 The Zimmerman Three Phase Model

Zimmerman describes three phases in SRL. The *forethought phase*, which involves goal setting and strategic planning, the *performance phase*, which occurs when the learning behaviour is taking place, and the *self-reflection phase* which covers processes which occur after the learning effort. Within each phase there are multiple self-regulatory behaviours which can take place. Zimmerman describes many of those which have already been investigated in [82].

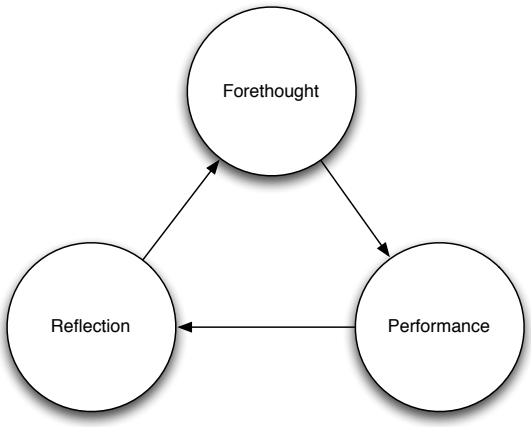


Figure 2.1: The Zimmerman Three Phase Model of SRL

Social Cognitive

Social cognitive theory considers learning to be situated in and affected by physical and social environmental factors as well as by internal cognitive factors. The social cognitive perspective considers SRL skills to be context dependent. Some SRL skills are expected to generalize, such as goal behaviours, but students are not expected to be equally self-regulated in all situations and at all times. Self-regulation can exist to the extent that the learner has some control over the factors of learning, including social, physical, and cognitive environments; if all of the factors are specified externally then a learner cannot self-regulate.

A main cognitive component of SRL for social cognitive theory is self-efficacy beliefs. Self-efficacy beliefs, the belief of learners that they can successfully execute a strategy and accomplish their goal, are central to the social cognitive perspective. Self-efficacy beliefs have been found to influence performance and coping even when controlling for multiple additional factors [7].

2.1.2 Winne and Hadwin - Four Phase Model

Winne and Hadwin proposed a four phase model (see Figure 2.2) with a consistent within phase structure of conditions, operations, product, evaluations, and standards (COPES) [66, 68]. COPES distinguishes this model from others by providing a more detailed view

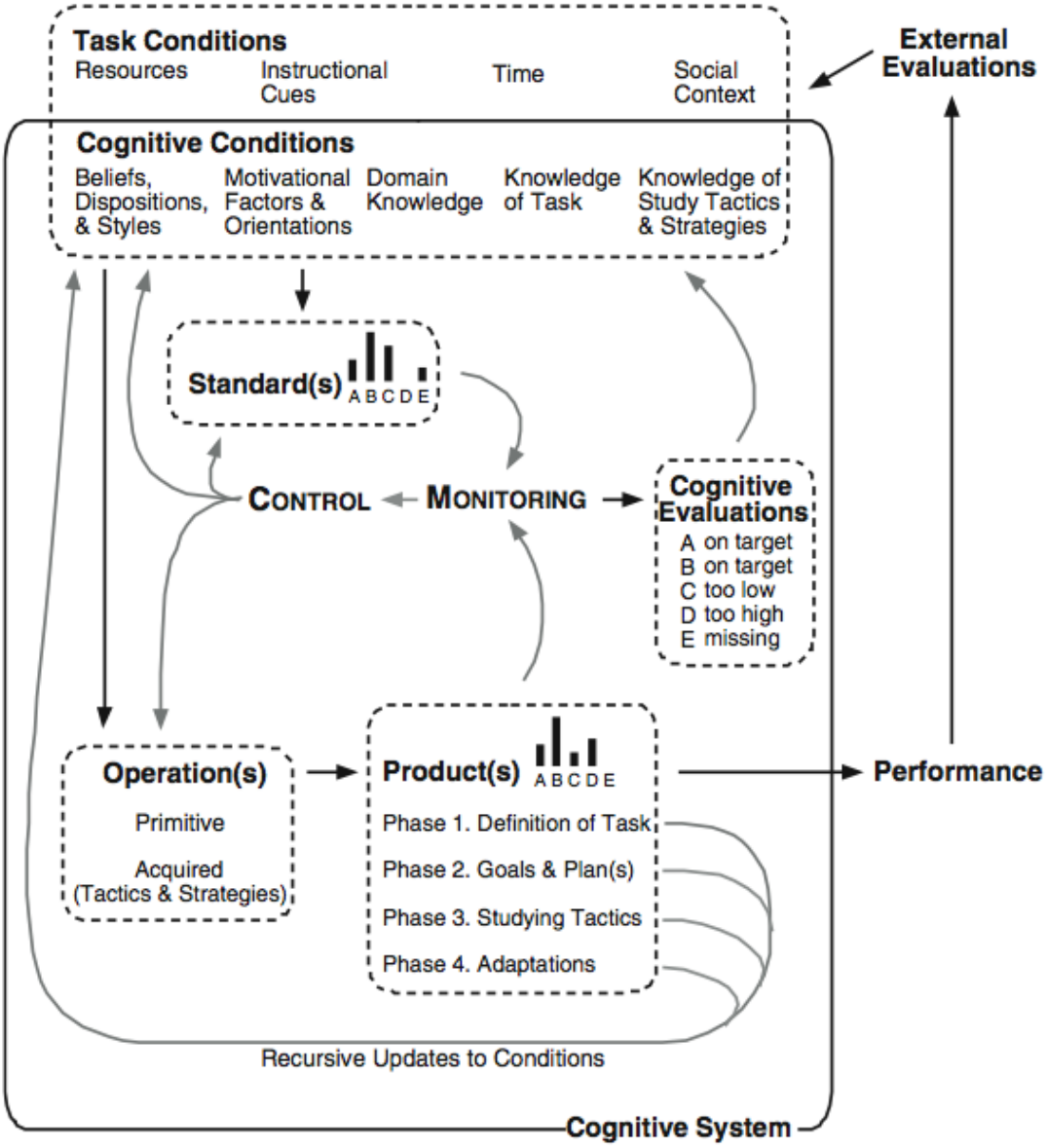


Figure 2.2: Winne and Hadwin's - Four Phase Model of SRL

of what is happening at each stage. The operations in COPES are cognitive operations, strategies, and tactics of the individual; the other components are all the various information inputs used by the operations or, in the case of products, created by operations.

The model also embeds meta-cognitive monitoring as a key part of the operations at all points, evaluating the differences between products and standards to note discrepancies and feed that information back, allowing it to be used again. This produces a feedback loop within and throughout the stages which seeks to better match products to standards or to change standards. Additionally the stages are not seen as strictly ordered, and the student may move around between stages, with information flowing between them in other than the canonical ordering.

Information Processing

Information processing theory grew out of early results in coding theory by Shannon and others who defined the information content of a communication. These ideas were loosely applied by researchers seeking to establish connections between mind and brain accounts of cognition. The human mind was described in terms of two modules, memory and information processing, which operated on symbols. Self-regulation within this environment was based on a negative recursive feedback loop attempting to reduce discrepancies between information on conditions and standards by which they are evaluated [76].

2.2 Methods, Models, and Interventions

2.2.1 Measurement Methods

Boekaerts and Corno describe eight major categories of measurement methods used in SRL research [10]; I reproduce her categories below with brief descriptions of the approaches.

Self-Report Questionnaires attempt to measure learners' SRL behaviour via a series of questions which ask the learner to describe self-regulatory responses to various learning situations. *Observations of Overt Behaviour* record ongoing behaviour and 'score' it according to predefined coding system which determines what variables will be included for consideration. Counts of scores can be subjected to statistical analysis. Qualitative data (e.g. recordings of actions) can also be interpreted, but not statistically. *Interviews* generally seek to obtain qualitative information about experiences during SRL. Interviews take

several common forms, including unstructured interviews where learners tell stories of their behaviour, structured interviews where the interviewer asks questions which build upon each other, guiding the respondent, semi-structured interviews which allows researchers to adaptively select from a pre-defined list of questions as the interview progresses, and stimulated recall, in which students comment while watching a recording of themselves working. In *Think Aloud Protocols* students verbally report their thoughts/strategies etc while they work. This process is limited by the need to train learners beforehand, and it may impact the tasks due to the increased cognitive load experienced. It may also force the students to be more cognitively aware than they might otherwise be. With *Traces of Mental Events and Processes* the research attempts to identify observable traces (evidence) of student learning processes and code them for what each trace indicates. *Situational Manipulations* are experimental studies where students' actions in the learning phase are connected with their performance. In *Recording Student Strategies During Work* students report their mental state with regard to particular variables at regular intervals. Finally with *Keeping Diaries* students keep a diary where they report their SRL behaviour, knowledge, and skills. Some students are more capable writers, which can affect the data. Like interviews, this method produces qualitative data.

2.2.2 Experimental Methods

While difficult to conduct, experimental investigations in SRL research are certainly possible. Bandura describes methods of experimental control and manipulation in investigations of self-efficacy beliefs [7]. Experimental investigations which attempt to investigate impacts while in an authentic educational environment are made more difficult by the complexity of the environment and the possibility for interactions between the groups.

2.2.3 Statistical Methods and Models

Many of the measurement methods described above provide qualitative data about phenomena. Qualitative research can provide researchers with insight into which variables are important, which variables seem to have relationships, and give theoreticians an idea of the structure of the domain. However, most research in psychology, including educational psychology and SRL research is dependent on assessing quantitative data using statistical methods to draw conclusions about the relationships between its variables.

The power of these methods to draw conclusions accurately, and what conclusions can be represented with the methods are then essential to the ability to advance knowledge in SRL.

Existing statistical techniques used in education research include

- Descriptive statistics
- Correlation
- Regression
  - Linear and nonlinear
  - Univariate and multivariate
  - Multi-level hierarchical
- Structural Equation Models
  - Path diagrams
  - Factor Analysis

Descriptive statistics and correlations make no claims about causation, only statistical conclusions, i.e. correlations and conditional probabilities, can be drawn from descriptive statistics and correlation without additional assumptions or experimental results. *Meta-Analysis* is a popular technique for combining data and statistics, including descriptive statistics and correlations, recorded in multiple studies in the literature. Meta-analysis has the benefit of increasing statistical power of tests by increasing the effective sample size, and of consolidating information from many sources. It considers effect sizes, and uses simple techniques to combine them. However it does not directly allow one to draw causal conclusions unless experimental studies are being meta-analyzed.

Regression testing is sometimes used to attempt to identify causal relationships by analyzing how much variance variables 'explain' about each other. However, in general, regression analysis per se makes no supportable claims about causation. Spirtes et al. [60] argue strongly that regression, as used, is in fact poorly suited to causal analysis in most cases.

Structural Equation Models (SEM) are graphical representations of relationships between variables which have been in use in the social science for nearly 90 years. Linear SEM represents linear relationships between variables generally assuming a normal (Gaussian)

distribution. The relationship between variables is represented as a linear equation of the form $y = \beta x + u$ where y is the effect, x the cause, u the error term represents the effect of all other variables, and β is the path coefficient which quantifies the strength of the relationship between the variables [41]. They can be used to represent both measured variables and latent variables. The measured variables are often the results of surveys or other instruments where multiple measures are intended to estimate an unmeasured (latent) variable. In this case the measured variables are called the measurement model and the latent variables are called the structural model. In SEM an undirected edge represents a correlation relationship, and a directed edge represents a 'directed relationship' [25]. Structural Equation Models original interpretation did include causation, however that interpretation has fallen out of favour over time [39].

Path diagrams, also called path analysis, are a special case of SEM which excludes latent variables. In the terminology of Spirtes et al. [60] this is known as a pure measurement model. Pearl notes [41] that the causal assumptions in path diagrams are represented by the absence of links, which represents a definite absence of a relationship, whereas the presence of a link only represents the possibility of a cause.

In psychology, SEM are often employed to analyze the results of observational studies, though they can represent experimental studies as well [25]. Social scientists generally use SEM in a confirmatory approach. The researcher proposes a model based on a theory or other considerations and then either tests the fit of the model to the data, or compares the fit of the proposed model against another baseline model. The a priori proposal of models presents a serious difficulty as model fit to data does not imply correctness of the model, and large sets of models which are statistically equivalent may exist. It appears that researchers are often unaware of or discount the existence of equivalent models, and are likely to overestimate the likelihood of the proposed model being correct [25]. The techniques I describe in the following chapters address this issue.

2.2.4 Results of Educational Interventions

There have been many different educational interventions made on the basis of different models of SRL. These vary from individual interventions to help with remediation of struggling students via strategy instruction to attempts to restructure classroom environments, to full school interventions, and the creation of computer mediated learning environments. Methods of academic strategy instruction have successfully been demonstrated to increase

achievement in multiple domains using path analysis methods, and have been incorporated into mainstream education [10].

2.3 Summary

SRL is a successful educational theory with a large literature that has contributed to a theoretical understanding of academic learning and to practical improvements in education. Researchers have attempted to derive causal relationships about SRL using the best statistical tools available to them combined with experimental investigations whenever possible. Unfortunately experimentation is very often not possible, and the statistical techniques for drawing causal conclusions otherwise are limited, prone to error, and require a great deal of subjective treatment leading to possible bias. Finally the scale of the literature itself and its narrative presentation presents a barrier to integrating, understanding, and communicating results clearly and accurately. The methods I describe in the following chapters can help address these issues by improving the formal statistical basis of causal inferences in SRL, offering a means to integrate results from multiple studies into causal conclusions, and providing a clarifying representation for theoretical and empirical causal claims.

Chapter 3

Causality and Causal Models

In this chapter I review causal models in some detail and motivate the use of causal modelling in SRL research.

3.1 Causality

Causality has been a troubling concept for philosophers and scientists for generations. An intuitively obvious concept, it has been remarkably difficult to define clearly and formally. In fact some statisticians, starting with Pearson, completely reject the concept as artificial, ill defined, and unnecessary. These statisticians define correlation as the fundamental relationship between entities [39], and causation to be nothing more than convenient shorthand. Those who do not reject causation outright advance several potential definitions; I will focus on two: counterfactual causation and interventional causation.

Causation has been defined in terms of counterfactual relationships. A *counterfactual* is a statement that is literally counter to fact, that is, something which did not happen. Counterfactual questions are about what would have happened had some factor been different from what in fact occurred. From the counterfactual perspective a causal relationship exists between two variables if the 'cause' occurred differently would have resulted in the 'effect' occurring differently. Examples of counterfactual questions include: asking whether a poorly achieving student would be a good student if he had been in a smaller class, asking if the Iraq war would have happened if the September 11th attacks had been averted, and asking if Vancouver would be less livable if a third crossing of the Burrard inlet had been constructed. In causal language those questions are: is class size a cause of performance

gains in students, were the September 11th attacks a cause of the Iraq war, and is bridge and freeway construction a cause of decreased livability. Counterfactual questions are common in our everyday lives, in political discourse, in the legal system where they are used to assign blame, and in making policy decisions. I address counterfactuals in more depth in the review however, my primary focus will be the interventional account of causation.

Causality can also be defined in terms of intervention, also called manipulation. Essentially, a variable A is considered a cause of variable B if changing (manipulating) A results in a change in B. For example, we say that impending rain, and the accompanying increase in air pressure causes a barometer to rise, but the rising barometer does not cause the rain, or the air pressure. From the manipulation perspective this means that manipulating air pressure changes the barometer level but manipulating the barometer level does not change the air pressure or the rain. This practical definition is essentially what is used in randomized controlled experiments throughout the sciences to detect causation. An experiment attempts to manipulate one or more variables while ensuring through randomization that the other variables have the same characteristics they would normally have. If an unmanipulated variable changes with our manipulation we ascribe a causal relationship between the manipulated and unmanipulated variable.

Spirtes et al. define causation thus:

> "We understand causation to be a relation between particular events: something happens and causes something else to happen. Each cause is a particular event and each effect is a particular event. An event A can have more than one cause, none of which alone suffice to produce A. An event A can also be overdetermined: it can have more than one set of causes that suffice for A to occur. We assume that causation is (usually) transitive, irreflexive, and antisymmetric. That is, i) if A is a cause of B and B is a cause of C, then A is also a cause of C, ii) an event A cannot cause itself, and iii) if A is a cause of B then B is not a cause of A."[60]

Maes et al. refer to Neapolitan defining causality thus:

> "Our operational definition of causality is as follows: a relation from variable C to variable E is causal in a certain context, when a manipulation in the form of a randomised controlled experiment on variable C, induces a change in the probability distribution of variable E, in that specific context." [26, 35]

These definitions are more specific than the examples given previously, but greater precision is required. The formalization of such definitions of causality has allowed causality to be analyzed mathematically. I present formal definitions in Section 3.2.2.

