Comparing Model Reuse with Model Building: An Empirical Study of Learning from Simulation

Thomas Monks

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Warwick Business School, University of Warwick

February 2011

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisor Professor Stewart Robinson for his constant guidance and support during the PhD. Stewart has helped me 'keep it simple'; something I initially thought I was good at, but quickly learnt that I was not. I have no doubt that without him I would still be languishing in an overcomplicated analysis in a darkened corner of the business school. Not only has Stewart improved my understanding of analysis and simulation he has also shown me that I will never beat (or get near) his marathon and half marathon times.

I would also like to thank my second supervisor Dr Kathy Kotiadis. Like Stewart, Kathy has provided a huge amount of support and guidance during my PhD and was always available if needed. I am especially grateful for the time and support Kathy gave me during Stewart's study leave (walking in the mountains in New Zealand).

Although they are all anonymous, thanks must be given to my research participants. They delighted, frustrated, fascinated and amused me in equal measure. I believe that a lot of them enjoyed the research and that they learnt something from it. I have at least two pieces of evidence to support these conclusions. Firstly, I saw a large number of the participants all took the elective simulation course run by WBS in the months following the experiment. I like to think that the experiment may have contributed to this decision in some manner. Secondly, the participant who compared using dedicated resources to manage patient emergency flows to 'chopping up a cucumber' (I still don't understand it) eventually conceded that pooled

resources worked better.

To my good friends - Rupal Rana, John Broomfield, Emma Stringfellow, Stavrianna Dimitriou and Antuela Tako - all I can say is that the time went too quickly. Thanks for making the experience such an enjoyable one. Thanks to my colleagues and friends in the OR group and elsewhere: the group secretaries, Sue Shaw and Racheal Monnington, who have helped me countless times over the years as my brain gradually turned to PhD mush; Etienne Rouwette for his invaluable help and advice on experiments and psychology; Les Oakshot and Katy Hoad for braving the simulation coursework sessions with me and many more that no doubt I have missed.

Thanks must also go to the ESRC whose financial support allowed me undertake a PhD and travel to a conference in the U.S, while all the time hiding from a recession in the real world. Special thanks go to the WBS OR group for providing me financial support to pay my rent during those final months and the WBS doctoral programme for partly funding my experimentation costs.

Special thanks go to my parents and brother, Tim, for their encouragement throughout the PhD. Thanks to mum for cooking so many amazing dinners to keep me going over that last Christmas while I finished my thesis. Thanks to Tim for providing extremely valuable and amusing distractions from my PhD from time to time as well as understanding my topic. Thanks to Dad, well for just being Dad.

Finally, to my darling Stephanie thanks for your all of your patience with me. Thanks for putting up with me living away for three years, endless hours of talking about simulation (sorry) and my obsession with electric guitars. To put it simply, I couldn't have done it without you.

The blame for any mistakes lies firmly with me.

Declaration of authorship

I, Thomas Monks, declare that this thesis entitled: 'Comparing Model Reuse and Model Building: An empirical study of learning from simulation' and the work presented are my own. I confirm that:

- This work was done wholly while in candidature for this research degree;
- This thesis contains no material which has been accepted for the award of any other degree or diploma in any university;
- I have acknowledged all main sources of help;
- The work described in this thesis has served the material for several published conference papers and presentations which are listed below.

Refereed Conference Proceedings

- Monks, T., Robinson, S. and Kotiadis, K. (2009), Model reuse versus model development: Effects on credibility and learning, in M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds, 'Proceedings of the 2009 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey.

- Monks, T and Robinson, S and Kotiadis, K. (2010). Model Reuse versus Model Development: Effects on Single-Loop Learning. In Proceedings of the OR Society Simulation Workshop 2010

Conference Presentations and Seminars

- Monks. T, Robinson, S and Kotiadis K, (2009). Model Reuse Versus Model Development, Young Operational Research Society Conference 2009 (YOR16).
- Monks. T, Robinson, S and Kotiadis K, (2009). Model Reuse Versus Model Development.Operational Research Society Conference 2009 (OR51).
- Monks. T, Robinson, S and Kotiadis K, (2009). Model Reuse Versus Model Development: Effects on Credibility and Learning. Winter Simulation Conference 2009
- Monks. T, Robinson, S and Kotiadis K, (2010). An Empirical Study of Learning from Simulation. Warwick Business School Seminar Series.
- Monks. T, Robinson, S and Kotiadis K, (2010). Model Reuse Vs Model Development: Effects on Single-Loop Learning. Operational Research Society Simulation Workshop 2010 (SW10).
- Monks. T, Robinson, S and Kotiadis K, (2010). Transfer of Learning in Simulation Studies. Operational Research Society Conference 2010 (OR52).

Signed:

Date:

Abstract

What are the benefits of involving decision makers in simulation model development? Do decision makers learn more about their problem if they are involved in model development than if they had been excluded? This thesis presents an experiment which compares decision maker learning outcomes and process in two different types of discrete-event simulation (DES) study. The first is a traditional simulation project where decision makers take the role of domain experts and are involved in the building of a simulation model through to its use. The second is where a model is reused rather than built. Sixty four undergraduate participants were individually involved in one of three experimental conditions: development of an A&E simulation model and its subsequent use; development of the same model, but with less time for model use; or were presented with the model already developed and asked to reuse it. Participants of each condition were then allowed to run the model, change variables and review results in an attempt to improve the performance of the system.

Learning was measured at two levels: attitude change, to infer learning about a business problem, and transfer of learning, to infer a deeper learning. Results indicated that, firstly, model building aided participant's discovery of aspects of the problem that were previously unconsidered. However, attitudes about these novel aspects of the problem were only converted to transferable knowledge when experimentation was not limited. Secondly, participants that reused the model learnt about the model through quick cycles of experimentation followed by validation,

although these tended to be focused on factors with which participants were most familiar. In fact, model reuse participants learnt more following this approach than by scrutinizing the results of each scenario in detail.

Little empirical evidence exists to support the discussions and the view that involvement in model building aids learning. This thesis contributes to this debate by providing insight into the mechanisms that influence learning. Moreover, results suggest that learning from experimentation when reusing a model is also valid, although the process followed may be different. Of course, there are limitations to the approach used to perform the comparison. For instance, the experiment uses novice decision makers and measures attitude immediately after the experiment. Reflection on such points is used to aid the generation of testable hypotheses that can be explored in future research.