3.1.1 Importance of Causal Relationships

The main argument I make for the importance of causal relationships and causal modelling is their usefulness. A model of the world is useful if it helps us understand and answer questions about the world. Causal representations allow us to answer a broader set of questions than stochastic representations. In a purely probabilistic model, represented by the joint probability distribution, we can answer questions about the correlation between any sets of variables. Probabilistic representations such as Bayesian networks can answer questions about the likelihood of one set of variables taking particular values, given that we have observed another set, and about which set of variables we should observe to obtain the most information about the likelihood of another set of variables taking particular values.

Causal models add the ability to answer questions about the likelihood of a set of variables taking particular values if we *intervene* to change the value of another set of variables. The distinction may not seem large, but it has major implications. This type of question is asked often, for example, in trying to improve education, we are interested in how various factors relate to learning achievement because we want to take some action to improve achievement. Knowing the correlation of achievement with high self-efficacy beliefs is interesting, but what we want to know is if intervening to increase self-efficacy beliefs will *cause* an increase in achievement. These questions cannot be answered by purely stochastic models.

The discovery of causal relationships is fundamental to science and to explaining how the world works in terms of the underlying mechanisms. In the social sciences such as economics or education, scientists try to determine which variables cause other variables and the strength of those relationships, such as the effects of taxation policies on the economy, or of negative feedback on academic performance. In medicine, we seek to understand whether a drug causes the remission of a disease, or whether a toxin causes an ailment. Historians also looks for causes, asking counterfactual questions about what historical events caused other events of interest, and how things might have turned out differently.

One additional argument that has been made is that causal relationships are a natural knowledge organization for human beings. Pearl argues that the causal organization

of knowledge is the most natural and fundamental [39], that it is how people organize our knowledge, and that people tend to ignore probabilistic information once they have learned the underlying causal relationships. This is in contrast to the traditional empiricist view of statistics, which defines correlation as the most fundamental relationship. Whichever relationship is fundamental it is clear that we need techniques for answering both correlational and causal questions.

3.2 Probabilistic Models vs Causal Models

The most important differences between probabilistic models and causal models are the type of information they can represent and hence the type of questions they can answer. Probabilistic models can be used to perform probabilistic inference, answering questions such as "What is the likelihood of X if I observe Y?", "What is the prior probability of X", and "Which variable should I observe to gain the most information about X (other than X)?" These questions can either be answered directly from a joint probability distribution, or by using a probabilistic model which represents the joint probability distribution more efficiently, such as Bayesian Networks.

Perhaps less obvious is what questions *cannot* be answered by stochastic models, but are answerable by causal models. Causal models allow for causal inference, answering questions such as "What is the likelihood of X if I *do* Y?", equivalently "What is the likelihood of X if I force Y to take on a particular value?" The difference between doing and seeing, elucidated by Pearl [39], is essential to the manipulative account of causation. Causal models allow us to answer questions about what will happen if we take action to change the world, instead of what will happen if we simply observe it. If we know that drinking and driving causes increased likelihood of car collisions, and car collisions cause increased likelihood of death, and we observe two cars get in a head on collision we have reason to increase our belief that the driver is both drunk and dead, but if we take control of the car remotely and cause the head on collision we still expect the unsuspecting driver to die, but we would have no reason to suppose him to be drunk. Our, rather immoral, intervention has broken the regular causal connection between drunk driving and traffic collisions, and trying to draw a conclusion solely from the correlation and drunk driving is erroneous. The change in circumstances invalidates our existing correlation and we have no way of recording the difference using probability alone. There is no way to write 'The barometer does not cause

the rain.' using the language of probability. We can only denote the dependency and only under static conditions. Causal models can also answer questions about counterfactuals such as "What is the likelihood that X would have been x if Y was y given that X was not x and Y was not y?" or more clearly "What would have happened if X had been different?"

A wide variety of policy issues require answering causal questions. To take an example in education, a probabilistic question might be "If we observe that a student has high self-efficacy beliefs, what is the likelihood he will be a high-achiever?" a causal question "If we intervene to increase a student's self-efficacy beliefs, what is the likelihood he will be a high-achiever?" or "Will intervening to increase a student's self-efficacy beliefs cause an increase in his achievement?" or the counterfactual "Would a student's achievement have been greater if his self-efficacy beliefs had been higher?" The causal questions are clearly what we want an answer to, and those answers are not available from purely probabilistic models. Knowing that self-efficacy beliefs are correlated with performance does not tell us whether acting to increase an individuals self-efficacy beliefs would have an effect on their performance, and this is precisely what we want to know.

Work in probabilistic models over the last twenty years, in particular the development Bayesian Networks underpins much of the subsequent work in causal models.

3.2.1 Bayesian Networks

According to Pearl the role of graphs in probabilistic and statistical modelling is threefold:

> "1. to provide convenient means of expressing substantive assumptions;
> 2. to facilitate economical representation of joint probability functions; and
> 3. to facilitate efficient inferences from observations." [39]

Bayesian Networks (BN), also called Bayesian belief networks or just belief networks, are a graph based representation of a joint probability distribution over a set of variables which accomplishes each of the above points. The BN consists of nodes and links between the nodes. The nodes represent variables, either discrete or continuous, and have attached conditional probability distributions. The links represent probabilistic dependencies between the nodes.

The conditional probability tables attached to each node provide the probabilities of the node taking its different values given the different possible values of its parent nodes. BN

are a parsimonious representation of the joint probability distribution because they require less specification of probability values for basic events, but can be used to compute the full joint probability distribution. The complete joint probability distribution over n binary variables requires the specification of 2^n probabilities. If the network is sparse, that is, each variable depends directly on a small subset of the complete set of variables, then the BN is much more efficient. For example if each variable depends only on 3 other variables, then each CPT requires only 2^3 probabilities.

The ability to efficiently represent the joint in this way is given by the Markovian Parents definition, due to Pearl [39], which states essentially that we may compute the conditional probability of a variable based only on the values of variables which it is dependent on conditional on any subset of the variables. A probability distribution is then related to the Bayesian network by either the Parental Markov Condition, which requires that every variable be independent of its non-descendants (excluding itself) in the graph, conditional on its parents.

d-separation *d-separation* (for directional separation) is a graphical criterion which allows us to read off the conditional independencies from a BN. If a probability distribution P and a graph G are compatible then d-separation implies conditional independence. d-separation is important because it gives us a purely graphical method of determining conditional independence. We represent d-separation as $(X \coprod Y|Z)_G$ and conditional independence as $(X \coprod Y|Z)_P$

A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if 1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or 2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middles node m is not in Z and such that no descendant of m is in Z. A set Z is said to d-separate X form Y if and only if Z blocks every path from a node in X to a node in Y.

d-separation makes sense if we consider the links between the variables to be causal. In the case of a causal chain $i \to m \to j$ the i causes m which causes j. Initially i and j are dependent, because learning about either of them changes the probability of the other, but once we learn the value of m it 'screens off' the i from j, because the only way i influences j is through m, of which we already know the value. In the case of the the fork $i \leftarrow m \to j$ i and j are initially dependent because of their common cause m. If we learn about either i or j it affects the probability of m and hence the other child. However, if we learn about m, we

can gain no further information about i from j or vice versa, the dependence is 'blocked' or 'screened off'. Finally with the inverted fork $i \rightarrow m \leftarrow j$ we have a common effect. In this case i and j are independent as long as we know nothing about m or its children. However, if we learn the value of m or any of its children we 'open' a path between the common causes, because now learning about one of them would affect our belief in the other.

Observational Equivalence Observational equivalence occurs when two networks cannot be distinguished based on probabilities alone. Just as one graph can be compatible with multiple probability distributions, one probability distribution can be compatible with multiple graphs. If two graphs are compatible with the same probability distribution, then we cannot distinguish them based on the probability distribution. In order to distinguish the graphs we must include further background information or assumptions about the relationship of the probability distribution to the graph. The two primary sources of background information are temporal information, about the ordering of the variables, and experimental results.

Inference

Inference is the process of computing results based on the model. From the complete joint distribution it is trivial to answer questions about the probability that a variable will take a particular value whether or not values are specified for any subset of the other variables. The joint distribution acts as a lookup table, we simply find the appropriate location in the table and read off the result. Unfortunately the joint distribution grows so quickly in the number of variables that it is completely impractical. Bayesian networks allow us to represent the joint efficiently, but at the cost of having to compute the results to particular queries when we want to know their values.

Algorithm exist for BNs to compute the likelihood of variables taking on a particular value given the values of other variables, and which variable we should observe to gain the most information about the value of a particular variable. The algorithms are dramatically more efficient than computing the joint distribution directly, however both the exact inference problem and the approximate inference problem are NP. It is possible to determine ahead of time the time and space requirements of the exact algorithms, which allows us to use the more efficient approximate algorithms.

3.2.2 Causal Models

Graphical causal models represent the causal relationships between variables graphically in a manner directly analogous to Bayesian networks representing probabilistic relationships. Nodes represent variables, but now edges represent a causal relationship of some kind. In fact Bayesian networks can be given a causal interpretation, though it is possible to have a BN which isn't causal. There are many different types of graphical causal models, which correspond to different sets of assumptions about the underlying distributions and different axioms for associating causal and statistical relationships with graph elements. In this chapter I cover models which represent causal relationships with discrete or continuous variables, and linear relationships between the variables in the continuous case. The graphs presented can also represent the existence of confounding variables.

Probability has long been subject to formal mathematical treatment, with axioms and a symbolic language. Causality has not had this benefit and has been forced to make do with vague natural language descriptions and varying definitions. The work of Pearl, Verma, Spirtes, Glymour, Richardson, and others over the twenty years has provided a formal representation and semantics for causal relationships and their connections to graphs. This account of causation, based on manipulation, has begun to gain recognition and use in philosophy of science, econometrics, psychology, and artificial intelligence. Research into efficient algorithms, formal correctness, necessary assumptions, and appropriate axioms is ongoing actively.

Definitions

At the opening of this chapter I provided several approximate definitions of causation. These approximate definitions capture the intuition behind the manipulative account of causality, but we require a formal mathematical treatment if we are to calculate the results of causal relationships.

Potential Causal Influence A variable X has a potential causal influence on another variable Y (that is inferable from P') if the following conditions hold.

1. X and Y are dependent in every context.

2. There exists a variable Z and a context S such that

(i) X and Z are independent given S and

(ii) Z and Y are dependent given S.

Genuine Causal Influence A variable X has a genuine causal influence on another variable Y if there exists a variable Z such that either:

1. X and Y are dependent in any context and there exists a context S satisfying

   (i) Z is a potential cause of X

   (ii) Z and Y are dependent given S

   (iii) Z and Y are independent given the union of S and X

2. X and Y are the transitive closure of the relation defined in criterion 1

Causal mechanism A causal mechanism is the physical means or mechanism by which one variable influences another. Mechanisms are assumed to be independent, stable, and autonomous. By stable we mean that the mechanism remains regardless of changes to parameter values or context. Autonomous means that the mechanism remains invariant to changes in other mechanisms, changing the environment in which the mechanism does not change the mechanism unless we change it directly, Pearl calls this 'transportability'.

Causal effect A causal effect of variable X on variable Y is the probability distribution of Y given that we have intervened on X. In Pearl's notation this is $P(Y = y|do(x)) = \sum(z)P(z, y|do(x))$ [39, 41]

Direct/Actual Cause A direct cause is one which, given the set of variables under consideration, is never independent of its effect, conditioned on any subset of the variables. Whether a variable is a direct cause of another variable is dependent on the set of variables being considered. We can often introduce a mediating cause by adding another variable which is in some sense at a 'lower level' than the existing variables.

Latent variables and Confounding Latent variables are those which are not directly measured in a study; they are present in most real world problems. Latent variables may be variables we are not interested in and hence have not measured, variables we are not aware of, variables we are not practically able to measure, or variables which are by definition unmeasurable.

Latent variables become a problem for studies attempting to determine causality when they impact two or more of our measured variables; this problem is known as confounding. A latent variable which is a cause of two or more measured (observed) variables can cause the observed variables to be correlated even though the observed variables may not cause each other. It can also effect the degree of correlation between variables which are causally related. Confounding is a fundamental problem in attempting to derive causal relationships in any kind of study, whether experimental or observational.

The following definitions due to Pearl, or Spirtes et al. as noted.

Causal Structure (Pearl) A causal structure of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V, and each link represents direct functional relationship among the corresponding variables.

Causal Model (Pearl) A causal model is a pair $M = \langle D, \Theta_D \rangle$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $P(u_i)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$, independently of all other u.

Latent Structure (Pearl) A latent structure is a pair $L = \langle D, O \rangle$, where D is a causal structure over V where $O \subseteq V$ is a set of observed variables.

Assumptions

There are three main assumptions which relate a causal structure represented by a graph G to a probability distribution. These assumptions, or axioms, are the foundation for representing causality graphically.

Causal Markov Condition (Spirtes et al.) The first assumption is the causal Markov condition. This is the same assumption that allows a Bayesian network to efficiently represent the joint probability distribution, but with a causal interpretation.

"Let G be a causal graph with vertex set $\mathbf{V}$ and P be a probability distribution over the vertices in $\mathbf{V}$ generated by the causal structure represented by G. G and P satisfy the Casual Markov Condition if and only if for every W and $\mathbf{V}$, W is independent of $\mathbf{V} \backslash (\mathbf{Descendents}(W) \cup \mathbf{Parents}(W))$ given $\mathbf{Parents}(W)$." [60]

Minimality (Pearl) The minimality condition moves us beyond the assumptions inherent in Bayesian networks. The minimality condition holds if the removal of any edge from the model will cause the violation of the causal Markov condition. The minimality condition subsumes the Markov condition; a model may be Markov but not minimal, but it cannot be minimal and not Markov. The minimality condition follows the standard scientific principle of Occam's Razor, limiting the inclusion of unnecessary edges in the graph.

Faithfulness and Stability Stability (also called Faithfulness) is the principle that a model is stable only if varying the parameters of the model does not destroy any independencies. The reason is that independencies induced by specific parametrisations are unlikely to be produced by data, and that all real independencies are assumed to be structural.

There are several circumstances in which the faithfulness condition may be violated. In particular it may be violated if two variables with different causal relationships are aggregated, if the population is a mixture of sub-populations with different causal structures, or if there are deterministic relationships between variables. Recent work has been done on establishing weaker versions of the faithfulness condition which are still sufficient for the algorithms to be correct [75].

"Let $I(P)$ denote the set of all conditional independence relationships embodied in P. A causal model $M = \langle D, \Theta_D \rangle$ generates a stable distribution if and only if $P(\langle D, \Theta_D \rangle)$ contains no extraneous independences - that is, if and only if $I(P(\langle D, \Theta_D \rangle)) \subseteq I(P(\langle D, \Theta'_D \rangle))$ for any set of parameters Θ'_D." [39]

"Let G be a causal graph and P a probability distribution generate by G. $\langle G, P \rangle$satisfies the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition applied to G." [60]

Observational Equivalence

Also known as Markov equivalence and statistical indistinguishability, observational equivalence is the point at which no additional information about a graphical causal model can be obtained from statistical data given the assumptions employed. There are different classes of observational equivalence for different assumptions and restrictions on the presence of latent variables.

When unmeasured common causes are not permitted observational equivalence is fairly well understood. Assuming the Markov and Faithfulness conditions, running the SGS or

PC algorithm on a distribution will produce a *pattern* which represents the equivalence class for the distribution.

A *pattern* P is a partially directed graph which represents a class of DAGs. A DAG G is in the class represented by a pattern iff

1. G and P have the same skeleton
2. If an edge is directed in P it is also directed in G and
3. if a unshielded collider exists in G it also exists in P.

When unmeasured common causes are permitted the issue grows more complex. The FCI algorithm (Algorithm 3.3.2) produces a representation of the equivalence class in the form of a *Partial Ancestral Graph* (PAG). PAGs are describe below.

Graphical causal models may be broadly arranged into two types: stochastic, and functional. Stochastic causal models represent relationships between nodes as indeterministic. The relationships between the nodes are defined by the probability tables and the sets of parents. Functional causal models posit a deterministic functional relationship between each variable and its parents, and include a randomly distributed error term as input to each function. The causal interpretation of BNs is a stochastic causal model. Structural Equation Models, described in the previous chapter, are linear functional models.

Stochastic Causal Models

Causal Bayesian Networks *Causal Bayesian Networks* are Bayesian Networks in which the links are given a causal interpretation instead of just a correlational interpretation. In a causal bayesian network a directed link between two variables represents a direct causal relationship. The Causal Markov Condition and Causal Faithfulness Condition are assumed.

Functional Causal Models

Semi-Markovian Causal Models Semi-Markovian Causal Models (SMCM) are graphical causal models in which a directed link between two variables represents a direct functional causal relationship between the variables. All edges are directed, and there are no directed cycles permitted. Each variable is associated with a single exogenous error variable, and the error variables may be correlated. The presence of correlated error variables represents

confounding by latent common causes. SMCM assume that the Causal Markov Condition and the Causal Faithfulness Condition hold.

Markovian Causal Models Markovian Causal Models are equivalent to Semi-Markovian Causal Models except that they require that the error terms are independently distributed, thus assuming the absence of confounding.