Chapter 1

Introduction

1.1 Overview

Computer simulation is one of the most widely applied Operational Research (OR) techniques. Like all other applications of OR its purpose is to aid individuals, management or otherwise, make decisions through model building and use. Due to the focus on decision making, simulation modellers often discuss the benefit of modelling in terms of learning as opposed to finding optimal solutions to problems. That is, there is a strong belief within the simulation community that computer simulation provides a platform for a decision maker to understand both the consequences of their actions and the dynamic behaviour they observe in the real system. There are three often quoted reasons why simulation models benefit learning.

The first reason is that a model is a simplification of a system (Pidd, 2009). This is advantageous primarily as the real world is far from simple. This complexity does not just stem from the size of the real system; more likely the behaviour of a real system over time is opaque to those that work within it. This can be due to the delays between cause and effect, variation in event occurrences, and interaction or feedback between parts of the system. In contrast to the uncertainty of the real

world a model, simulation or otherwise, is always a concrete simplification of the system it represents (Pidd, 2009). It has a specific model objective(s), explicitly stated assumptions, observable logic and is built by a modeller. Hence the value of simplification as an aid to learning is that the model should be well understood and transparent to modellers and decision makers alike.

The second reason quoted is that a simulation model has a specific objective and can be used to search for ways to meet that objective. Experimentation with model inputs such as available resources, business rules and policies can help identify actions that can be taken in the real world to help improve performance (Pidd, 2004; Robinson, 2004). Of course, one might argue that this experimentation could also be conducted on the real system. If the real system exists, this is true, but data collection may be extremely costly, slow or even dangerous. Model use aids learning as it provides a repeatable procedure, collecting data in a relatively efficient manner, that can be used to explore the consequences of what if scenarios.

The final argument quoted is that in order to build a simulation model it is often *necessary to involve the decision makers* and this can be particularly beneficial for learning (e.g. Robinson, 2004). Initially a simulation modeller may know very little about the problem and the system of study. To build the model the modeller must elicit and validate relevant data from the decision makers and other individuals who work within the system. There is a strong belief that this involvement aids decision maker learning by helping them *think the problem through* more thoroughly or in ways not previously considered. In fact, it is often stated that a large proportion of the learning occurs during model building. In addition to changing individual's attitudes towards action it is believed that this process can aid 'deeper learning' - knowledge that can be transferred and used again. This belief might be described as a *high involvement hypothesis*.

Although there are clear benefits of the approach it is also often acknowledged

that the building of a *computer simulation can be a time consuming process* (Robinson et al., 2004). In fact, given the budget restrictions often found in industry and public sector projects a simulation modeller may seek ways to reduce the scope of a simulation study; for example, only perform a limited amount of experimentation after the model is built. Thus many research projects in computer simulation have considered the problem of how models can be built quicker. One possible approach that has gained popularity is the reuse of an existing simulation model specially designed to be used with multiple similar systems. For example, an NHS trust may wish to evaluate the impact and side effects of a policy in operating theatre management on the rest of the hospital. To develop a whole hospital model from scratch is clearly an expensive and time consuming exercise and may not even be deliverable in time to aid a decision. If, however, a suitable model already existed (Günal and Pidd, 2010, currently provide the only example) that could be reused for the new objective this would benefit the decision maker in two ways:

1. study feasibility would be increased: it is possible to complete the study to time and cost constraints;

2. the time saved could be used to increase the scope of simulation study: more scenarios can be simulated;

In particular, it would seem that the reuse of the hospital model would aid learning based on the two points listed above. The model is relatively simple; hence understandable compared to the real hospital and can be used for extensive experimentation.

At this point it is worth highlighting that model reuse incorporates model use, but the two should not be considered as the same thing. Model use, as already discussed, refers to the process of using a validated simulation model to experiment with system inputs to learn about the impact on system performance. Model reuse is

the process by which the simulation study is run; i.e. the identification, acquisition, validation and use of a suitable simulation model. This process, of course, influences the time available for model use. Thus more experimentation should be possible relative to if the model was built from scratch. One problem, however, may be that reuse of full models appears to be at odds with the high involvement hypothesis: that involvement of decision makers in model building substantially aids their learning.

To date no studies have considered reuse from the perspective of learning. This thesis describes an experiment to test the high involvement hypothesis in a discrete-event simulation (DES) context. If the hypothesis is true then individuals involved in the building of a DES model before its use should learn more than when they are not involved i.e. if they simply reuse a full model for experimentation. Another useful output of such a comparison should be the identification of mechanisms that aid and inhibit decision maker learning. Of course, as the experiment is a simplification of a real simulation study there will be important differences that must be reflected on before results can be interpreted in any useful manner. These reflections should give rise to some new more focused hypotheses of learning in studies where decision makers are involved in model building, use and reuse.

The remainder of this introductory chapter is organised as follows. The following two sections provide introductory discussions of two key concepts used throughout the thesis. The first of these introduces details on the type of learning that is theorised to occur during involvement in model building and involvement in simulation projects in general. The second provides an introduction to DES model reuse, with a particular emphasis on the reuse of generic models. The section illustrates that only a very limited number of model reuse studies have considered the impact of reuse on learning and decision making. The final two sections of the chapter list the research objectives, expected contributions and the structure of the thesis.

1.2 DES Study Involvement: Learning Outcomes

In order to conduct an experiment of learning from simulation, the type of learning outcomes from a DES study must be explored. Within the OR and simulation literature the term learning is often expressed at two levels of outcome, sometimes linked to the concept of transparency. The first is a change in attitude towards strategy or action. For example, a manufacturing manager's attitude towards achieving 100% utilisation of machines in his or her production line may change dramatically after a DES study aiming to increase throughput. The second outcome is inherently linked to the first and represents a change in problem understanding for a decision maker. For example, the same manufacturing manager may initially see that aiming for 100% utilisation of machines is counterproductive for increasing throughput in the factory, but then also gain an understanding of why performance behaves in this manner. This second level of learning should be transferable, to some extent, to similar situations encountered in the future.

Discussions of the benefit of involving decision makers in modelling can be traced back to the 1960's. Churchman and Schainblatt (1965) emphasise that a position of mutual understanding is essential for implementation success. That is, involvement of decision makers or clients is important as modellers must gain an understanding of the decision maker's position as well as the decision makers learning about the model and the modeller's position. This involvement is believed to give rise to transparency (Ward, 1989; Lane, 1994): an understanding of the computer and mental models of the problem by decision makers. It is theorised that this transparency aids reflection on the differences between computer and mental models of the problem to create attitude change and a 'deeper' transferable knowledge. Section 2.2 provides a more detailed literature review of this area.

These types of learning are often referred to as single-loop and double-loop learning respectively (Argyris and Schön, 1996). Single-loop learning is really short term

learning that is used instrumentally to solve an immediate business problem i.e. 'what should we do to improve performance'. Double-loop learning incorporates attitude change, but adding understanding about the causes of the problem. Argyris and Schön (1996) state that double-loop learning is much more difficult for individuals or organisations to achieve than single-loop (the theory is that as it is more comfortable and less challenging to solve instrumental problems) and that often some help is required to stimulate double-loop inquiry. It is often believed that simulation provides this 'assistance' and fosters a degree of double-loop learning (Lane, 1995; Senge, 1990). Chapter 3 explores double-loop learning in detail and provides a framework for understanding attitude change and transfer of learning.

1.2.1 Queuing concepts used in the thesis

This thesis focuses on decision maker involvement in DES modelling i.e. models of queuing systems. The concepts that participants in the experiment learn about related to two key concepts in all queuing systems, resource utilisation and process variation. This section provides an overview of the key queuing concepts that participants learn about in the experiment.

The relationship between resource utilisation, variation and performance

Often in a queuing system, for example a manufacturing line or an A&E department, it is desirable to have a high utilisation of expensive resources, such as machines or staff. However, when a system is subject to process variation, for example, when machine cycle times vary or arrivals of patients are unpredictable, there is a trade-off between the speed at which an entity can travel through the system and the utilisation of resources. In particular at high levels of resource utilisation and process variation a small change in utilisation can lead to large changes in performance.

Reducing process variation

The relationship described above often means that it is desirable to reduce the process variation in queuing systems; for example, reducing the variation in machine breakdowns by implementing a routine maintenance program. The results of such initiatives reduce the severity of the resource utilisation and performance trade-off; meaning that the same performance can now be achieved with higher utilisation of expensive resources.

1.3 DES Study Efficiency: Model Reuse

The opening paragraphs to this chapter discussed the potential benefits of simulation models as an input to a decision making process. That is, learning is aided through the simplified nature of models, through their use and through involvement in their construction. The previous section expanded on this general discussion by providing an overview of the DES approach to simulation, its use for modelling the effects of process variation and its relationship to queuing theory, i.e. the potential for learning about queuing processes through DES.

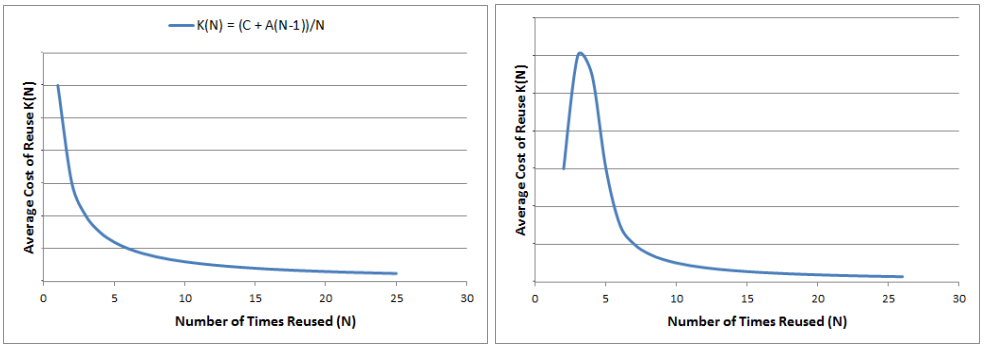
Although DES offers these benefits one problem with the approach in practice is that it can take a long time to design, build and use the model relative to a simple analytical model (Robinson et al., 2004; Pidd, 2004). Indeed some limited data exists to support this view: Cochran et al. (1995) found that a typical project, included in their sample, lasted between one to three months. This duration isn't always acceptable to a decision maker with a limited budget or limited time available; for example, a healthcare professional or a small manufacturing firm may need to make a decision quite quickly. Thus in practice the time and budget available dictates the design, objectives and level of detail included in a model (Law, 1993).

This issue has led to much research into the problem of improving the efficiency of constructing DES models. For example, the three phase simulation method (Tocher

and Owen, 1960), component based simulation (Pidd and Carvalho, 2006), component storage (Arons and Boer, 2001), discovery and retrieval using web services and ontologies (Bell et al., 2008), grabbing and gluing simulations (Paul and Taylor, 2002), reuse at the conceptual level (Balci and Ormsby, 2007; Balci and Nance. N, 2008), reuse of full, possibly 'generic', models (e.g. Günal and Pidd, 2010, 2009; Kaylani et al., 2008; Fletcher et al., 2007; Günal and Pidd, 2006; Sinreich and Marmor, 2004; Brown and Powers, 2000; Pidd, 1992; Pierce and Drevna, 1992) and the validity of reuse (e.g. Spiegel et al., 2005; Malak and Paredis, 2004; Tolk, 1999).

Figure 1.1a illustrates a typical financial argument for reuse adapted from Pidd's model in Robinson et al. (2004). The chart illustrates a smoothed (i.e. assuming a fixed adaptation cost) reduction in costs as the model or piece of software is reused. Although Figure 1.1a intuitively makes sense, quite often software code requires a substantial amount of refactoring (a reduction in internal dependencies of the code, but still producing the same behaviour) before it is reusable in new projects (Fowler, 1999). Dependency problems may also be found in a simulation project where a full model is reused. For example, Robinson et al. (2004) briefly discuss the maintenance planning model originally built for a telecommunications provider reused with a water utility company. However, much of the simulation code was not directly reusable in the new context, possibly due a high dependency of the code to the original problem. Therefore a substantial amount of rework was required. Figure 1.1b illustrates this issue through an alternative hypothetical reuse cost curve. Here a number of costly refactoring efforts are necessary in the early reuse attempts before benefits can be achieved.

A number of DES studies have focussed on developing *generic* models to try and solve the refactoring problem and useability problem. Models can be defined at four generic levels in terms of abstraction and transportability (Fletcher and Worthington, 2009); these are listed in Table 1.1. The first level is the most abstract



(a) Illustration of the cost reduction expected from reuse (adapted from Robinson et al (2004)).

(b) Alternative view of reuse costs

Figure 1.1: Example Reuse Cost Curves. K(N) = average cost/use. C = cost of developing for first use, A = adaptation cost, N = number of times that software is reused. Illustration of the extra effort needed to reuse a model if it is not designed for reuse.

and transportable and refers to *generic principles*; for example, the Department of Health demonstrates a generic lesson to all NHS primary care trusts that patients queue substantially longer when utilisation of beds/nurses and so on rise above 80%. This level is closer to DES and its relationship to queuing theory than an actual computer model analysing the performance of a system.