Partial Ancestral Graphs Partial Ancestral Graphs are a graphical representation in which a directed edge between two variables represents an *ancestral* relationship instead of a direct causal relationship. That is, in the underlying causal graph G which the PAG represents, there is a directed path between variables A and B if there is a direct link between variables A and B in the PAG. Bi-directed edges are permitted and indicate the there is a latent common cause for the two connected variables. Partial Ancestral Graphs assume both the Causal Markov Condition and the Causal Faithfulness Condition.

3.3 Structure Discovery Algorithms

Structure discovery algorithms, also called causal learning, structure learning, or causal search algorithms, attempt to discover the causal structure which generates a set of data from the data itself. There are two main approaches to causal discovery. The first approach is constraint based algorithms. Constraint based algorithms attempt to combine a set of assumptions about the relationship between causal structure and conditional independence relationships present in the data to constrain the set of allowable structures that could be correct given the data. The second approach is Bayesian structure learning. Bayesian structure learning assigns subjective prior probabilities to complete structures and then updates the likelihoods of the models to select the model with the maximal likelihood. Both methods have had success, and the same assumptions about causal structure relating to statistical data are used by both. In this thesis I shall only consider the constraint based approach.

In the remainder of this section I describe several constraint based algorithms for discovering causal structure and the assumptions they rely on.

A) Form the complete undirected graph H on the vertex set V.

B) For each pair of vertices A and B, if there exists a subset S of V $\{A,B\}$ such that A and B are d-separated given S, remove the edge between A and B from H.

C) Let K be the undirected graph resulting from step B). For each triple of vertices A, B, and C such that the pair A and B and the pair B and C are each adjacent in K (written as $A - B - C$) but the pair A and C are not adjacent in K, orient $A - B - C$ as $A \rightarrow B \leftarrow C$ if and only if there is no subset S of $\{B\} \cup V \backslash \{A, C\}$ that d-separates A and C.

D) repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient $B - C$ as $B \rightarrow C$.

If there is a directed path from A to B, and an edge between A and B, then orient $A - B$ as $A \rightarrow B$.

until no more edges can be oriented

Algorithm 3.1: SGS Algorithm - [60]

3.3.1 Algorithms

Algorithms Without Latent Variables

There are several constraint based algorithms for discovering causal structures given causal sufficiency. The algorithms of Verma and Pearl are very similar to those of Spirtes, Glymour, and Shienes. The IC algorithm was originally proposed by Pearl and Verma, and the SGS and PC algorithms proposed by Spirtes et al. are more detailed specifications and extensions.

The basic proceedure of the SGS and PC algorithms is similar. Initially the complete undirected graph over the variables is formed. Then each adjacency is tested for d-separation and removed if it is found to be d-separated by any subset of variables not including the endpoints. Next orientation begins based on discovering unsheilded colliders. Finally additional edges are oriented as possible by repeatedly applying additional edge orientation rules.

A.) Form the complete undirected graph C on the vertex set V.

B.) n = 0.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that Adjacencies$(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset S of Adjacencies$(C, X)\backslash\{Y\}$of cardinality n, and if X and Y are d-separated given S delete edge $X - Y$ from C and record S in Sepset(X, Y) and Sepset(Y, X);

until all ordered pairs of adjacent variables X and Y such that Adjacencies$(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets of S of Adjacencies$(C, X)\backslash\{Y\}$ of cardinality n have been tested for d-separation; $n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, Adjacencies$(C, X\backslash\{Y\}$ is of cardinality less than n.

C.) For each triple of vertices X, Y , Z such that the pair X, Y and the pair Y, Z are each adjacent in C put the pair X, Z are not adjacent in C, orient $X - Y - Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in Sepset(X, Y).

D.) repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient $B - C$ as $B \rightarrow C$.

If there is a directed path from A to B, and an edge between A and B, then orient $A - B$ as $A \rightarrow B$.

until no more edges can be oriented.

Algorithm 3.2: PC Algorithm

The output of the PC algorithm is a pattern, which is a partially oriented graph representing the class of observationally equivalent models.

While SGS is not viable for significant models due to complexity constraints, PC makes use of local information in order to be more efficient. PC takes advantage of the fact that with no latent variables, if a pair of variables A,B are d-separated, then they are d-separated given some subset of Parents(A) or Parents(B), so only nodes which are adjacent to A or B need to be considered, and the size of this set decreases as the algorithm runs.

The complexity of PC is still significant, in fact it is exponential in the average in-degree of its nodes.

Let n be the number of vertices and k be the maximal degree of any vertex

Then the algorithm is bounded in the worst case by

$$2(\frac{n}{2})\sum_{i=0}(\frac{n-1}{i})$$

Which is bounded by

$$\frac{(n^2)(n-1)^{(k-1)}}{(k-1)!}$$

This worst case is rare, but a formal expected complexity analysis is not reported. According to Spirtes et al. [60] it is possible to recover graphs of up to 100 variables, given that they are reasonably sparse.

3.3.2 Algorithms With Latent Variables

The IC* algorithm was proposed by Verma and Pearl and the CI and FCI algorithms were created by Spirtes, Richardson, et al. [60]. These algorithms are constraint based and make use of conditional independence relationships plus a set of assumptions to discover causal relationships from data. The CI algorithm is not suitable for use on large data sets due to its computational complexity. The FCI algorithm produces a PAG representing the equivalence class.

Background Knowledge

Background knowledge can be incorporated into the algorithms as additional constraints on the relationships between variables. Temporal information about the time ordering of variables is the most common form of background information and has historically been

A.) Form the complete undirected graph Q on the vertex set $\mathbf{V}$.

B.) If A and B are d-separated given any subset $\mathbf{S}$ of $\mathbf{V}$, remove the edge between A and B, and record $\mathbf{S}$ in $\mathbf{Sepset}(B, A)$.

C.) Let F be the graph resulting from step B). Orient each edge $o - o$. For each triple of vertices A, B, C such that the pair A,B and the pair B, C are each adjacent in F but the pair A, C are not adjacent in F, orient $A * - * B * - * C$ as $A* \rightarrow B \leftarrow *C$ if and only if B is not in $\mathbf{Sepset}(A, C)$, and orient $A * - * B * - * C$ as $A * - \underline{*B*} - *$ C if and only if B is in $\mathbf{Sepset}(A, C)$.

D.) repeat

If there is a directed path from A to B, and an edge $A*-*B$, orient $A*-*B$ as $A* \rightarrow B$,

else if B is a collider along $< A, B, C >$ in π, B is adjacent to D, and D is in $\mathbf{Sepset}(A, C)$, then orient $B * - * D$ as $B \leftarrow *D$,

else if U is a definite discriminating path between A and B for M in π, and P and R are adjacent to M on U, and $P - M - R$ is a triangle, then

if M is in $\mathbf{Sepset}(A, B)$ then M is marked as a noncollider on subpath $P * - \underline{*M*} - *R$

else $P * - * M * - * R$ is oriented as $P* \rightarrow M \leftarrow *R$.

else if $P* \rightarrow \underline{M*} - *R$ then orient as $P* \rightarrow M \rightarrow R$.

until no more edges can be oriented.

Algorithm 3.3: Casual Inference (CI) Algorithm

A). Form the complete undirected graph Q on the vertex set $\mathbf{V}$.

B). $n = 0$

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in Q such that **Adjacencies**$(Q, X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**$(Q, X))\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given **S** delete the edge between X and Y from Q, and record **S** in **Sepset**(X, Y) and **Sepset**(Y, X)

until all ordered variable pairs of adjacent variables X and Y such that **Adjacencies**(Q,X))$\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**$(Q, X))\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$

until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(Q, X))\backslash\{Y\}$ is of cardinality less than n.

C). Let F' be the undirected graph resulting from step B). Orient each edge as $o-o$. For each triple of vertices A, B, C such that the pair A, B and the pair B, C are each adjacent in F' but the pair A, C are not adjacent in F', orient $A * - * B * - * C$ as $A* \rightarrow B \leftarrow *C$ if and only if B is not in **Sepset**(A, C).

D). For each pair of variables A and B adjacent in F', if A and B are d-separated given any subset **S** of **Possible-D-SEP**$(A, B)\backslash\{A, B\}$ or any subset **S** of **Possible-D-SEP**$(B, A)\backslash\{A, B\}$ in F remove the edge between A and B, and record **S** in **Sepset**(A, B) and **Sepset**(B, A).

The algorithm then reorients an edge between any pair of variables X and Y as $Xo-oY$, and proceeds to reorient the edges in the same way as steps C) and D) of the Causal Inference Algorithm

Algorithm 3.4: Fast Causal Inference (FCI) Algorithm

assumed to be necessary, though not sufficient, to deduce causal relationships. Incorporating temporal information can restrict the possibility of some orientations of adjacencies, thereby allowing the algorithms to be sure about more edges. Additional information about specifically required or disallowed links can also be incorporated in the same fashion.

3.4 Applicability to SRL

As Green and Azevedo put it "Self-regulated learning (SRL) theories attempt to model *how* each of these cognitive, motivational, and contextual factors influences the learning process" [21] (emphasis mine). The question is not *if* various factors influence learning, it is *how* they influence learning, and that is a causal question. Given that SRL attempts to answer causal questions, the remainder of this thesis constitutes an attempt to demonstrate the applicability of causal modelling and causal structure discovery from observational data to the domain of SRL.

3.4.1 Discovering Causal Structures in Education

Causal discovery algorithms can be used to identify causal relationships between variables and to find equivalence classes. Causal discovery algorithms such as FCI provide a formal means of identifying causal relationships between variables of interest from observational data, possibly combined with experimental results.

Compared to existing methods used in the social sciences this methodology is most closely related to an exploratory use of structural equation modelling. The very important difference is that as conventionally used SEM provides only a single model, which in some sense 'fits' the collected data. The FCI algorithm provides a representation of the equivalence class of causal models which can produce the given statistical data.

An exploratory approach with SEM is difficult to justify for two reasons. Firstly, only a single model out of the many models which are indistinguishable given the evidence is considered. Assuming the justification for the model is observational data or statistics, there is no reason to prefer one model from the equivalence class to another. Thus the 'fit' of the model to the data does not provide a strong reason to believe that the relationships represented are true, that is, that they exist in the world.

The class of indistinguishable models provided by the FCI algorithm indicates which

relationships are common in all models which can produce the data under the given assumptions. Given the accuracy of the data, and the validity of the assumptions, this does give us a strong reason to believe that those relationships are true, and hold in the world.

The second reason the existing exploratory approach is difficult to justify is the possibility of over fitting, as it is called in machine learning. A model is over fitted if it assumes that either random variance in the sample data or peculiar characteristics of the sample apply to the wider population where they do not. In the exploratory approach we are attempting to generalize from sample data to population characteristics and risk assuming that relationships which appear to exist in the sample do no exist in the population.

Attempts to specify a completely oriented causal model from statistical data when there are many models which can produce the same statistics is one form of over-fitting which results from the exploratory approach where a single model is proposed to fit the data. The FCI algorithm avoids this problem by only specifying the equivalence class.

Several approaches are used in scientific work to avoid or overcome over-fitting. The first is the use of Occam's Razor which, to paraphrase, states that all things being equal we should prefer explanations which make less assumptions. This is embodied in the assumptions made by the FCI algorithm, that the model be minimal [39].

Another essential step is attempting to falsify models by testing their predictions. A model which is over fitted to the sample data will fail when applied to data gathered from another sample, or a somewhat broader population. This confirmatory or, more appropriately, disconfirmatory approach is used in SEM when a proposed model or theory is tested against data. The same approach can and should be applied to models which are discovered using this approach.

3.4.2 Testing Relationships

Not only does an equivalence class indicate which relationships are supported by the data, it also indicates which relationships cannot be determined from statistics alone. Statistically indistinguishable models may be distinguished by incorporating background information, temporal information, or experimental results to determine the nature and orientation of the remaining relationships. By indicating which relationships require experimentation to determine the equivalence class acts as a guide to which relationships to test experimentally. The clear implications of the model also allow a researcher to see what relationships are being asserted, and devise tests for those relationships if they seem suspect.

By clarifying exactly what relationships are claimed, this formal representation allows the claims to be more easily understood and argued for or against than comparatively vague natural language specifications or informal diagrams. Describing relationships in such formal detail leads directly to testable predictions. Fully parametrised causal models can be used to make testable predictions about the values of variables in different circumstances. A causal model which is fully oriented and parametrised can be used to infer the outcomes of different circumstances and interventions. If the predictions are being made about measurable variables then these predictions are testable against real world data.

Education researchers can apply the models to different sets of circumstances and compare the results to their intuition, experience, theories, and evidence. If the results appear to be questionable, or if we simply wish to verify the results, observational or experimental studies can be undertaken to falsify or augment the model. The results of any such studies, data or additional constraints on the structure of the model, can be supplied to the causal discovery algorithm to refine or falsify portions of the existing model.

Presuming the model is valid it can be used to guide educational practice and policy by computing the expected results of educational policies. For example a causal model which correctly identifies the relationships between study skills and learning could guide policies towards teaching such skills in the classroom.

Chapter 4

Detailed Design and Methods

4.1 Engineered Model Construction

The first contribution of this thesis is the creation of theoretical and empirical models of SRL. The models are intended as proofs of concept, demonstrating that the relationships in SRL can be represented in the form of graphical causal models. If we are to use causal models of SRL we must first demonstrate that SRL variables and relationships can in fact be represented in this formalism. The creation of a causal model of SRL from the literature acts as a kind of existence proof, demonstrating that the causal structure of SRL can be captured in this way.

Both models have been constructed beginning with a literature review. For the theoretical model I considered theory and review papers from the SRL literature, which I read for definitions of variables and claims or predictions of causal relationships between the variables. For the empirical model I considered empirical papers directly reporting results from observational or experimental studies. The literature on SRL is expansive and, of necessity, both of these models are built from subsets of the available literature on SRL.

When creating a causal graph the appropriate identification of variables is necessary. There is almost always some choice in how to define a variable, in terms of the discretization of continuous variables or aggregating lower level variables into more abstract variables, and these choices change the structure. Omission of relevant variables can conceal causal effects and aggregation of variables which have different causal structures can result in an unfaithful distribution [60]. Identifying the relationships between variables also depends on the presence of other related variables because we determine causal structure by considering

how two variables relate in the presence of additional variables. Therefore, when using only observational data, excluding some variables from the model can limit the ability to discover causal relationships between modeled variables.

In constructing the variables for these models I have attempted to be guided by the common uses and definitions provided by the literature. However, particularly in the case of the theoretical model, it is not expected that the model created corresponds perfectly to either the theory of SRL, or to the correct underlying structure. Even approaching such precision in the theoretical model would require enlisting multiple experts in SRL in a knowledge engineering effort. In the case of the empirical model it would require the inclusion of a larger proportion of the published empirical literature.

4.1.1 Theoretical Model

To construct the theoretical model I performed a literature review of theoretical and review papers in Self-Regulated Learning. Papers were found by searching major journals of educational psychology, following references to additional papers, and searching article databases. 10 theoretical and review papers, published between 1990 and 2007, were selected and used as the basis for the model.

Each paper was processed by a careful reading of the paper, recording any variables or theoretical constructs mentioned, particular definitions of such variables, and any assertions made about correlational or causal relationships between variables, as well as assertions about the absence of a relationship. The relationships within each paper were then graphed as a Semi-Markovian Causal Model (SMCM) for clarity.

Finally an initial complete model was created with all variables for which relationships had been found by straightforward inclusion of all relationships asserted in individual studies. The initial complete model was cyclic, and our methodology only applies to acyclic graphs. When cycles were found they were broken by removing links which conceptually represented links between multiple passes through the SRL phases. When links were proposed in both directions between a pair of variables there are several options. If there is a cyclic relationship between the variables, then one link can be selected for exclusion as above. In any cases where this is not apparent the most commonly suggested link can be chosen. This situation did not occur in practice.

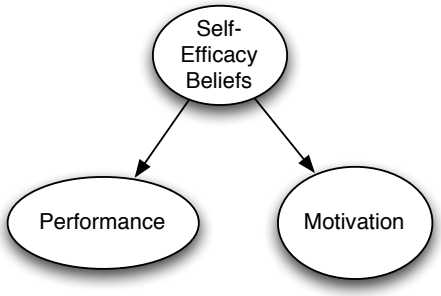
Figure 4.1 on page 43 and Figure 4.2 on page 44 present several of the particular relationships identified in the theoretical literature. In each, a quote from the literature makes

an assertion about the relationship between two or more variables, and is accompanied by a small causal graph, representing the structure suggested by the statement. For example, Figure 4.1(a) presents a quote from [7] that self-efficacy beliefs contribute to both motivation and performance, and a graph showing causal links from self-efficacy to motivation and performance.