Table 1.1: Fletcher and Worthington's levels of genericity

Level 1:	A generic principle model; e.g. patients queue longer when resources are utilised at $> 80\%$
Level 2:	A generic framework; e.g. a healthcare toolkit in a simulation package
Level 3:	A setting-specific generic model; e.g a simulation model A&E departments in the England.
Level 4:	A specific model; e.g. a simulation model A&E department x

At the other extreme the fourth level has the lowest level of abstraction and transportability; here models require substantial refactoring before they can be reused

with anything more than the *specific* system modelled. This might be viewed as a model found in a more traditional DES study such as the maintenance model discussed in Robinson et al. (2004).

In between these extremes come generic *frameworks* and *setting-specific generic* models. Frameworks provide the building blocks for models within a domain; for example a simulation package may come with healthcare specific resources and activities included. More generally commercial DES packages encapsulate a generic model, e.g. 3 phase (Tocher and Owen, 1960), and components comprising a framework relevant for all queuing systems and problems. A setting-specific generic model might be, for example, a hospital (Günal and Pidd, 2010) or an A&E department (Günal and Pidd, 2009; Fletcher et al., 2007) that can be reused at multiple NHS Trusts across England. These models tend to be data driven (Pidd, 1992) and can be parameterised with local data for reuse.

Although work has been done to create generic models, the costs and effort reduction benefits of reuse may be difficult to achieve due to the perceived credibility of the simulation model by the decision makers (Robinson et al., 2004). This seems to be particularly problematic when the modeller(s) and decision makers did not build or were not involved in the creation of the model. The so called 'not invented here syndrome' (Sutcliffe, 2003) requires effort in validation and verification by the modellers and decision makers. In fact, it is often believed that this effort may be more than if the model was built from scratch. Hence something similar to the cost curve in Figure 1.1b may be seen in a model reuse study even if a generic model is employed.

Only two DES studies have considered the decision maker or simulation user aspect of reuse. These exceptions are the suggestion to review generic model assumptions with the client by Ozdemirel (1991) and the discussion of client based issues surrounding reuse by Fletcher et al. (2007).

Ozdemirel (1991) presents an expert system called GUIDES that allows users of generic simulation models to review model assumptions. Users either accept or reject assumptions leading to a measure of 'model acceptance' calculated as the percentage of assumptions accepted. The method is presented as a practical approach to measuring acceptance of a generic model. However, this recommendation seems to have little difference to the standard guidance for a traditional simulation study. For example, Law (2007) emphasises that it is essential to review model assumptions with clients in order to build credible simulation models.

Taking a case study approach Fletcher et al. (2007) discuss the use of a generic A&E model with ten NHS trusts across the UK. An unexpected benefit of initial demonstrations of the generic model was that it was the first time that many of the stakeholders had met together to discuss the issues facing their department. Although the model was not parameterised with local data, and instead used high level national data, it was used to facilitate these discussions and help stakeholder gain some insight into the local issues. Fletcher et al. (2007) also discuss issues and obstacles in reusing the model. For example, they note that barriers to successful reuse typically were data quality or low client motivation. This adds an interesting dimension to reuse beyond model validity, but the issues they bring out, such as low data quality, could plausibly appear in any traditional study where a model is developed.

These two studies highlight two points. Firstly, that generic models have been used as an input into a decision making process where stakeholders had no involvement in development. Secondly, variables other than cost should be considered in simulation studies where model reuse is possible. Another factor to consider is proposed by many experienced DES and System Dynamics (SD) authors: high involvement of clients in model development leads to more learning for decision making (e.g. Churchman and Schainblatt, 1965; Robinson, 2004; Rouwette, 2003). If this

high involvement hypothesis is true then we would expect studies where involvement in modelling is minimal, for example, the reuse of a generic model, to produce less learning relative to studies where involvement is high.

1.4 Research Objectives and Contribution

As is evident from the previous section there are competing goals within a DES project. It is important that decision makers and stakeholder are involved and engaged in the building and use of a simulation model to aid learning. In fact, many simulation authors discuss a *high involvement hypothesis*: namely that a large proportion of the benefit of simulation is delivered through involvement of decision makers in model building. At the same time the realities and pressures of decision making push modellers and researchers to seek ever more efficient methods of delivering models to clients. Even the most involved decision maker will have little use for a model that is delivered after the world has moved on.

Due to these competing goals a lot of research has been conducted into model reuse. The existing work has largely been of a technical nature, but has been fruitful in, among other areas (see section 1.3 for a summary), delivering feasible ways to reuse specially designed models. These *generic* models are parameterised with data rather than built from 'scratch' and hence offer the hope of drastically increasing the chances of gaining modelling insights in a useful amount of time.

This thesis focuses on these two competing objectives and explores the differences in learning between traditional DES studies and those where models are reused. At a general level the objective of this thesis is to test the high involvement hypothesis against the benefit of increased experimentation offered by reusing a model assuming a fixed budget of time. If the high involvement hypothesis is true and substantial in effect size, it should be possible to recreate the effect in a laboratory setting. This is done via a comparison of model building (with limited and extended experimen-

tation) against a model reuse process. In addition to the general test of the high involvement hypothesis this research aims to:

1. Empirically identify mechanisms that aid decision maker learning within model building;

2. Empirically identify mechanisms that aid decision maker learning within model reuse and experimentation;

3. Empirically identify mechanisms that inhibit learning from DES models;

Although some limited literature exists exploring learning in simulation studies, little research has been conducted to explicitly compare learning from involvement in simulation model building and (re)use. An explanation for this may the difficulty in finding or gaining access to real comparable simulation studies of model building and reuse. Another reason may be the need for knowledge about the large body of learning research within psychology; a research field in which OR and simulation has, understandably, had limited involvement. To achieve the objectives outlined above this thesis provides an account and details results of an alternative approach to studying learning from DES modelling using a psychology style laboratory experiment methodology. The expected contributions from achieving these objectives are:

1.) An empirical comparison of the impact of model building and reuse on learning

Little empirical evidence exists to support the discussions and view that involvement in model building aids learning. This thesis provides a simple empirical test of the hypothesis that has not been attempted before. Analysis of the experimental results and differences between the experiment and real simulation studies should provide insight into the foundations of this belief.

2.) Insight into how generic models aid learning

Little is known about the impact of generic models on a simulation user's behaviour. The experiment offers a chance to observe the reuse of a model. Differences in the approach to experimentation used by participants involved in model building and reuse should offer insight into how generic models can be used to aid decision maker learning.