Figure 4.2(b) presents a longer quote and a more complex set of relationships including a statement of a relationship "...the motivation of novices can be greatly enhanced when and if they use high-quality self-regulatory processes, such as self-monitoring." and a statement of its mediation by another variable "...their motivation does not stem from the task itself, but rather from their use of self-regulatory processes, such as self-monitoring, and the effects of these processes on their self-beliefs." Here we can see the ambiguity present in constructing a causal model from the literature. The quote indicates a path from self-monitoring to motivation, and that self-monitoring can affect motivation through self-efficacy beliefs, but it does not require or preclude a direct relationship from self-monitoring to motivation. In this case I have chosen to leave out the direct relationship, opting for less relationships assumed if they are not explicitly required by the literature.

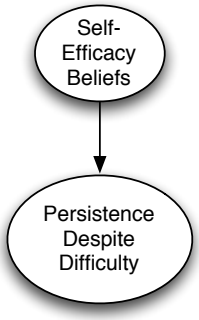
The process of combining the models created from each quote also leaves room for judgement. Figure 4.3 on page 45 shows a simple combination of relationships found in [7]. This version was created by including all of the variables and relationships suggested by the individual models, but without making any additional changes. Many possible changes might be made based on similarity of variables and background knowledge. For instance self-efficacy beliefs is a parent of performance, and also of grades, which are a measure of performance, but performance is not listed as a parent of grades. A common-sense interpretation might be that self-efficacy is not a direct cause of grades, but causes grades via performance. Another possible refinement arises from the multiple variables about goals. Goals, a general variable, is shown as a cause of performance, but goal challenge and self-set goals are not. Intuitively we might expect that if goals cause performance, then self-set goals should cause performance. The case for goal challenge to cause based solely on this graph is weaker and might require further evidence.

"The evidence from these meta-analyses is consistent in showing that efficacy beliefs contribute significantly to the level of motivation and performance. Efficacy beliefs predict not only the behavioral functioning between individuals at different levels of perceived self-efficacy but also changes in functioning in individuals at different levels of efficacy over time and even variation within the same individual in the tasks performed and those shunned or attempted but failed. Evidence that divergent procedures produce convergent results adds to the explanatory and predictive generality of the self-efficacy determinant."



(a) Self-Efficacy, Performance, and Motivation

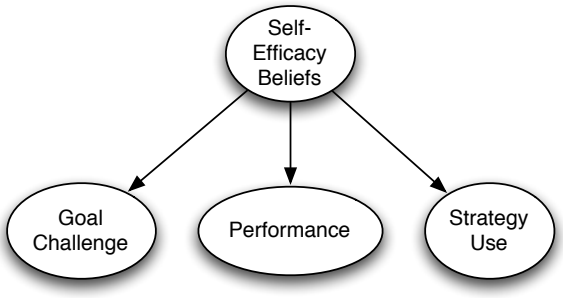
Result
Supported
Experimentally



"One direct way of altering perceived self-efficacy is to introduce a trivial factor devoid of any relevant information whatsoever but that can bias perceived self-efficacy. Studies of anchoring influences show that arbitrary reference points from which judgments are made bias judgmental processes because the adjustments from the arbitrary starting points are usually insufficient (Tversky & Kahneman, 1974). For example, people will judge a larger crowd at a major sports event from an arbitrary starting number of 1,000 rather than from an arbitrary number of 40,000, even though these anchoring numbers are completely irrelevant to judging the size of the crowd. Cervone and Peake (1986) raised perceived self-efficacy by having individuals rate their efficacy from a supposedly randomly selected high number and lowered their self-efficacy from a low arbitrary starting number. The higher the instated perceived self-efficacy was, the longer individuals persevered on difficult and unsolvable problems before they quit. Mediational analyses showed that the biasing anchoring influence had no effect on performance motivation when perceived self-efficacy was controlled. Thus, the effect of the external anchoring influence on performance motivation was completely mediated by the degree to which it changed efficacy beliefs."

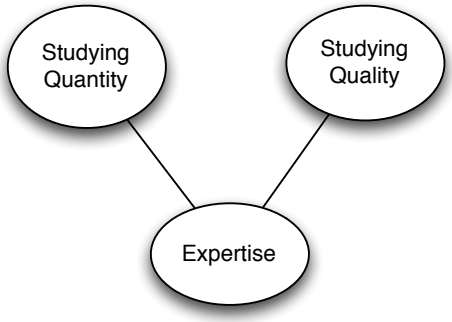
(b) Self-Efficacy and Persistence

Students whose perceived efficacy was illusorily raised set higher goals for themselves, used more efficient problem-solving strategies, and achieved higher intellectual performances than did students of equal cognitive ability who were led to believe that they lacked such capabilities.



(c) Self-Efficacy, Goal Challenge, Strategy Use, and Performance

Figure 4.1: Subset of relationships from [7]

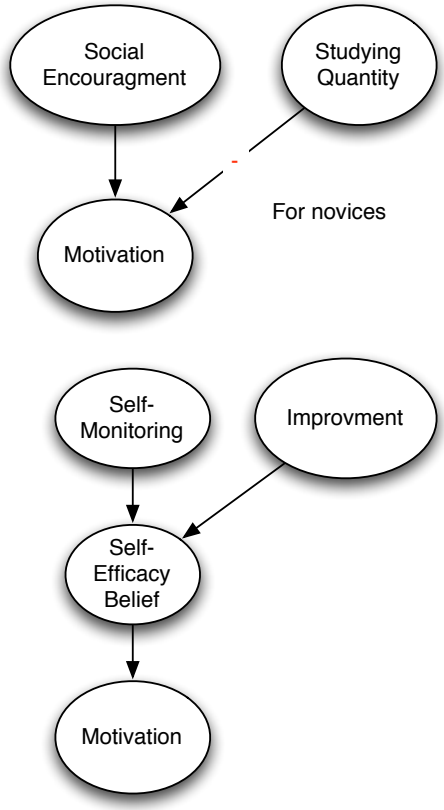


"With such diverse skills as chess, sports, and music, the quantity of an individual's studying and practicing is a strong predictor of his or her level of expertise. There is also evidence that the quality of practicing and studying episodes is highly predictive of a learner's level of skill (Zimmerman & Kitsantas, 1997; 1999)."

(a) Studying Quality, Quantity, and Expertise

However, few beginners in a new discipline immediately derive powerful self-motivational benefits, and they may easily lose interest if they are not socially encouraged and guided, as most music teachers will readily attest (McPherson & Zimmerman, in press).

Fortunately, the motivation of novices can be greatly enhanced when and if they use high-quality self-regulatory processes, such as close self-monitoring. Students who have the capabilities to detect subtle progress in learning will increase their levels of self-satisfaction and their beliefs in their personal efficacy to perform at a high level of skill (Schunk, 1983). Clearly, their motivation does not stem from the task itself, but rather from their use of self-regulatory processes, such as selfmonitoring, and the effects of these processes on their self-beliefs.



(b) Self-Efficacy and Motivational Factors

Figure 4.2: Subset of Relationships from [82]

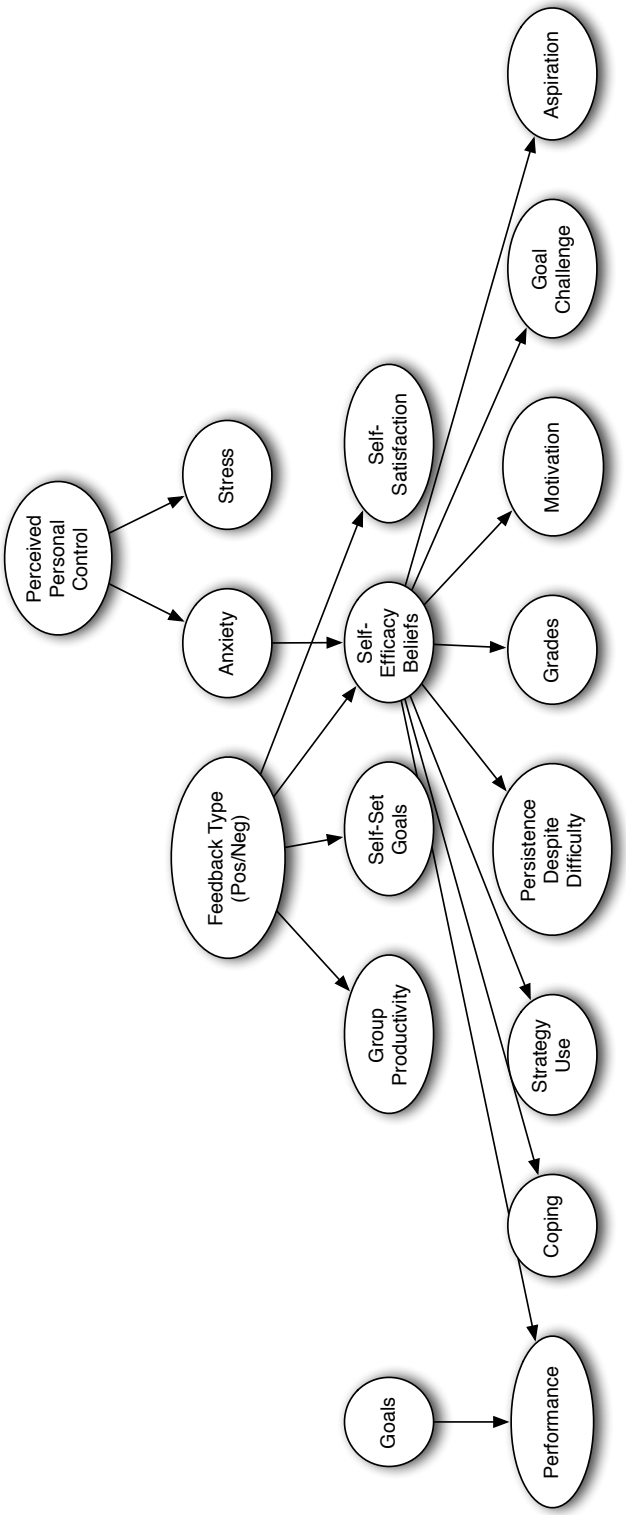


Figure 4.3: Combination of relationships from [7]

This methodology is neither formal nor complete and is not necessarily able to find a unique correct representation of the theory of Self-Regulated Learning. To my knowledge no formal correct method for constructing a graphical causal model from natural language exists, and it is unlikely that any method could reliably produce a single 'correct' model. The SRL literature contains multiple theoretical perspectives and claims, not all of which are compatible. Additionally claims of relationships made in theoretical papers and review papers are typically made in natural language and are not always well defined in terms of the causal or correlational implications of the claims are. For example, Boekaerts and Corno assert that

> "...all theorists assume that there are no direct linkages between achievement and personal or contextual characteristics; achievement effects are mediated by the self-regulatory activities that students engage to reach learning and performance goals."

[10]

This statement has several possible causal interpretations. One possible interpretation (see Figure 4.4 on page 47) is that all causal relationships between personal/contextual characteristics and achievement are mediated by self-regulatory activities. In our causal graphs this would prohibit any direct links between such characteristics and performance, making them independent conditional on self-regulatory activities. Read in this fashion this seems a very strong claim; that once we know the self-regulatory activities of a learner, knowledge of their personal characteristics should provide us no additional information about their academic achievement. However, the specific definition of the very broad terms "personal or contextual characteristics" is not clear, and it is difficult to assign this relationship directly to a particular set of variables. It is also possible that this statement is only intended to apply in situations where individuals are actively self-regulating, and that in less mindful contexts achievement might be directly related to personal or contextual characteristics.

Despite these limitations, what this methodology does produce is one perspective on the current theory of SRL, which could reasonably be construed to be similar to a correct model of SRL. As this model provides a more formal, concise representation of the causal claims of the theory than a narrative approach does, I hope it will be valuable for allowing education researchers to discuss theoretical claims in a formal yet understandable manner, thus enabling it to be improved and to be useful to researchers. The clarity of claims

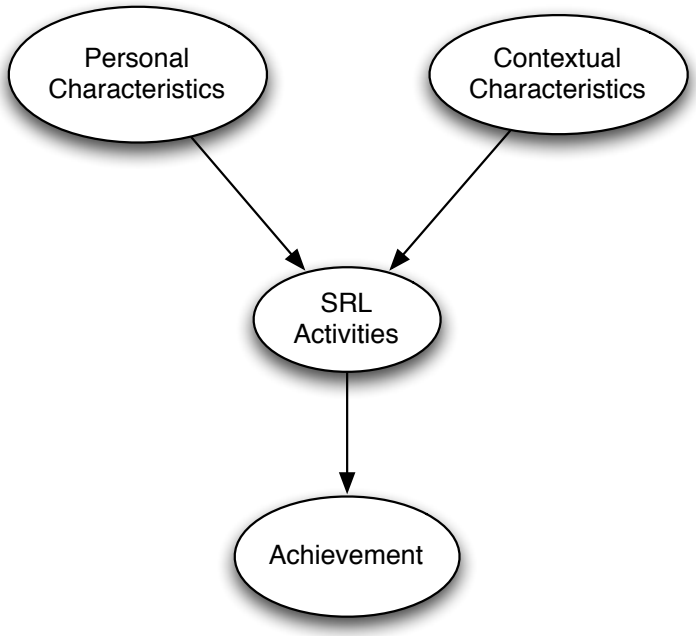


Figure 4.4: Possible interpretation of Boekaerts and Corno

formulated in this manner makes them particularly available for criticism, compared to more vaguely formulated claims.

4.1.2 Empirical Model

Structure discovery algorithms are used to discover the causal structure between variables from data about the variables. The FCI algorithm (Algorithm 3.3.2) and other constraint based algorithm need as their input a set of conditional independence relationships. These relationships can be determined from raw data using any standard statistical test for conditional independence or vanishing partial correlation or a correlation or covariance matrix can be supplied if known.

Existing observational studies can provide the necessary information about covariance or correlation, though the raw data is not typically available. We can also attempt to collect the results of multiple individual studies, conducting a meta-analysis of many studies to determine the conditional independence relationships more accurately by increasing the number of studies included, subject to the usual limitations of meta-analysis. The ability to combine multiple studies is especially important when considering large networks or highly improbable events both of which require larger samples to deal with effectively.

As an initial demonstration of the potential to discover an empirical model of SRL from existing empirical results I investigated existing empirical papers and meta-analyses. I present the results of the FCI algorithm when run on a correlation matrix from a single meta-analysis of SRL related variables found in the literature. The FCI algorithm is used because it is the most informative and best understood constraint based algorithm which allows for the presence of confounding variables and is efficient enough to be used on realistically sized models. The FCI algorithm, as described in Section 3.3.2, uses a correlation matrix to determine a Partial Ancestral Graph (PAG) over the variables, which represents the equivalence class for the supplied data.

4.2 Analysis

4.2.1 Equivalence classes

As described in Section 3.2.2 the equivalence class for a model provides the upper limit on what causal relationships can be discovered from observational data given the assumptions

made. The equivalence class for the theoretical model then informs us of the relationships which can be discovered using observational data, and which cannot. It provides the best case scenario for our search algorithms. I find the equivalence class of the theoretical model by using the FCI algorithm directly on the conditional independence relationships represented by the theoretical model. The FCI algorithm produces a Partial Ancestral Graph (PAG) which represents the equivalence class for the model. I then compare the equivalence class to the theoretical model, assessing the number of correct adjacencies and arrows (directionality).

4.2.2 Model Comparison

The empirical model is a PAG output by the FCI algorithm. In order to evaluate it I conduct a comparison of the empirical model against the PAG representing the equivalence class of the theoretical model over the same variables for adjacency and orientation matching. This comparison provides an idea of how well the two models match, given the data that was available to produce the empirical model. I also present a visual comparison in order to compare which specific relationships were found to be different.

4.2.3 Simulation studies

Algorithms for discovering causal structure from data are only useful if the amount of data required can be reasonably obtained in practical situations. Theoretical results and simulation studies have shown that simple structures can often be discovered up to the point of observational equivalence with sample sizes between 1000 and 10000 [60]. To evaluate the possibility of learning the causal structure of SRL theory from observational data, I conduct simulation studies over the engineered theoretical model to establish approximately the quantity of observational data required to correctly recover the equivalence class. If the model accurately reflects the theory, or has a similar structure and sparseness, this should provide an idea of the quantity of data required to learn the model from real data. The simulations are run using the TETRAD IV software package [52].

Since the engineered theoretical model is not parametrised, and to avoid any biasing effects from a particular parametrisation of the variables, each simulation run used a different randomly generated parametrisation. Each variable was assumed to be discrete, and to take between two and four values. 30 or more simulation runs each were done with samples of

1000, 2000, 5000, 10000, 20000, and 50000 complete data instances. A complete instance of data is a vector with one element for each variable in the model. Formally, for a model $< G, P >$ over $\mathbf{V}$, a complete data instance $\mathbf{D}$ is $\mathbf{D} =< V_1, \ldots, V_i >$. It can be considered a simultaneous measurement of all of the variables in the model. The PAGs produced at each sample size were compared with the PAG produced directly from the conditional independence relationships, as well as being compared with the theoretical model itself.

In the terminology used in the comparisons, an adjacency is an edge between two variables, possibly directed. An arrow point is an arrow mark as the end point on an edge.