3.) Generation of focussed hypotheses on learning mechanisms

The learning mechanisms identified in the experiment should also be testable and thus can be explored in more detail. Future research may use these results to structure an analysis or test methods for improving learning in simulation studies.

4.) A framework for future research

To date no research has attempted to involve participants in the building of a simulation model in a laboratory environment. Development of the materials for the experiment will provide a framework for involving participants in model building and measuring their learning that can be reused and adapted for future work.,

1.5 Thesis Structure

The thesis is organised as follows:

Chapter two provides a review of the OR and simulation literature that has considered learning from modelling. This is considered from three perspectives: learning from model building, learning from model use and relationship between credibility and learning. This review concludes that learning from models can be thought of as a process of improving decision makers understanding of a.) the computer model and b.) their own mental models of a problem. This understanding in turn can lead

to changes in attitude about the course of action to take and a deeper understanding about the underlying behaviour of a system that can be transferred elsewhere. The review also illustrates that although some empirical studies explore learning from quantitative modelling they are focussed on demonstrating that learning occurs rather than how or where decision makers learn in the process. Moreover, little learning research has been undertaken in the DES domain.

Chapter three frames the existing OR and simulation literature on learning in terms of Argyris and Schön (1996) concepts of single-loop and double-loop learning. The chapter links single-loop learning to attitude change and double-loop learning to both attitude change and transfer of learning. Several simulation examples of learning under this framework are provided. This is followed by a review of the attitude change/persuasion and transfer of learning literature from psychology. The chapter concludes with a discussion of a measurement approach for single-loop and double-loop learning.

Chapter four details a psychology style laboratory experiment to measure single-loop and double-loop learning. The chapter firstly describes how participants in the experiment, undergraduate business students, are either involved one of three experimental conditions: the reuse or building (with high or low experimentation) of a DES model of a case study A&E department in England. This is followed by a description of the research hypotheses, the single (attitude) and double-loop (attitude and transfer) dependent variables and the materials used for measurement. The chapter ends with a detailed procedure for conducting the experiment.

A descriptive account of single-loop learning in the three experimental conditions, described in chapter four, is presented in *chapter five*. The aim of this chapter is to familiarise readers with the results by condition without any formal testing of hypotheses. Conclusions of the analysis serve as a basis for a detailed comparison between conditions.

Chapter six provides a formal comparison of the single-loop learning results in respect to the research hypotheses. For clarity these are presented as both graphical and inference tests. The chapter also details the issues in undertaking the analysis and procedures used to help e.g. bootstrapping. The results are followed by a discussion focussing on the possible learning mechanisms found in the experiment in *chapter seven*.

A descriptive account of double-loop learning in the three experimental conditions is presented in *chapter eight*. The aim of this chapter is to familiarise readers with the transfer of learning results and correlations with single-loop variables by condition without any formal testing of hypotheses. Conclusions of the analysis serve as a basis for a detailed comparison between conditions.

Chapter nine provides a formal comparison of the double-loop learning results in respect to the research hypotheses. For clarity these are presented as both graphical and inference tests. The results from this chapter are discussed with respect to the single-loop results (chapters five and six) in *chapter ten*.

The thesis is concluded with *chapter eleven*. In addition to a summary of the main findings this chapter provides a discussion of the contribution of the research to the simulation literature on the value of modelling and model reuse as well as the limitations. Further work is discussed under three headings: work to address limitations, work to test the hypotheses generated by the research and work to refine the experiment. The thesis ends with a brief discussion on how the simulation community view model building and use and the potential for generic models to make an impact in OR interventions via increased experimentation.

Chapter 2

Simulation and Learning

2.1 Introduction

This chapter provides an overview of the simulation literature, DES or otherwise, that is relevant for understanding learning that can occur in a modelling intervention. As the aims of the research are to explore decision maker learning given involvement in model building and model reuse, the chapter tackles these areas separately.

The chapter starts with a review of the conceptual and empirical literature relevant to model building. This argues that involvement of decision makers in model building aids transparency of both the computer model *and* their own mental models of the problem. The learning outcomes frequently discussed are attitude change towards action and development of a deep transferable knowledge, although there is little empirical data available to support these claims.

This is followed by a review of DES and SD literature that explores learning from a use perspective. In fact, much of the simulation learning research falls into this category. Furthermore, although many authors are positive about the potential for learning from gaming and model use, results have been mixed.

In order to learn from a simulation model it seems necessary that decision makers and users feel the model is credible. Thus the third section of this chapter focuses on simulation study credibility and studies exploring it. Although much has been written on methods to build credibility in a simulation study little has been done to measure and test it. Thus a number of other fields that are interested in credibility are briefly reviewed. The chapter concludes with a summary of the main points covered in the review.

2.2 Model Building and Learning

2.2.1 Introduction

The OR and simulation literature on the benefits of model building can be divided into three categories: conceptual, empirical studies with a user focus and empirical studies with a modeller focus.

Firstly, there is the discussion or conceptual category (Churchman and Schainblatt, 1965; Radnor et al., 1970; Urban, 1974; Ward, 1989; Senge, 1990; Lane, 1994; Alessi, 2000; Dhir, 2001; Robinson, 2004). This literature stresses the importance of involvement of decision makers in model building and its benefits for learning (both by the modeller and the decision maker) (e.g. Churchman and Schainblatt, 1965; Alessi, 2000; Dhir, 2001; Robinson, 2004), as well as improving the chances of implementation (e.g. Churchman and Schainblatt, 1965; Radnor et al., 1970; Urban, 1974). Underpinning both these areas seems to be the idea that involvement improves model transparency (Ward, 1989) as well as transparency of decision makers' own mental model of the problem (Lane, 1994). In turn this improved transparency aids attitude change towards action and creation of deep transferable learning.

The second category of literature contains empirical studies of decision maker learning from involvement in simulation model building. These are both field based

(Thomke, 1998; Rouwette et al., 2010) and experimental laboratory based (Shields, 2001). There are two important observations about these studies. Firstly, they are not DES based; they are either quantitative/qualitative System Dynamics or Finite Element simulation. Secondly, the two field work studies, although data rich, do not have the specific objective of identifying learning during different phases of a modelling intervention (i.e. model building versus model use). Hence it is difficult to pull out the relative benefits of model building from the studies.