4.2.4 Recorded Factors

Correct Adjacencies (ADJ_COR) The number of adjacencies which are present in the true graph or reference graph, which are also present in the discovered graph.

False Positive Adjacency (ADJ_FP) The number of errors of commission in the discovered graph. That is, the number of adjacencies present in the discovered graph which do not exist in the true graph.

False Negative Adjacency (ADJ_FN) The number of errors of omission in the discovered graph. That is, the number of adjacencies present in the true graph which are absent from the discovered graph.

Correct Arrow Points (APT_COR) The number of arrow points which are present in both the true and discovered graphs.

False Positive Arrow Points (APT_FP) The number of errors of commission of arrow points. That is, the number of arrow points which exist in the discovered graph, but not in the true graph.

False Negative Arrow Points (APT_FN) The number of errors of omission of arrow points. That is, the number of arrow points which exist in the true graph which do not in the discovered graph.

False Positive Arrow Points on correct adjacencies (APT_AFP) The number of arrow points which do not exist in the true graph but are present in the discovered graph, considering only adjacencies which are present in the true graph.

False Negative Arrow Points on correct adjacencies (APT_AFN) The number of arrow points which exist in the true graph but are not present in the discovered graph, considering only adjacencies which are present in the true graph.

The median values and the standard deviation was recorded for each of these measures at each of the levels of data generation. Additionally representative individual graphs are compared.

4.2.5 Theoretical analysis of experimental information

The equivalence class indicates which relationships cannot be oriented completely using observational data alone, hence telling us exactly which relationships require experimental investigation. We can assess the maximum number of perfect single experiments needed to completely orient the PAG models using the rules provided in [26]. Because the edge orientation rules could be re-run after each experiment is conducted, the number is a worst case bound on the necessary number of experiments. We compare this with the worst case scenario given no causal relationships being discovered from observational data. These considerations apply only to ideal experimentation. Non-ideal experiments may contain errors or sampling variation which requires repetition of those experiments to verify. The need for replication of results is of course a requirement of any scientific study, not just those employing the methods described herein.

Chapter 5

Results and Discussion

In this chapter I present the results of my investigation into discovering causal models of SRL.

5.1 Engineered Network (Theoretical)

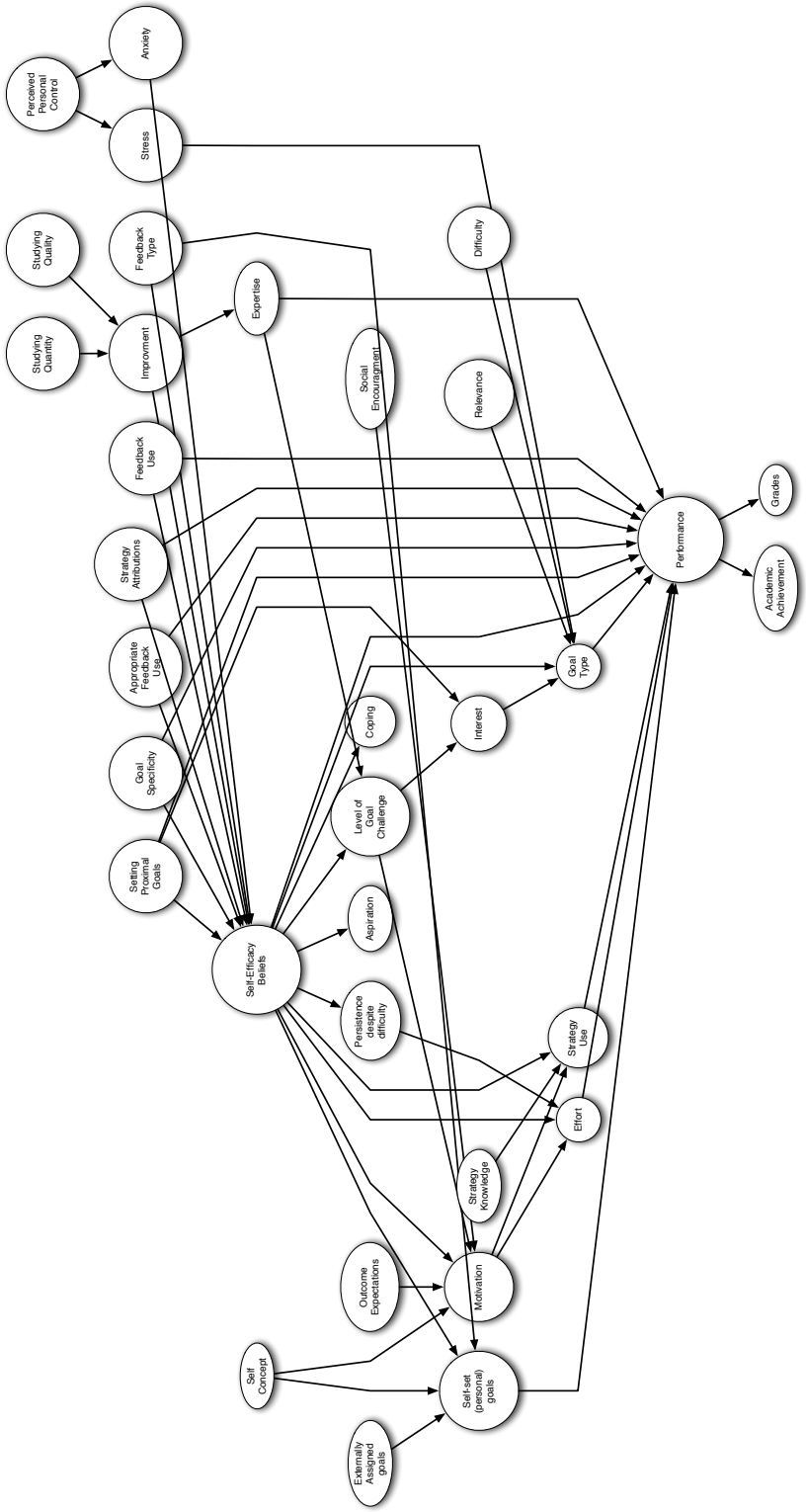


Figure 5.1: Full Engineered Theoretical Model

The initial form of the theoretical model based is directly on the complete set of claims from the literature review. The graph is presented in Figure 5.1 on page 53. This model depicts the structure derived from relationships asserted in the theoretical literature. The variables included are a significant subset of the possible variables in SRL. This model does not include all SRL variables and is not a unique description of the variables involved. In addition to the multiple possible interpretations of the SRL theories, there are multiple ways of operationalizing and aggregating the variables depending on the area of interest.

For instance, in the presented model their are numerous variables representing different aspects of goal setting. These variables could be aggregated into fewer variables by combining variables representing the quality of goals from an educational standpoint into a single scale variable, or divided into more variables of interest, for example by breaking down goal orientation into the various orientations and their degrees, as needed for representational power and computational efficiency. My goal is not to provide a definitive model, but to provide a demonstration of the ability to learn such representations in SRL, and of the understandability of such models.

Unfortunately a large number of variables were described as causing self-efficacy beliefs and performance without indicating intermediate relationships. As a result the complexity of the initial model was too great to analyze with the FCI algorithm on available hardware within a reasonable period. The process of reducing the complexity of the original model to a feasible level is describe in Section 5.3.4. Figure 5.2 is the first reduced structure created, with 24 variables and 32 relationships included. Deriving this reduced model was necessary for finding the equivalence class and running simulation studies. The full model is still the best representation to use for purposes of discussing SRL in causal terms and evaluating the relationships found

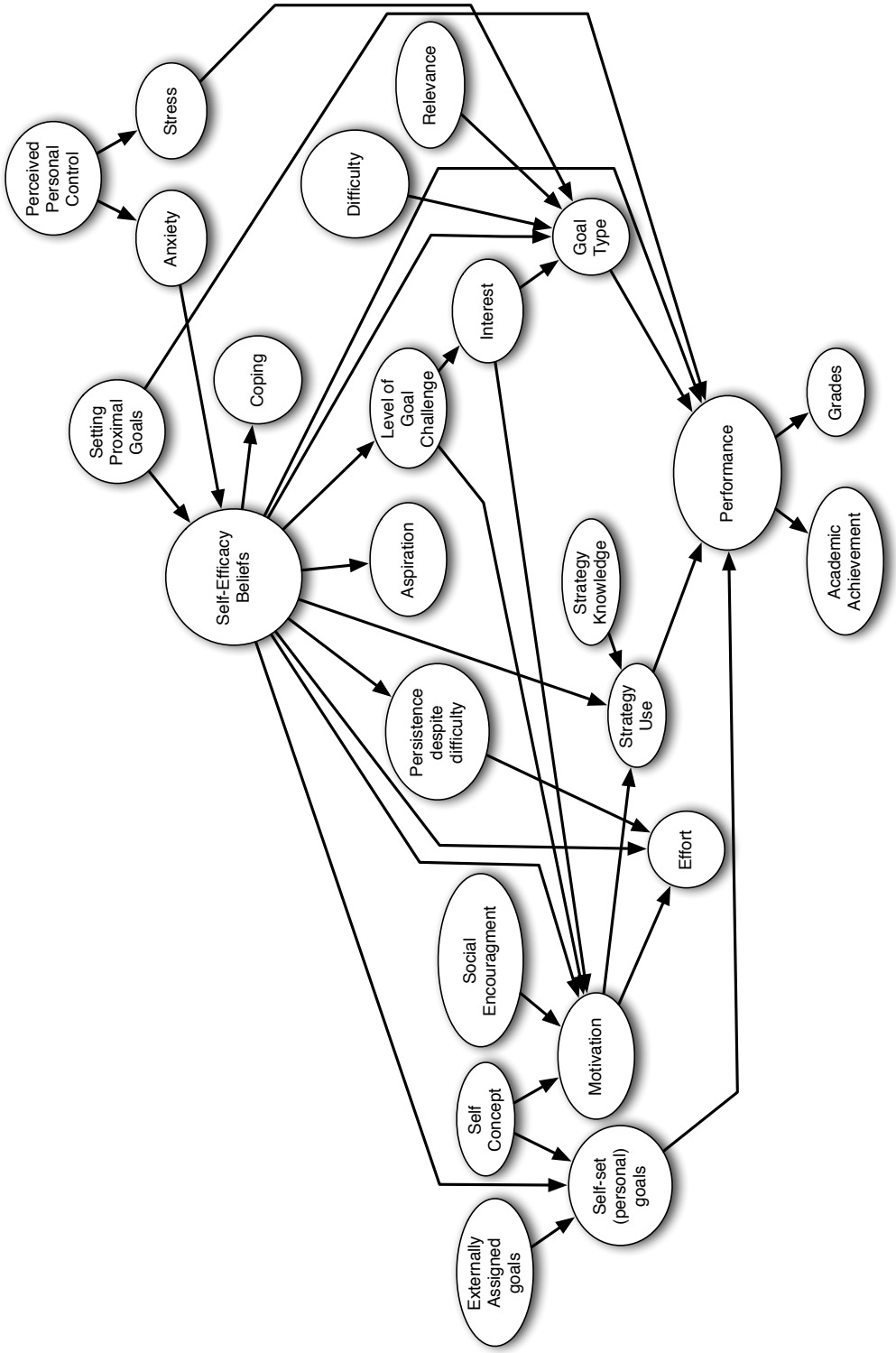


Figure 5.2: Engineered Theoretical Model

5.1.1 Equivalence Classes

The equivalence class for the model tells us the most we can hope to discover from purely observational data given the assumptions we are making. This demonstrates the ability to discover some causal relationships from observational data without experimentation in a principled fashion. It also allows us to see what relationships require experimentation or background information to orient. Figure 5.4 depicts the PAG produced by running the FCI algorithm directly on the conditional independence relationships represented by the theoretical model from Figure 5.2. This simulates running the algorithm on 'perfect' data without sampling error or bias. When the data inputs to FCI algorithm accurately represent the underlying model the result is the correct equivalence class. Therefore the PAG presented in Figure 5.4 provides an ideal case of what the algorithm can produce as the sample statistics converge on the population values.

The equivalence class is presented in Figure 5.4 and the results of the edge comparison analysis in Table 5.1 and Table 5.2. The algorithm successfully discovered the complete set of 34 adjacencies present in the original graph, representing all the dependency relationships between the variables in the graph and no spurious edges were added to the graph.

All but two arrow points from the original graph were recovered correctly, with the arrow points from *PerceivedPersonalControl* to *Anxiety* and *Stress* being labeled as unknown. No spurious arrowheads were added. The algorithm also managed to fully orient 18 of the 34 adjacencies without any need for background knowledge, temporal data, or experimental results. 18 endpoints remain to be labeled and require additional information. This clearly demonstrates the ability of the FCI algorithm to discover many of the causal relationships directly from observational data.

For the reduced theoretical model created for conducting simulation studies, shown in Figure 5.3, the equivalence class (Figure 5.5) edge comparison results are presented in Table 5.3 and orientation results in Table 5.4. The algorithm in this case again recovered the complete set of adjacencies with no false positives, and recovered 16 arrow points with 3 false negatives and no false positives. Incorporating background knowledge that *SelfEfficacyBeliefs* precede *PersistenceDespiteDifficulty* and *LevelOfGoalChallenge* in a given phase, the complete set of arrow points are recovered. Additionally, without the background knowledge 13 endpoints are unable to be oriented, with the background information 9 endpoints remain unoriented.

As expected, the edges whose orientation was completely undetermined had fewer relationships with other variables than the edges which were oriented, reflecting the fact that the algorithms orient edges based on constraints in the patterns of relationships between variables. This reinforces the need to include potential causes of variables, even if we are only interested in the effects of those variables.

These results indicate that given the assumptions and data which accurately represents the population our adjacency relationships and their absence are likely to be correct. The orientations are less certain, however the propensity for false negatives in the ability to orient edges, and the lack of false positives indicates that the inclusion of an end point as oriented seems to be reliable. These results do not hold perfectly when data is not ideal, as described in Section 5.1.2.