The third and final category of literature contains empirical studies with a modeller focus. These will be mentioned briefly, but are not the focus of the current study and hence will not be taken further. The most widely recognised study into model building within OR literature was conducted by Willemain (1995). This examined the process that expert modellers used in model building. Findings indicated a highly iterative process. More recent studies have shown that expert modellers may follow a more linear process when building DES models, at least relative to expert System Dynamic modellers (Tako, 2009) and that expert modellers may spend more time on conceptual modelling and validation than novices Wang and Brooks (2011). The remainder of this section reviews the conceptual and empirical literature in greater detail.

2.2.2 Conceptual Literature

The argument advocating the importance of involving decision makers in model building can be traced back to the 1960's. Churchman and Schainblatt (1965) discuss the 'problem of implementation' and advocate a position of mutual understanding between OR modellers and OR decision makers. That is, the modeller needs to do more than just design solutions and communicate these effectively or simply attempt to persuade the decision maker. A modeller must attempt to understand the decision maker's problem and the decision maker needs to understand the modeller's model

building process and eventual model.

This message has been repeated in the OR and simulation literature over the years with the emphasis on involvement of decision makers in the model building process. During these discussions the topic of the learning is typically touched upon. Robinson (2004) provides a nice summary of the argument:

> 'The development and use of a simulation model forces people to think through issues that otherwise may not have been considered. The modeller seeks information, asks for data and questions assumptions, all of which lead to an improved knowledge and understanding of the system that is being simulated.' (Robinson, 2004, :10)

This is described as implementation as learning; the process of developing and experimenting with the simulation model has led to a shift in beliefs, attitudes and behaviour of simulation clients (Robinson, 2004) as well as changes for the system process or policy under study. The mechanism for learning during the model building stage of an intervention seems to be linked to two key factors of the decision makers. Firstly, the decision makers take the role of the major knowledge repository for the modeller (Pidd, 1988). Secondly, much of the desired information, such as management approaches and system workings, are tacit knowledge (Forrester, 1994; Ford and Sterman, 1998) used by the decision maker. To build the model, the modeller must access this tacit knowledge through the procedures Robinson lists above. Thus the model building process forces 'people to think through issues that otherwise may not have been considered' Robinson (2004).

The learning mechanism outlined above may well arise from what is commonly referred to in the OR literature as model transparency. Ward (1989) defines a model as transparent if it is well understood by the decision maker. Ward argues a decision maker is less likely to regard a (quantitative) model as opaque (i.e. not understood) if they are involved in the model building process. Furthermore he lists several

other factors relating to transparency that would seem to stem from or are aided by involvement in the model building process:

- The decision maker receives clear explanations of the model workings;
- The decision maker is addressing what they regard as an important problem and sets aside time to examine the analysis;
- The decision maker is familiar with management science methods
- The decision maker is concerned with a continuing or recurring decision problem and has helped evolve a model of this over time.

Thus the information seeking and questioning of decision makers by modellers during model building leads to improved model transparency. Model transparency equates to improved understanding about the model and its relation to the real world problem by decision makers.

The discussion of transparency can be extended to include transparency of the mental models of decision makers. Lane (1994) provides an overview of a (System Dynamics) simulation process called modelling as learning that has a strong underlying ethos of involving decision makers in model building to create transparency. Lane's process views a model as a transitional object: 'a model that allows users to feed in their assumptions and have played back to them the consequences of these assumptions' (Lane, 1994). Lane argues that that formulating the decision makers assumptions into a SD model increases the transparency of both the computer model and the decision makers' own mental models.

In addition to the conceptual discussion of the approach and its benefits, Lane provides an overview of several studies where the process was applied. For example, Lane discusses one case where two groups of stakeholders within a manufacturing organisation disagreed about the effect of a maintenance shutdown policy on production capacity. Involving the stakeholders in the model building process allowed the 'disentangling of their high level hypotheses of behaviour to find the basic assump-

tions' (Lane, 1994) that the two groups were working from. The learning generated from the model building was the reflection on these assumptions and a change in attitude by the groups about the proposed maintenance policy.

Alessi (2000) discusses the relative educational benefit of building versus using simulation models. He points out that the 'deeper' knowledge created by building a model is largely an untested assumption, albeit one that he accepts. Deeper knowledge is defined as 'knowledge that can be used, in contrast to inert knowledge that is difficult or impossible to use outside of the immediate situation' (Alessi, 2000). This transferability of learning would again seem to be linked to the transparency of a mental model. Alessi (2000) explains the reasoning behind this argument:

> 'The activities of model building require a learner to use and manipulate the knowledge not just observe or repeat it. This is likely to improve not only the recall of the relevant knowledge, but its application and transfer'

Some care must be taken about Alessi's discussion as he is referring to individuals who directly build the computer model. In the context of OR and DES this will typically be a modeller with the relevant expertise in facilitation, software, statistics and methodology. Learning for a modeller may be quite different than that for the decision maker. In a DES study, for example, a simulation modeller may have to develop novel modelling approaches to handle issues with missing data or to enable the model to be built quickly using available data; much of which might be transferable to future projects with the same or similar companies. It is less likely that a decision maker will undergo this type of learning. However, it is plausible to assume that if the real world problem is related to queuing, as in a typical DES project, then the decision maker may learn some general queuing theory results. For example, a decision maker may learn about the trade-off between resource utilisation and service level. This type of learning may also be transferable.

In summary, and for the sake of understanding, it is useful to separate the learning discussed into two categories. The first type of learning refers to attitude change towards, for example, policy actions, intentions and beliefs about system behaviour. Clearly, transparency of computer and mental models is important to form appropriate attitudes. However, a well understood mental model also implies a level of transferability. That is, the decision maker should be able to apply what they have learnt elsewhere. Of course, this new application context must be analogous or, more likely, contain analogous parts to the first context for the decision maker to recognise the opportunity to transfer what they have learnt from the modelling intervention. The OR and simulation papers discussed take the position that involvement of decision makers in model building improves the likelihood of this deeper learning, i.e. attitude change and transferability of learning, through increased transparency of the computer and mental models.

So far the OR and simulation literature discussing the benefits of involving clients in model building for learning has been discussed. The next section introduces the literature that has taken an empirical approach to explore learning from modelling.

2.2.3 Empirical Studies

The second category of literature contains empirical studies of decision maker learning from involvement in simulation model building. Although the studies discussed are useful in understanding learning from the involvement of building simulation models; these in general do not directly compare quantitative simulation model building to model (re)use; they are either qualitative model building (Shields, 2001, 2002) or based on a full modelling interventions including substantial experimentation (Thomke, 1998; Rouwette et al., 2010). Moreover, none of the studies to be discussed is focussed on discrete-event simulation. This section examines the support these studies provide to the conceptual discussions of learning in the previous

section.

Before reviewing this literature in detail it is worth pointing out that a number of experimental SD studies have compared the effect of the level of model transparency on learning without including participants in the building of the model. Rouwette et al. (2004) conduct a literature survey of the factors influencing individual decision maker rationality when using SD simulation models. They find four studies that support the conceptual discussion regarding transparency aiding learning. These experiments provide transparency via introduction to the conceptual model (Young et al., 1992), teaching of SD and the model (Größler, 1998; Größler et al., 2000) or providing details of the conceptual model as the participants struggle with the model (O'Neill, 1992). These studies indirectly support the belief that model building aids learning through increased transparency.

Turning to the literature that explores learning from involvement in the modelling process, the most recent example is an extensive study into attitude change conducted by Rouwette et al. (2010). They develop and test a conceptual model for attitude change in group model building (GMB) interventions based on the Theory of Planned Behaviour (Ajzen, 1991) and theories of persuasion (Petty and Cacioppo, 1986; Chaiken et al., 1989). These theories are discussed in detail in Section 3.3.

Rouwette et al. use data from seven SD field studies they conduct for different clients; for example, a housing association, an oil company and the ministry of transport. The study uses a pre-test post-test (i.e. before and after the intervention) design making use of content analysis (for project reports), questionnaires and post-test interviews. In total 14 variables are measured across three areas: context, mechanism and outcome. A full list of these along with an explanation can be found in Rouwette et al. (2010); a more detailed account can be found in Rouwette (2003). Here only an overview of the most relevant findings is given.

The mechanism variables are based on long standing social psychology theories

of persuasion: the elaboration likelihood model (Petty and Cacioppo, 1986) and the heuristic-systematic model (Chaiken et al., 1989). Rouwette et al. measure the variables theorised to be the most important for systematic processing and evaluation of information: argument quality, assessed by questionnaire and interview, and ability to process information. These are assessed by a questionnaire taken from Vennix et al. (1993) and Rouwette et al. (1998). Outcomes contained seven variables based on another social psychology theory widely used to predict and explain individuals intentions and behaviour called the Theory of Planned Behaviour (Ajzen, 1991). This is reviewed in detail in Section 3.3. For now it enough to say that Rouwette et al measure a change in participant's attitudes towards options to improve system performance. These attitudes are related to participant's views about the other members of the groups and the perceived control they have over the implementation of the options they are investigating.

Rouwette et al. (2010) have three findings that are relevant to learning. Firstly, they find that ability to process information has only a weak relation to pos-test attitude; and no relationship with either the participant's views of other members of the group or the perceived control they have over implementing the options. Secondly, the intervention appears to have no effect on the perceived control participants have over implementing the options. Rouwette et al. remark that this is surprising as most SD practioners would expect simulation to identify control variables with a large impact on the problem. Lastly, they find that the participants often could not identify learning outcomes in the study - even if their attitude measures showed substantial difference from pre to post-test. This agrees with the general view of social psychologists that individuals struggle to understand their own learning (Nisbett and Wilson, 1977). Rouwette et al. recommend that participants in a group model building study are asked to list three options to improve performance before and after the study. The expected differences in these should help reflection on learning.

In a different simulation domain - finite element simulation - Thomke (1998) explores learning during the research and development (R&D) of vehicle crash worthiness at BMW. To illustrate learning over the course of the R&D project, Thomke employs content analysis and extensive interviewing to construct time series illustrating engineers' perceptions of the impact, both magnitude and direction, of different variables on crashworthiness. The three examples included in the paper illustrate the types of learning that could occur. For some variables, attitudes could remain the same for the majority of the project and then experience a sudden jump in importance. Others might fluctuate throughout the project.

Given the R&D setting and finite element approach in Thomke's case study, the model building approach may be similar to a DES study where the system being modelled does not already exist. Building can largely be thought of as design: engineers brainstorm a design for a vehicle based on their assumptions about how the system will behave, outsource the construction of the model (called meshing) and analyse results from the new model. Experimentation also clearly plays a role in the learning process. Thomke points out that numerous learning points for the engineers were based around the interaction of variables in the model that gave rise to the crash dynamics observed.

In his closing comments Thomke reiterates the points outlined in the conceptual discussion of the benefit of involving decision makers in model building.

> 'Thus, I propose that the discipline required in developing computer models will lead to advantages that go beyond making simulation models available to users. It may also take a firm's R&D knowledge from tacit to explicit and, as a result, making it easily transferable within and between firm boundaries'. (Thomke, 1998).

Although the Rouwette et al. (2010) and Thomke (1998) studies are data rich their objectives are not specifically focussed on comparing model building to model

use. Thus it is difficult to pull out exactly what is learnt in building and what is learnt in experimentation. For example, in Thomke's study it might be that the experimentation with variables in a crash model was the most beneficial for learning.

The one study found that directly compares model building and model use is concerned with building qualitative models versus simulation use. Shields (2001) performs an experimental study comparing the building of qualitative SD models (i.e. casual loop diagrams) from a case study to the use of a SD simulation model (generally referred to as a management flight simulator in the SD literature). Shields manipulates these processes by either providing or not providing scripted facilitation. Facilitators ask participants to describe their assumptions about interactions in the model, predict what will happen when a chosen strategy is implemented and explain results and outcomes following feedback on performance.

The study measured pre to post-test learning in five ways: an open ended question asking participants to describe the system; two rating tasks where participants rate the impact of variables on performance; and a diagramming task where participants are provided with pre-labelled variables within the system. Marks were given for the number of connections (complexity) and for inclusion of the direction of feedback (SD).

Findings showed that the qualitative model building groups where a facilitator is not provided demonstrated increased understanding of the complexity and system dynamics of the problem. However, it was the facilitated simulation group that demonstrated increased performance.

2.2.4 Conclusions on Model Building and Learning

The argument for the learning benefits of involving decision makers in model building dates back to the 1960's. Much of the discussion and reference to this learning is compelling and plausibly suggests that involvement in model building can influence

decision maker attitudes and facilitate 'deeper' transferable learning. The limited empirical studies that investigate the area of learning within simulation interventions are also supportive of this. However, none of this literature explicitly compares involving decision makers in the building of a quantitative simulation model to involvement in model use. Additionally, none of the empirical literature explores learning from model building in DES studies.

2.3 Model Use and Learning

The previous section reviewed the general OR and simulation literature regarding learning from the process of building a model. One problem with the empirical studies in that literature is that it can be difficult to draw out decision maker learning that occurs primarily from involvement in model building. This section reviews the literature addressing model use and learning. Perhaps unsurprisingly, it is easier to identify empirical studies primarily focussed on model use. This would seem to be due to relative simplicity in the procedure for laboratory based studies focussed on use compared to building.