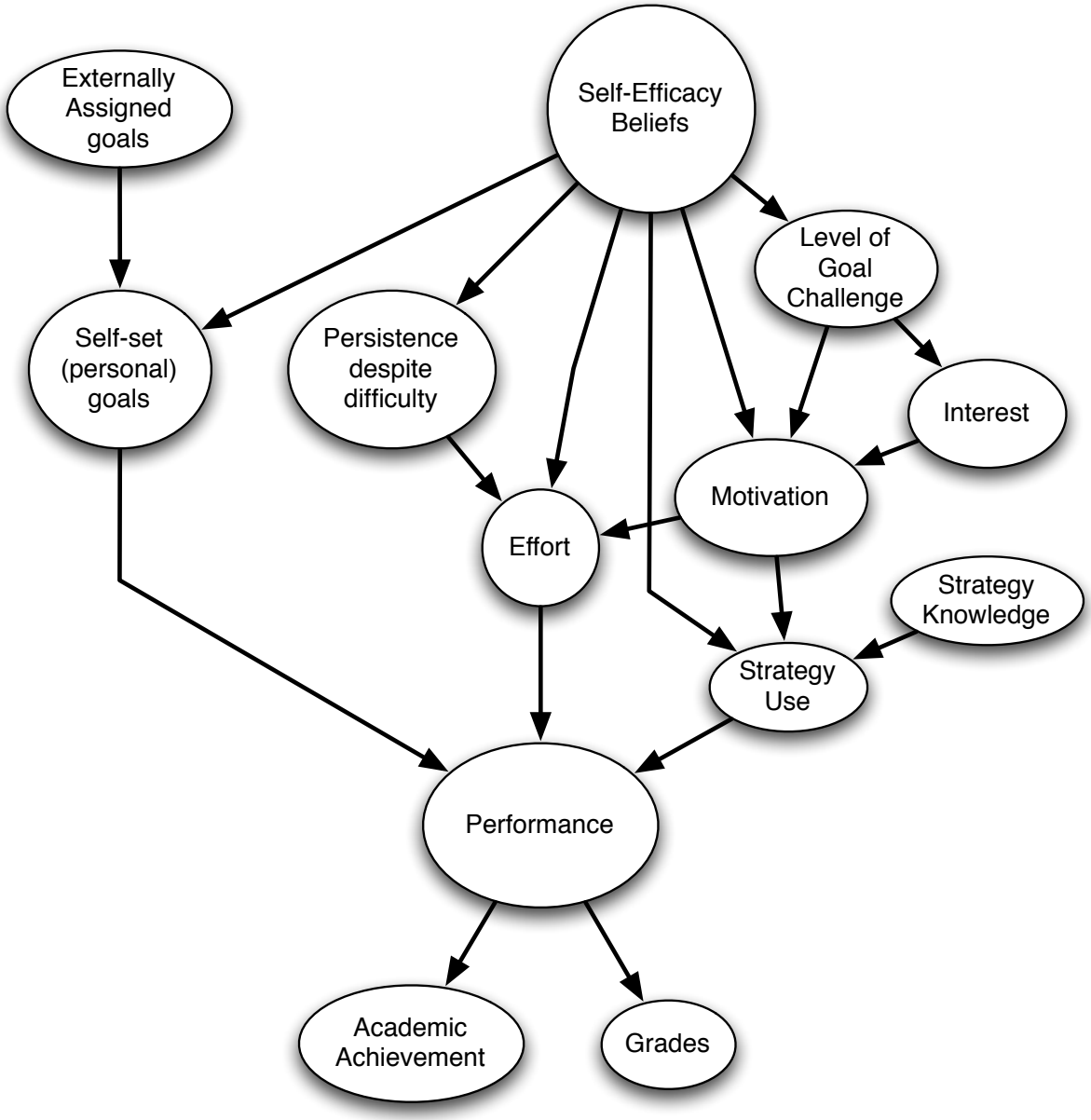


Figure 5.3: Reduced Theoretical Model

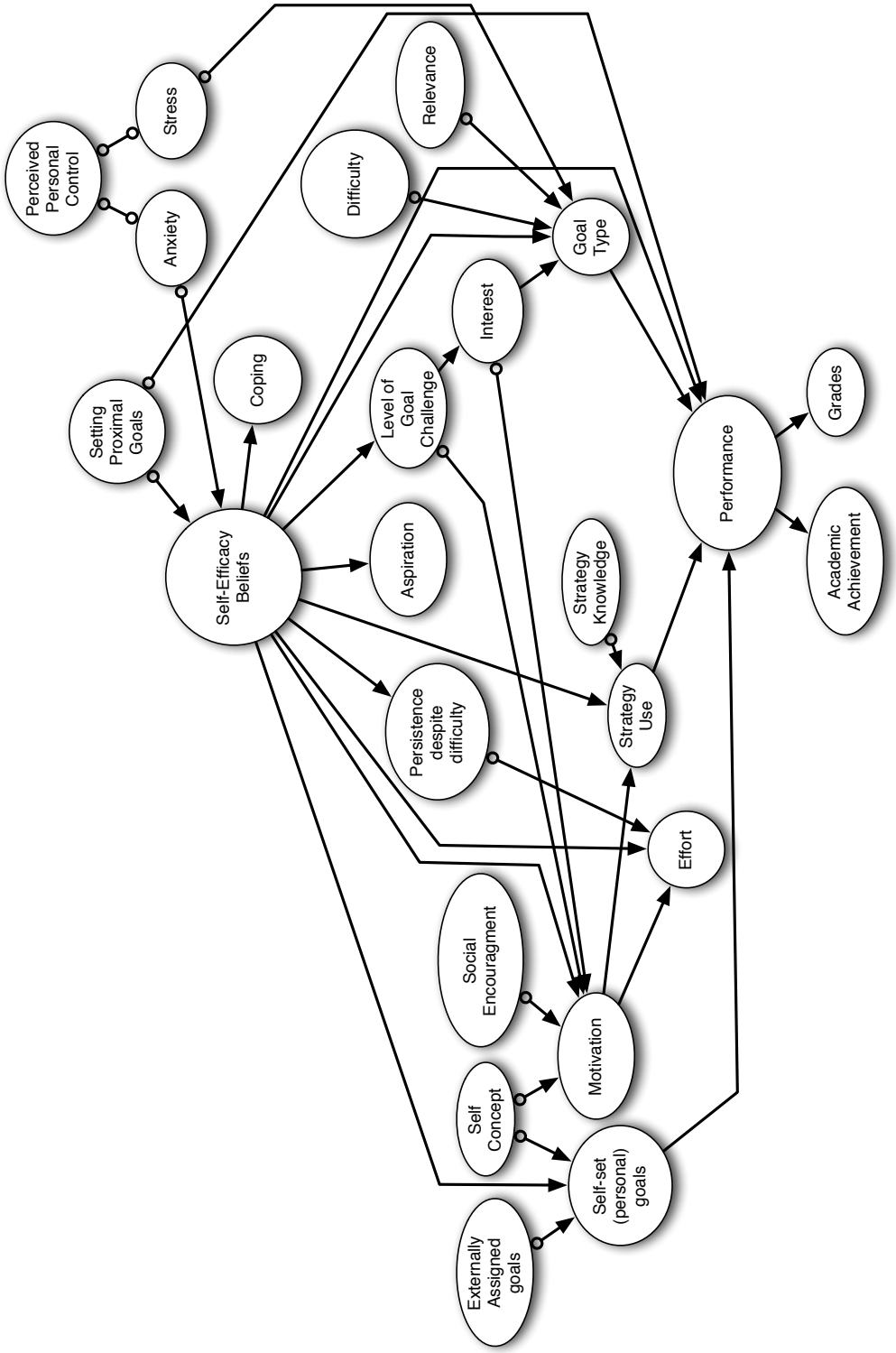


Figure 5.4: Equivalence Class (PAG) For Theoretical Model

ADJ_COR	ADJ_FP	ADJ_FN
34	0	0

Table 5.1: Theoretical Equivalence Class Adjacency Comparisons

APT_COR	APT_FN	APT_FP	APT_AFN	APT_AFP
32	2	0	2	0

Table 5.2: Theoretical Equivalence Class Orientation Comparisons

ADJ_COR	ADJ_FP	ADJ_FN
19	0	0

Table 5.3: Reduced Theoretical Model Equivalence Class Adjacency Comparisons

5.1.2 Simulation Studies

For these causal structure discovery algorithms to be practically useful, them models must be able to represent the information we are interested in, the algorithms must be theoretically correct and able to discover the information, and they must be reliable when given reasonable amounts of real data. The first two criteria have been demonstrated in the previous sections. I now turn to evaluating the amount of data required to reliably discover the equivalence class by running simulation studies on data generated by the theoretical model.

The simulation studies were conducted on the reduced theoretical model due to computational constraints. Given reasonable computing resources the engineered theoretical model would also be possible to compute. A model such as the 'full' theoretical model, with high in-degrees to several variables, would require large scale computing resources to compute effectively, and assuming additional variables with 5 or more parents would rapidly become completely infeasible with existing algorithms and technology.

The data for the simulations was generated and the FCI algorithm was run using TETRAD 4.3.8 on PCs running Windows XP. Simulated data was generated based on the reduced theoretical model at sample sizes from 1000 up to 50,000 data items. At least

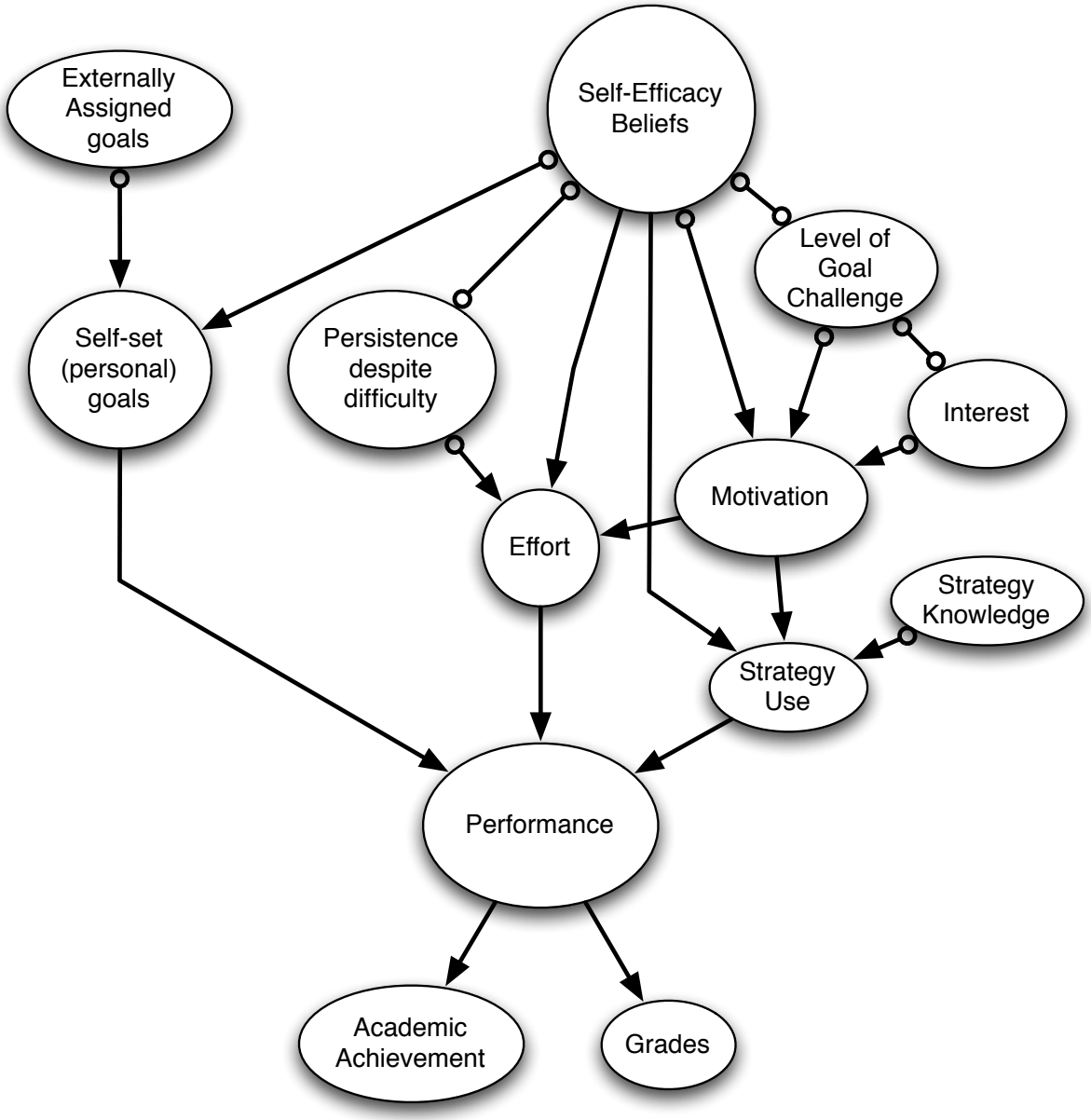


Figure 5.5: Equivalence Class (PAG) For Reduced Theoretical Model

APT_COR	*APT_FN*	*APT_FP*	*APT_AFN*	*APT_AFP*
16	3	0	3	0

Table 5.4: Reduced Theoretical Model Equivalence Class Orientation Comparisons

30 simulation runs were completed for each sample size. Edge comparisons for each run were produced using TETRAD, and the median and standard deviations of the comparison scores were calculated using Microsoft Excel. The total computing time used for the simulations was roughly 100 hours.

Data Tables

	ADJ_COR		ADJ_FP		ADJ_FN	
	Md	Std	Md	Std	Md	Std
1000	13	1.37	0	0.30	6	1.36
2000	10	3.12	0	0.00	4.5	1.56
3000	13	3.47	0	0.24	3	1.76
5000	17	1.29	0	0.15	2	1.33
10000	18	1.15	0	0.17	1	1.11
20000	19	0.71	0	0.24	0	0.70
50000	19	0.49	0	0.15	0	0.54

Table 5.5: Simulation Adjacency Comparisons with Reduced Theoretical Graph

Analysis

The graphs discovered from the simulated data were compared both with the true graph and with the equivalence class for the graph. The results of the simulation comparisons against the reduced theoretical model are presented in Table 5.5 and Table 5.6. These results indicate the reliability of the algorithm in discovering the full set of relationships from the graph. The comparisons against the equivalence class are presented in Table 5.7 and Table 5.8. These results indicate the reliability of the algorithm against its theoretical best performance.

With sample sizes of up to 5000 the majority of the adjacencies are correctly identified

	APT_COR		APT_FN		APT_FP		APT_AFN		APT_AFP	
	Md	Std	Md	Std	Md	Std	Md	Std	Md	Std
1000	9	1.94	10	1.94	6	2.46	10	1.94	6	2.52
2000	12	2.30	7	2.30	7	2.17	7	2.30	7	2.17
3000	13	2.52	6	2.52	4	2.85	6	2.52	4	2.71
5000	15	2.05	4	2.05	6	2.12	4	2.05	6	2.12
10000	16	1.97	3	1.97	6	2.85	3	1.97	6	2.85
20000	16	1.83	2	1.76	4	3.24	2	1.76	4	3.27
50000	16	1.41	3	1.41	2	2.43	3	1.41	2	2.44

Table 5.6: Simulation Orientation Comparisons with Reduced Theoretical Graph

	ADJ_COR		ADJ_FP		ADJ_FN	
	Md	Std	Md	Std	Md	Std
1000	13	1.37	0	0.30	6	1.36
2000	10	3.12	0	0.00	4.5	1.56
3000	13	3.49	0	0.24	3	1.76
5000	17	1.29	0	0.15	2	1.33
10000	17	2.26	0	0.17	1	1.11
20000	19	0.71	0	0.24	0	0.70
50000	19	0.00	0	0.13	0	0.60

Table 5.7: Simulation Adjacency Comparisons with Equivalence Class

	APT_COR		APT_FN		APT_FP		APT_AFN		APT_AFP	
	Md	Std	Md	Std	Md	Std	Md	Std	Md	Std
1000	8	1.77	8	1.77	7	2.75	8	1.77	7	2.80
2000	10	2.20	6	2.20	8	2.62	6	2.20	8	2.62
3000	12	2.43	4	2.43	6	3.59	4	2.43	6	3.51
5000	13	1.94	3	1.94	9	2.82	3	1.94	8.5	2.82
10000	14	1.72	2	1.72	8	3.75	2	1.72	8	3.75
20000	15	1.13	1	1.13	5	4.02	1	1.13	5	4.05
50000	15	0.95	1	1.11	2	3.13	1	1.11	2	3.14

Table 5.8: Simulation Orientation Comparisons with Equivalence Class

but there are large number of false negatives where adjacencies were not correctly identified. False positives were rare, with no false positives being the most common case at all sample sizes. This lends credence to the result from the equivalence class that the presence of an adjacency in the discovered graph is strong evidence for its existence. As sample sizes increased the results for adjacency detection begin to converge to correctly identifying the complete set of adjacencies with very low rates of both false negatives and false positives.

The orientation results are less consistent. The equivalence class recovers 16 arrow points with 3 false negatives and no false positives. At very low sample sizes of 1000 and 2000 only approximately 9 to 12 of the 16 possible arrow points are correctly identified. As sample sizes increase to 5000 and above the algorithm begins recovering all of the arrow points it can correctly recover, matching the equivalence class. However, the algorithm produces a large number of false positive arrow points at low sample sizes, and false positives continue to occur, even at sample sizes of 50,000. This limits the confidence we can have in the orientations produced by the algorithm, particularly at low sample sizes. One possible technique to reduce the number of false positive orientations is to include background information about links which we are certain are not allowable due to temporal relationships or other constraints between the variables. Further study should investigate the effects of including such temporal information on simulation results.

5.1.3 Experimental Requirements

Using the results of [26] we can evaluate the PAGs discovered by FCI to see how many experiments are necessary in the worst case and which ones are required. The number of experiments required is roughly equivalent to the number of undetermined endpoints (endpoints labeled with a o). Evaluating the PAG learned from the conditional independence relationships we see that we need experiments for 18 relationships.

In the worst case assuming the theoretical graph is correct we would require $(n-1)$ experiments, in this case 68 experiments. Both of these estimates are worst cases, and actual results should require less ideal experiments due to application of orientation rules, though possibly more experiments due to the need for repetition and verification of results. The reduction is ideal experiments required is dramatic, with 73% of the experiments no longer necessary.

5.2 Engineered Network (Empirical)

We have now demonstrated that graphical causal models can represent many of the relationships in SRL, and the reliability of the techniques for discovering those models and relationships with simulated data. For this technique to be useful we still require some means of gathering sufficiently large samples for the algorithm to be useful. To begin to demonstrate this I consider a correlation matrix from an existing meta-analysis as the source for the FCI algorithm.

Robbins et al. present the results of a meta-analysis of 108 papers relating psychosocial and study skill factors to college outcomes [50]. From the correlation matrix they present it is possible to directly run the FCI algorithm and investigate the results. Figure 5.6 presents the results from running the algorithm on a subset of the variables they present, excluding two variables due to insufficient sample size, and one due to irrelevance to SRL. The aggregate sample sizes for the correlations vary from as low as 110 up to approximately 17,000. The correlations for variables are low enough that we do not expect the inferences to be extremely reliable. In particular this is the case for relationships between *AcademicRelatedSkills* and *AcademicSelfEfficacy* and between *AcademicRelatedSkills* and *GeneralSelfConcept*.

As can be seen from the double headed arrows in Figure 5.6, multiple variables are identified as having latent common causes. From the provided data the algorithm is unable to completely orient any edges other than those which represent the presence of confounding variables. For many of the variables it is reasonable to expect confounding in this example, given the presence of several indicators of performance or ability such as *GPA*, *ACT/SATScores* and *HighSchoolGPA*. Overall, it is difficult to draw conclusions from such a study network due to insufficient data. Given the known convergence of the algorithms at large sample sizes and the demonstrations of the results from the simulation study, I suggest that given increased data the results would be more informative. Sample sizes of 17,000 as available for some correlations are sufficient to be reliable for the adjacencies, and becoming reliable for the presence and absence of orientation information, unfortunately the low sample sizes for the other correlations can produce errors which propagate throughout the network.

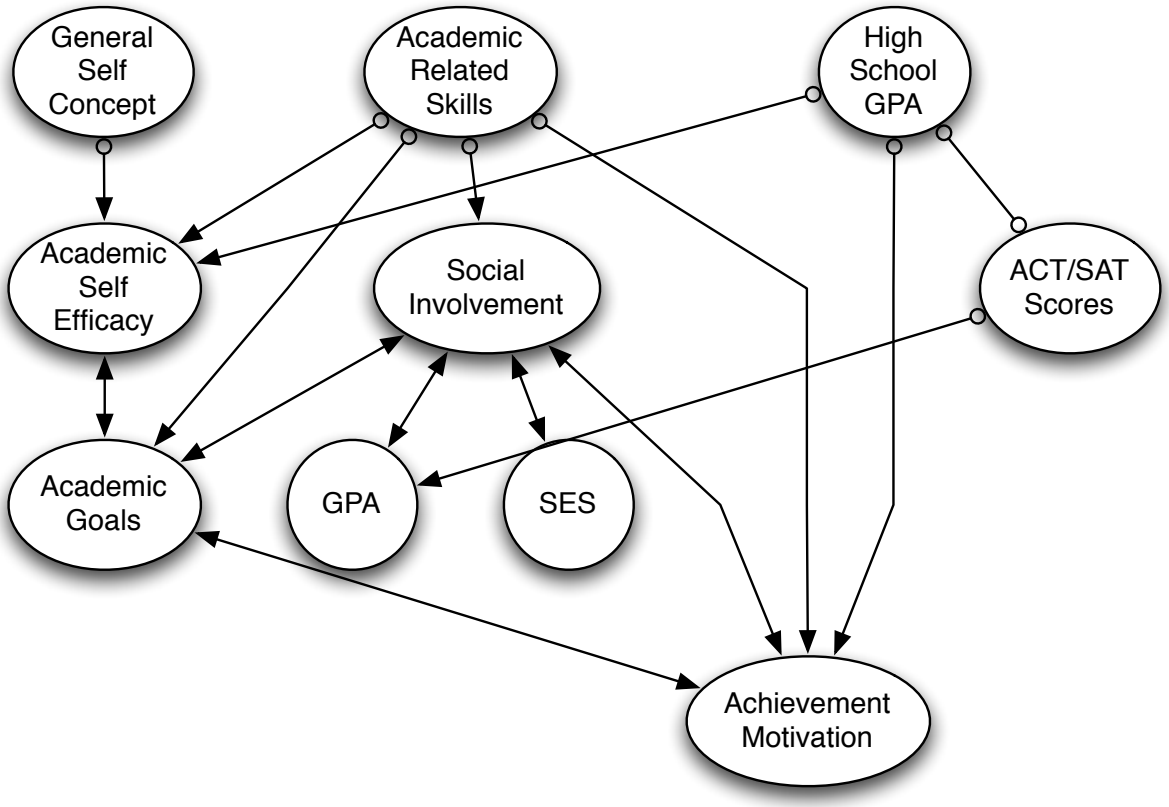


Figure 5.6: Equivalence Class (PAG) derived from Robbins et al.

5.3 Limitations

5.3.1 Quality of Data

One of the foremost limitations that these methods face common to any data based investigation, that is, the quality of the conclusions is limited by the quality, and quantity, of the data. When using causal discovery algorithms we are limited by the size of the sample, the statistical power of the simple correlations, the accuracy of the measurements, and the difficulty of evaluating high order conditional independence relationships from a reasonable amount of data. The solution to these problems is the same as in any observational study: collect more data, and collect better data. One means of attempting to gather more data is to conduct a meta-analysis of many studies of the same variables. By increasing the effective sample size we address the issues above, increasing the reliability of the conclusions reached by the algorithms. When attempting to combine data from many studies by reviewing the literature we face the same difficulties as in conducting a meta-analysis. Studies may be drawn on different populations, may have different biases, use different instruments, and differing definitions. In the ideal case we would collect a large amount of data using consistent techniques in a single study, either over time from a smaller sample, or from a large number of individuals. Computer technology for monitoring students during their learning process offers a potential solution to this need for data.

5.3.2 Assumptions

Whether the assumptions required by the representation and algorithms are valid is a major question which must be addressed when employing these methods. The first assumption, that the Causal Markov Condition holds for the underlying distribution, is the least controversial. The second assumption, the Causal Faithfulness Condition, also called the Stability condition, has more exceptions. The Causal Faithfulness Condition in essence requires all independencies which hold over the distribution to be structural. That is, they should result from stable mechanisms in the data generating process as opposed to coincidental combinations of parameter values which perfectly cancel out to produce independence. For a detailed consideration of this assumption, see for example [75]. The other assumptions of the algorithms are less philosophical and more practical.

The algorithms require that the same causal relationships hold for the entire population

under study. Mixing sub-populations with different causal relationships can result in a graph which is inaccurate and often complete. For example in [82] studying is described to have a negative impact on the motivation of novices, and a positive impact on the motivation of experts. Given the variables included, this indicates different (opposite!) causal relationships between studying and motivation for experts and novices which may result in a causal relationship which is not correct or even a conclusion of independence. Such difficulties can be often be resolved by including appropriate variables and mechanisms for consideration, in this case a variable representing degree of expertise. A more problematic case occurs when the direction of a causal relationship may change, with variable A sometimes causing B and sometimes caused by B.

The algorithms will only produce correct results given correct statistical decisions. That is, if the determinations made of conditional independence or vanishing partial correlation are incorrect, errors may be made in edge inclusion and/or orientation. This is a general problem faced by any statistical procedure.

5.3.3 Needed Theoretical Advances

For a model with a large number of variables, running the tests for conditional independence at conventional significance levels may result in multiple incorrect results given the large number of such tests required. Increasing the thresholds for significance of the statistical decisions changes the type of mistake likely to be made, as correct results may not meet significance thresholds. Given the reliance of the algorithms on patterns of such results, changing the significance of the decisions can produce very different results from the algorithms. At present I am not aware of any exact characterization of the reliability of the FCI algorithm or related algorithms in the face of inaccurate data or violation of the assumptions.

When causal sufficiency is assumed, results exist for calculating bounds on the number of experiments necessary and sufficient, and which experiments are most informative. However no such results have currently been reported for the general case where confounding variables are allowed.

5.3.4 Computational Complexity

The FCI algorithm is exponential in the in-degree (number of parents) that nodes have in the graph. For a sparse graph the algorithm runs in a reasonable amount of time, however the algorithm quickly becomes infeasible for graphs with many parents. This is directly related to the issue Bayesian networks face with large conditional probability tables with graphs have high average in-degree.

For example, in the version of the engineered theoretical model shown in Figure 5.7, the $Performance$ node has in-degree of 13. On a current modern PC, graphs with average in degree of up to approximately 5 may be run in a reasonable period of time. For a graph which a large number of nodes (say 100 or more) and an average in degree of 10, a supercomputer would be required to run the FCI algorithm. There are several means of mitigating this limitation.

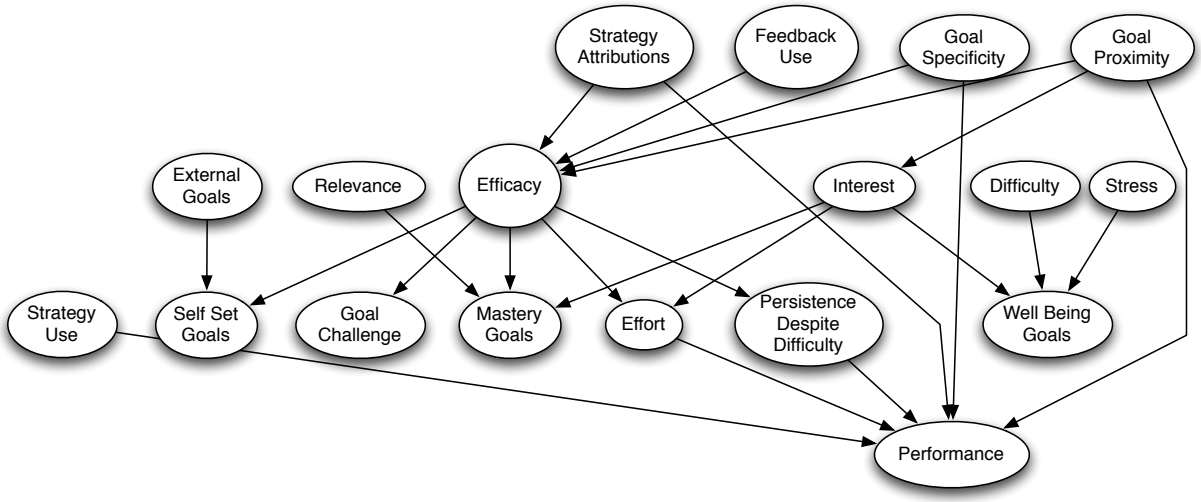


Figure 5.7: Theoretical Model with High in-degree

Removal of variables

Removal of variables can simplify the model by eliminating edges between the removed node and other variable which may have high in-degree. The effects of removing variables depends on their relationship to other variables in the model. The least problematic case is when the

removed variable is an exogenous variable with only one child in the network. In this case the effects of the variable become part of the error value for its child variable, which prevents us from directly considering the effects of the removed variable. A second undesirable effect is that the removal of a parent may prevent us from drawing a conclusion about causal relationships between the child which remains in the network and other variables.

If an exogenous variable that is a parent of two or more variables is removed from the model it becomes a latent common cause (confounder). In addition to the problems which result from removing a exogenous variable with one child, we now face the situation in which we may not be able to orient relationships between the children of the removed variable if they exist, and a confounding relationship may appear even if those variables have no direct causal relationship.

A more problematic case is the removal of endogenous variables, that is, variables which have one or more parents in the model. If the endogenous variable has no children it is called a leaf or a sink and its removal will generally simplify the network. If the leaf has only one parent, then removing it simply eliminates our ability to consider that variable in the network. If it has more than one parent it is a common effect of those parents and by d-separation the parents become dependent when we measure the common effect. By removing it we lose this information.

If such a variable also has one or more children in the model it is called a mediating variable. Removing a mediating variable may actually cause increased complexity, as relationships from its parents are then 'direct' to its children.

Aggregation of variables

Several variables can be aggregated into a single variable which represents them collectively either exactly or approximately. Many variables may also be disaggregated into multiple variables. Both cases may either increase or decrease the average in-degree of the model. If several direct causes of a single variable are aggregated the resulting variable will have only a single relationship with the effect variable, reducing the complexity as well as our ability to reason about the relationships. If several variables which have different parents are aggregated the in-degree of the aggregated variable will be increased, and with it the complexity of the model. Aggregation of variables has an additional risk in that if variables are aggregated incorrectly the resulting model may violate the causal faithfulness condition which is required for the discovery algorithms to produce correct results.

Reducing number of values for discrete variables

For discrete variable the number of possible values the variable can take on is important to the space complexity of the representation and the time complexity of both discovery and inference algorithms. Reducing the number of values possible for a discrete variable thus reduces the space and time requirements of the algorithms, however the time and space requirements grow more slowly in the number of values than in the in-degree so this effect is quickly overwhelmed.

Model Reduction Considerations

The considerations for what methods to use for reducing complexity as necessary vary depending on our purpose. When building a theoretical model for discussion and understanding, the computational characteristics are not important. However if we attempt to use that model for causal inference they become important. In this case we can see the relationships and the effects of modifications by inspection.

If we are attempting to discover causal relationships directly from data we do not yet have information about the causal structure which prevents us from optimizing the structure in advance. However, if we are investigating a small number of new relationships in addition to an existing structure the situation is similar to the theoretical model.

The process I undertook to reduce the complexity of the model for purposes of simulation was a simple process of removing exogenous variables, that is, variables which have no parents included in the network. Since there were many exogenous variables they were prioritized for removal based on being parents of variables with a high in-degree. To some degree this process of limiting the scope of the model will always be necessary due to limitations both human and computational. When designing observational and experimental studies, all of these factors need to be taken into account.

5.3.5 Comparison to Theories

It is worth considering how these causal models relate to the various theories of SRL. I have not limited the models to the consideration of variables emphasized in any particular theory of SRL, instead choosing to include all of the variable I reasonably could. The theories propose variables to be measured and what relationships are expected. The causal discovery techniques provide another way of analyzing the correctness of those theories and

suggesting new possible relationships. The exploratory methods of the causal discovery algorithm contrast to the confirmatory approach typically taken in SEM studies where a particular model is suggested by the theory and is subject to confirmatory analysis checking model fit against data.

The causal models include only causal relationships between the specified variables. These causal relationships are a subset of the information contained in the theories of SRL. For instance, Winne and Hadwin's four-phase model of SRL is inspired by information processing and theory suggests the relationships between the variables is the flow of information and its processing. The social cognitive theory instead focuses on the social context and interactions between individuals to motivate its proposed mechanisms. Both of these models suggest additional variables for consideration, and possible mechanisms which underlie the relationships between those variables.

These are differences in level of abstraction and areas of focus do not impact the correctness of particular causal relationships, though they may suggest other variables of interest which may mediate the relationships if included. For instance, it is reasonable to expect that all social interactions are eventually mediated by the processing of information by an individual whether conscious or not, and that the information processing steps are composed of relationships at the biological level. Any of these levels of abstraction may inspire us to consider new variables and relationships which can then be evaluated and employed using the same techniques discussed throughout this thesis.

Chapter 6

Conclusion and Future Work

In this thesis I have argued for the use of graphical causal models and structure learning to be applied to Self-Regulated Learning and demonstrated the viability and usefulness of such a course. I conducted a literature review of the theoretical and empirical literature of SRL to engineer a theoretical causal model of SRL and generate an empirical model using structure discovery algorithms over the data from the literature. I find that graphical causal models provide a useful means of representing the causal claims underlying SRL theory in a formal and computable form, but they are limited by the availability of sufficient quantities of accurate data.

Using the engineered theoretical model I produced the equivalence class of relationships which can be discovered from ideal statistical data, and generated simulated statistical data of varying sample sizes. I then employed the FCI algorithm to discover the models back from the data. I compared the equivalence class of the theoretical model to the models discovered at different sample sizes by accuracy in adjacency inclusion and orientation. The results indicate that at sample sizes of approximately 5000 complete data items most adjacencies are recovered with several false negatives. The results for recovering arrow points are less positive, with significant numbers of false positives and false negatives at low sample sizes. The accuracy of the methods gradually increased with the sample size, with complete recovery of adjacencies at approximately 10,000 data items and arrow point identification converging on the equivalence class results. However, even at large sample sizes of 10,000 to 50,000 data items the algorithms continue to generate false positive arrow points. The application of more conservative algorithms may alleviate the errors of commission, but at the cost of increases in false negatives. This may be acceptable as it increases the reliability

of relationships which are discovered.

Finally I analyzed the number of experiments necessary to fully orient the model from the equivalence class versus from just the adjacency set normally considered discoverable from probabilistic data and found that the worst case number of experiments for the equivalence class was less than one third of the worst case for the undirected adjacency model. This represents a large improvement in the number of experiments necessary to fully orient such a model.

The exploratory approach taken by the causal discovery algorithms stands in contrast to the confirmatory approach to SEM. The use of a confirmatory approach in which a model is proposed a priori has the considerable limitation of ignoring the equivalence class of models which can equally account for data. The confirmatory approach is appropriate for disconfirming proposed models, but cannot confirm one model over another equivalent model. The exploratory structure discovery approach has the benefit of discovering the complete equivalence class for the available data. A standard challenge of data based methods in machine learning and in science is over-fitting of a model to idiosyncrasies of the data. The FCI algorithm and related algorithms partially overcome this difficulty by incorporating the faithfulness assumption, but may fail to correctly evaluate relationships when this assumption is violated. The models must of course be tested repeatedly in the same fashion as any proposed theory in order to be considered valid.

The creation of graphical causal models representing educational theories offers multiple benefits. They require a clear and precise specification of the claims of a theory and the definitions of the variables, and represent those claims in an understandable form. This formal, understandable representation should allow for clearer specifications of causal claims in the theoretical literature.

We have also shown that it is possible to use causal structure learning to process the results of a meta-analysis of empirical results in SRL and create a causal structure directly from existing observational results. The results were limited by small sample sizes and an incomplete set of variables. Given sufficient time and computational resources the literature of SRL can be meta-analyzed to collect a large proportion of observational and experimental results which provide correlational data for the variables and then the FCI algorithm or a similar algorithm for the cyclic case can be used to discover an equivalence class of causal models. Such a representation has the benefits of indicating what experiments are necessary in order to evaluate un-oriented edges and clearly showing the relationships which can't be

derived solely from observational data. Additionally any new results can be incorporated into the model, producing a integrated, continuously improving model.

6.1 Future Work

In future I plan to improve the models created in this work by a more comprehensive review and analysis of the literature, and by involving experts in the analysis of the theoretical model. By incorporating a large number of studies we can overcome the limitations on accuracy and confidence imposed by small sample sizes. Alternative techniques for causal discovery which make different assumptions should also be applied and their results evaluated. In particular, algorithms which allow for cyclic graphs or time series models may be fruitfully applied to SRL to represent the cyclic nature of SRL instead of considering only a single step. When evaluating existing empirical studies through meta-analysis it may prove effective to make use of related discovery algorithms to evaluate the relationships between multiple measurements of a latent variable and reduce error introduced via that source.

I intend to explore the use of discovered and engineered model as user models in educational technology systems, attempting to aid students both in improving their SRL behaviour, and in improving their knowledge in particular domains. I intend to evaluate the ability of computerized monitoring to collect large data sets which may improve the reliability of the methods. In the course of this work, I hope to perform automated observations and interventions to improve and refine our model of causation in SRL. We hope that by allowing the automated ongoing use of theory based causal models backing educational technology systems we will be better able to gather data and make use of it on an ongoing basis.

Chapter 7

Appendices

7.1 Formal Background

In this section I present a basic review of the background concepts necessary to understand causal modelling.

7.1.1 Graph Terminology

Graph A graph G is a set of vertices V and a set of edges E. An edge E is a pair (possibly ordered) of vertices from the set V. There are many refinements to this definition which specify different types of graphs.

Undirected vs directed edges An undirected edge is a unordered pair of vertices. A directed edge is an ordered pair of edges. Undirected edges are typically rendered graphically as a unmarked line between the vertices. Directed edges are rendered as a line with an arrowhead pointing to the latter vertex of the pair.

Skeleton The skeleton of a graph is a pair **V**, **E** where **V** is the set of vertices and **E** is the set of edges with any directional marks removed.

Cycle A cycle is a directed path with the same node as the start and end points.

Complete A graph is called complete if and only if every pair of nodes in the graph is connected by an edge.

Connected A graph is connected if there is an undirected path between every pair of vertices in the graph.

Path A path is a set of edges connecting two vertices. An undirected path ignores the directionality of the edges on the path, and a directed path follows the directionality of the edges.

Clique A clique is any subset of a graph which is complete. A maximal clique is a clique which is not properly contained within any other clique.

Directed Acyclic Graph A Directed Acyclic Graph or DAG is a graph in which all edges are directed, there are no bi-directed edges, and there are no directed cycles.

Into An edge is into a variable if the endpoint at the edge is marked by an arrowhead.

Collider A node V is a collider on an undirected path U if and only if V is on U and there exist two other distinct variables on U which are both into V.

Unsheilded Collider A collider is called unshielded if the nodes U_1 and U_2 which are into V are not adjacent.

Pattern A Pattern P is a partially directed graph which represents a class of DAGs. A DAG G is in the class represented by a pattern iff i) G and P have the same skeleton ii) If an edge is directed in P it is also directed in G and iii) if a unshielded collider exists in G it also exists in P.

7.1.2 Probability and Statistics

There are several philosophical interpretations of probability. In this work I shall take the Bayesian(subjectivist) interpretation of probability, as opposed to frequentist. In the Bayesian interpretation probabilities represent subjective degrees of belief in a proposition, as opposed to actual physical properties as in the frequentist philosophy. Note that the causal discovery algorithms discussed in this thesis do not require the Bayesian interpretation.

The Bayesian interpretation satisfies the three axioms of probability theory

$0 \leq P(A) \leq 1$

$P(certainevent) = 1$

$P(A \vee B) = P(A) + P(B)$ if A and B are mutually exclusive

The central formula of Bayesian probability is Bayes' theorem

$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$

Bayesian inference involves updating (subjective) probabilities based on new evidence, with Bayes theorem taken as a normative rule for updating the probabilities from data.

P(e) is a normalizing constant if P(H—e) + P(not H—e) required to be 1. P(e) = P(e—H)P(H) + P(e—notH)P(notH)

Conditional Probability is traditionally defined in terms of conjunction

$P(A|B) = \frac{P(A,B)}{P(B)}$

$P(B|A) = \frac{P(A,B)}{P(A)}$

however Bayesians take conditional probability to be more fundamental to human understand than conjunctions. Since human knowledge generally comes in the form of conditional statements about the probability of some phenomena give what else we know, joint probability (conjunction) can be calculated from the conditional probability.

$P(A, B) = P(A|B)P(B)$

Marginal Probability and Marginalizing

The marginal probability (also called the prior probability) of a variable A taking a particular value $A = a$ is the probability of that particular value over all the possible ways it can be realized. The marginal probability of $A = a$ over B is the given by $\sum_{b_i} P(A = a|b_i)P(b_i)$ equivalently $\sum_{b_i} P(a \wedge b_i)$. Calculating the marginal probability of $A = a$ over another variable or set of variables is called marginalizing over B.

Law of total probability

Independence (marginal & conditional)

We say that two variables are independent of each other if learning about one does not change our beliefs about the other. That is:

$(X \amalg Y) iff P(X|Y) = P(X)$

This is also called marginal independence.

We say that two variables are conditionally independent of each other if learning about one does not change our beliefs about the other, given that we know some third variable. That is:

$(X \amalg Y|Z) iff P(X|Y,Z) = P(X|Z)$

Joint Probability Distribution

The joint probability distribution specifies the complete state of probability information over a set of variables.

For discrete variables, the joint is a table which provides the probability of each possible instantiation of the variables.

For continuous variables, distribution functions represent the joint instead of tables.

The joint probability distribution quickly becomes unmanageably large, as its size is exponential in the number of variables.

Correlation

The correlation of two variables is a measure of the linear relatedness of the variables. Correlation ranges from 0 (no correlation) to 1 (perfect correlation). Correlation is typically denoted by $r_{x,y}$ where x and y are the variables whose correlation is being measured.

Partial Correlation

Partial correlation is the correlation between two or more variables after controlling for a third variable or set of variables. The partial correlation of A and B given C is denoted $\rho_{A,B|C}$

Statistical Tests

The causal discovery algorithms make use of statistical tests for dependence and independence when drawing conclusions from sample data. Any statistical test for vanishing partial correlation (in the continuous case) or conditional independence (in the discrete case) may be used. I cover two of the most common tests here.

t-tests

f-test

Regression

Regression is a statistical procedure which attempts to fit a curve to sample data of two or more variables. The most common form of regression is linear regression which attempts to

fit a line to the data.

In linear regression the formula is $y = \beta X + u$ where β is known as the regression coefficient and u is an error term.

7.1.3 Graphical Models

Graphical models represent relationships in terms of a graph structure according to a set of axioms.

Bayesian Networks are a graphical model of probability relationships between variables which is both easy for humans to understand and an efficient representation of probability information.

Graphical causal models relate variables and their causal relationships to nodes in a graph and links between them, via a series of axioms defining the relationship. These graphical models show great promise for representing and reasoning about causal information.

Bibliography

[1] Karen Ablard and Rachelle Lipschultz. Self-regulated learning in high-achieving students: Relations to advanced reasoning, achievement goals, and gender. *Journal of Educational Psychology*, 90:94–101, 1998.

[2] Ayesha Ali and Thomas Richardson. Markov equivalence classes for maximal ancestral graphs. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 1–9, San Francisco, CA, 2002. Morgan Kaufmann.

[3] Ayesha Ali, Thomas Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, page 10, Arlington, Virginia, 2005. AUAI Press.

[4] R. Ayesha Ali, Thomas Richardson, Peter Spirtes, and Jiji Zhang. Orientation rules for constructing markov equivalence classes of maximal ancestral graphs. Technical report, University of Washington, April 2005.

[5] John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, N.J., June 1998.

[6] Albert Bandura. *Self-Efficacy: The exercise of Control*. New York : W.H. Freeman, 1997.

[7] Albert Bandura and Edwin Locke. Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, 88:87–99, 2003.

[8] Monique Boekaerts. Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, 7:161–186, 1997.

[9] Monique Boekaerts. Motivated learning: The study of student * situational transactional units. *European Journal of Psychology of Education*, 14:41–55, 1999.

[10] Monique Boekaerts and Lyn Corno. Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology An International Review*, 54(2):199–231, April 2005.

[11] Monique Boekaerts, Els de Koning, and Paul Vedder. Goal-directed behavior and contextual factors in the classroom: An innovative approach to the study of multiple goals. *Educational Psychologist*, 41:33–51, 2006.

[12] Deborah L. Butler and Philip H. Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3):245–281, 1995.

[13] David M. Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, 2002.

[14] T Cook. Randomized experiments in education: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, 24:175–200, 2002.

[15] Gregory Cooper and Changwon Yoo. Causal discovery from a mixture of data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 116–12, San Francisco, CA, 1999. Morgan Kaufmann.

[16] Martin Covington. Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, 51:171–200, 2000.

[17] Michel C. Desmarais, Peyman Meshkinfam, and Michel Gagnon. Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction*, 16(5):403–434, December 2006.

[18] Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Proceedings of AI & Statistics 2007*, 2007.

[19] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 178–18, Arlington, Virginia, 2005. AUAI Press.

[20] Andrew Elliot and Todd Thrash. Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, 13:139–156, 2001.

[21] Jeffrey Green and Roger Azevedo. A theoretical review of winne and hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, 77:334–372, 2007.

[22] Allyson F. Hadwin, Philip H. Winne, and John C. Nesbit. Roles for software technologies in advancing research and theory in educational psychology. *British Journal Of Educational Psychology*, 75(1):1–24, March 2005.

[23] Eric D. Heggestad and Ruth Kanfer. The predictive validity of self-efficacy in training performance: Little more than past performance. *Journal of Experimental Psychology: Applied*, 11:84–97, 2005.

[24] Avi Kaplan and Martin Maehr. The contributions and prospects of goal orientation theory. *Educational Psychology Review*, 19:141–184, 2006.

[25] Robert MacCallum and James Austin. Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51:201–226, 2000.

[26] Sam Maes, Stijn Meganck, and Philippe Leray. An integral approach to causal inference with latent variables. Technical report, INSA Rouen Laboratoire LITIS, 2006.

[27] Subramani Mani, Gregory Cooper, and Peter Spirtes. A theoretical study of y structures for causal discovery. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, 2006. AUAI Press.

[28] Christine B. Mccormick. *Metacognition and Learning*. Handbook of Psychology. John Wiley & Sons, Inc., 2006.

[29] Stijn Meganck, Philippe Leray, and Bernard Manderick. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. In *Proceedings of MDAI 2006*, 2006.

[30] Stijn Meganck, Sam Maes, Philippe Leray, and Bernard Manderick. Learning semi-markovian causal models using experiments. In *Proceedings of the Third European Workshop on Probabilistic Graphical Models*, 2006.

[31] Alessio Moneta and Peter Spirtes. Graphical models for the identification of causal structures in multivariate time series models. In *JCIS-2006 Proceedings*, Advances in Intelligent Systems Research, 2006.

[32] Muis, R. Krista, Winne, H. Philip, Jamieson-Noel, and Dianne. Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology*, 77(1):177–195, March 2007.

[33] Krista Muis. The role of epistemic beliefs in self-regulated learning. *Educational Psychologist*, 42:173–190, 2007.

[34] Kevin P. Murphy. Active learning of causal bayes net structure. Technical report, UC Berkeley, 2001.

[35] Richard Neapolitan. *Learning Bayesian Networks*. 2003.

[36] Peter Nenniger. Commentary on self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology An International Review*, 54(2):239–244, April 2005.

[37] Richard Netemeyer and Peter Bentler. Structural equations modeling and statements regarding causality. *Journal of Consumer Psychology*, 10:83–85, 2001.

[38] Frank Pajares. Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice*, 41:116–126, 2002.

[39] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[40] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann, 2001.

[41] Judea Pearl. Statistics and causal inference: A review. *Sociedad de Estadistica e Investigacion Opevativa Test*, 12:281–345, 2003.

[42] Paul Pintrich. *Handbook of Self-Regulation*, chapter The role of goal orientation in self-regulated learning, pages 452–502. 2000.

[43] Paul Pintrich. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95:667–686, 2003.

[44] Paul Pintrich. A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16:385–407, 2004.

[45] Kevin Pugh and David Bergin. Motivational influences on transfer. *Educational Psychologist*, 41:147–160, 2006.

[46] Minna Puustinen and Lea Pulkkinen. Models of self-regulated learning: a review. *Scandanavian Journal of Educational Research*, 45:269–286, 2001.

[47] Joseph Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, 2006. AUAI Press.

[48] Thomas Richardson and Peter Spirtes. *Computation, Causation, and Discovery*, chapter Automated Discovery of Linear Feedback Models, pages 253–302. MIT Press, 1999.

[49] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30:962–1030, 2002.

[50] Steven B. Robbins, Kristy Lauver, Huy Le, Daniel Davis, Ronelle Langley, and Aaron Carlstrom. Do psychosocial and study skill factors predict college outcomes? a meta-analysis. *Psychological Bulletin*, 130:261–288, 2004.

[51] R Rosenthal and M DiMatteo. Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52:59–82, 2001.

[52] Schienes. Tetrad iv. Computer Software.

[53] D Schunk and P Ertmer. *Handbook of Self-Regulation*, chapter Self-regulation and academic learning: Self-effacacy enhancing interventions., pages 621–650. Academic Press San Diego, California, 2000.

[54] Dale H. Schunk and Barry J. Zimmerman. *Self-Regulation and Learning*. John Wiley & Sons, Inc., 2006.

[55] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning measurement models for unobserved variables. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 543–55, San Francisco, CA, 2003. Morgan Kaufmann.

[56] Herbert A. Simon. *Designing Organizations for an Information-Rich World*, pages 37–72. 1971.

[57] Peter Spirtes. Limits on causal inference from statistical data. In *Proceedings of American Economics Association Meeting*, 1997.

[58] Peter Spirtes. The limits of causal inference from observational data. 2000.

[59] Peter Spirtes. An anytime algorithm for causal inference. In *Proceedings of AISTATS2001*, 2001.

[60] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2nd edition, 2000.

[61] Daniel Steel. Indeterminism and the causal markov condition. *British Journal of Philosophy of Science*, 56:3–26, 2005.

[62] Dirk Temme. Constraint-based inference algorithms for structural models with latent confounders - empirical application and simulations. *Computational Statistics*, 21:151–182, 2006.

[63] Dinand Webbink. Causal effects in education. *Journal of Economic Surveys*, 19(4):535–560, September 2005.

[64] Kathryn Wentzel and Allan Wigfield. Academic and social motivational influences on students' academic performance. *Educational Psychology Review*, 10:155–175, 1998.

[65] Phil H. Winne. Key issues in modeling and applying research on self-regulated learning. *Applied Psychology An International Review*, 54(2):232–238, April 2005.

[66] Phil H. Winne and Allyson F. Hadwin. *Metacognition in educational theory and practice*, chapter Studying as self-regulated learning, pages 277–304. Lawrence Erlbaum, 1998.

[67] Philip Winne. Experimenting to bootstrap self-regulated learning. *Journal of Educatonal Psychology*, 89:397–410, 1997.

[68] Philip Winne. *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, chapter Self-regulated learning viewed from models of information processing, pages 153–189. Lawrence Erlbaum Associates, 2001.

[69] Philip Winne. A perspective on state-of-the-art research on self-regulated learning. *Instructional Science*, 33(5-6):559–565, November 2005.

[70] Philip Winne. How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, 41:5–17, 2006.

[71] Philip H. Winne. Inherent details in self-regulated learning. *Educational Psychologist*, 30(4):173–187, 1995.

[72] Christopher Wolters. Understanding procrastination from a self-regulated learning perspective. *Journal of Educational Psychology*, 95:179–187, 2003.

[73] Jiji Zhang and Peter Spirtes. A characterization of markov equivalence classes for ancestral graphical models. March 2005.

[74] Jiji Zhang and Peter Spirtes. A transformational characterization of markov equivalence for directed acyclic graphs with latent variables. In *UAI*, pages 667–674, 2005.

[75] Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. In *LSE-Pitt Conference: Confirmation, Induction and Science*, 2007.

[76] Barry Zimmerman. *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, chapter Overview. 2001.

[77] Barry Zimmerman and Albert Bandura. Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31:845–869, 1994.

[78] Barry Zimmerman, Albert Bandura, and Manuel Martinez-Pons. Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29:663–676, 1992.

[79] Barry Zimmerman and Anastasia Kitsantas. Developmental phases in self-regulation: Shifting from process goals to outcome goals. *Journal of Educational Psychology*, 89:29–36, 1997.

[80] Barry J. Zimmerman. A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3):329–339, September 1989.

[81] Barry J. Zimmerman, editor. *Self-Regulated Learning and Academic Achievement : Theoretical Perspectives*. Lawrence Erlbaum Associates, Incorporated, 2001.

[82] Barry J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41:64–71, 2002.