This section, firstly, reviews the DES literature on model use. This ranges from conceptual discussions of the benefits (or problems) associated with visual interaction modelling (Van der Zee and Slomp, 2009; Chwif and Barretto, 2003; Belton and Elder, 1994; Musselman, 1990) to empirical studies of learning and perceptions (Tako and Robinson, 2009; Bell and O'Keefe, 1995). Following this is a review of the SD literature. The SD literature exploring model use is much more closely linked to the history of simulation gaming (Lane, 1995) and appears to be much more mature than the field of DES. In fact, there are numerous well known experimental studies investigating learning from simulation type board games (Sterman, 1989*b*,*a*) and simulation models (Paich and Sterman, 1993; Bakken et al., 1994), as well as review papers summarising important insights in the SD simulation field (Lane, 1995;

Rouwette et al., 2004; Größler, 2004). The section closes with a summary of the main points covered in the review.

2.3.1 DES Research

Belton and Elder (1994) argue, based on their combined experience, that use of Visual Interactive Simulation (VIS) for simulation experimentation is beneficial for discovery, clarification, change and creation of a client's views and ideas about system management. Indeed these benefits would seem to suggest alternative uses for DES such as gaming and training (Van der Zee and Slomp, 2009; Chwif and Barretto, 2003). The reasoning behind this is similar to that expressed by Lane (1994). Decision makers are able to explore the consequences of their assumptions and decisions while receiving fast and understandable feedback from the model. This experience helps makes the tacit views they hold explicit and possibly reveals inconsistencies or incompatibilities between them (Belton and Elder, 1994).

Bell and O'Keefe (1995) reiterate the argument for visual interactive simulation and also provide an insight into an opposing view: namely the overemphasis of the result from a single run and subjective based analysis (Musselman, 1990, quoted in Bell and O'Keefe, 1995). Bell and O'Keefe point out that most proponents of VIS will use a combination of batch and interactive experimentation, but also that the majority of what is known about using VIS is anecdotal.

Hence to explore the proposed benefits of VIS, Bell and O'Keefe (1995) conduct an experimental study investigating the results users of interactive experimentation can achieve relative to a statistical analysis. Participants, namely MBA students, are provided with a model based around a queuing and resource-allocation problem at a mine. The model is simple and allows for a 'correct solution' to be found. Participants read the case study the night before and provide an initial solution before using the model. After interactive use of the model the participants provide

a revised answer. Statistics on usage are collected throughout.

Findings show that the participant's solutions were a substantial improvement on their initial efforts. However, only 39% provided the correct solution. Analysis of the usage statistics identified that the participants who made most use of the animated display found the correct results most quickly. Furthermore, participants that collected multiple data from a scenario (i.e. viewing the results in detail and drilling into data) before proceeding to the next, performed best.

The most recent study was conducted by Tako and Robinson (2009): an empirical comparison of the use of SD and DES models. Participants in the experiment, executive MBAs, take the role of a government consulting service and use either a SD or DES model of the UK prison population to explore and draw conclusions on how the process could be improved. At the end of the experiment participants fill in a questionnaire to elicit their perceptions of model understanding, complexity, credibility and interpretation of results. Interestingly, the results showed little difference between the measures. In fact, only slight differences were found in how representative of the case study participants found the models (favouring the SD model) and that the participants found the results of the DES model more difficult to interpret.

2.3.2 SD Research

The literature exploring and testing learning from using SD models is arguably more substantial and mature than the DES literature just discussed. The focus is largely on the use of management flight simulators or SD business games; essentially using simulation models as games to help the user, who may or may not be a manager, learn about a specific concept. Several review papers exist to summarise the key insights that have been discovered (Lane, 1995; Rouwette et al., 2004; Größler, 2004). Lane (1995) provides a historical perspective on business games in general dating

back to the 1960's. Business games appear to have had a mixed history with some early successes, but also failures including an infamous paper by Neuhauser (1976) declaring that it is the model design process, not use, where the majority of learning is found. Rouwette et al. (2004) review only the empirical studies of model use found within the SD literature. They provide an overview of the characteristics of models, simulators (e.g. a simulation business game) and players that influence decision making over time. A key finding, already discussed in the section on model building, was that model transparency has a positive relation to performance. Größler (2004) reviews the methodological issues in using simulation as an educational tool and lists 15 main issues. These include overcomplicated models, accounting for the different learning styles of participants and the lack of risk for participants' decision making.

Turning to specific relevant case studies, Bakken et al. (1994) conduct an empirical study concerned with transfer of learning using executive MBA and undergraduate students. In the experiment MBAs, working in groups of two, and undergraduate students, working individually, attempt to manage an SD simulation model of either a real estate or oil tanker system. The underlying feedback structure of these systems is identical. After a certain amount of time the participants use the other simulation model. To infer transfer of learning Bakken et al. look at the performance of participants using the second simulation model.

The results show that the students outperform the executive MBAs in the transfer model. This is particularly interesting as a number of the MBAs have substantial experience in one of the domains used in the experiment. To explain this result Bakken et al. provide data that shows that undergraduate students are much more exploratory than the executive MBAs in the first simulation model. In fact, the undergraduate students make approximately double the amount of decisions compared to the real world managers. This results in an increased bankruptcy rate in the first game, but allows the students to test a greater number of their assump-

tions. In contrast an underlying assumption of the executive MBAs with real world experience appeared to be 'I know a lot about this market; so I just do in the game what I would do in real life' (Bakken et al., 1994).

Learning problems are not just limited to managers using simulation games; students also have difficulties. Paich and Sterman (1993) conduct an empirical study where postgraduate students, including MBAs, use a SD simulation model to manage a product from launch through to maturity. The objective of the game is to minimise the boom and bust dynamics of the market using pricing and capacity decisions. Participants use the model over five trials and make a total of 200 pricing and capacity decisions. After each decision stage the simulation model provides results to the participant on current performance as well as the history of decisions and performance in the game. Furthermore, participants can take as long as they need to make decisions.

The results of the study found that on average the participants performed better over the five trials. However, the participants performed poorly relative to a naive benchmark: mean performance (profit) was 60% of the benchmark with only 17% of participants outperforming the benchmark in the final trial. This performance is explained by the student learning the average demand for replacement of the product in the market and matching their capacity to this area. However, they completely ignored other aspects of the market, for example, the growth rate of demand for new products. In fact their actions worsened the extent of the boom and bust dynamics.

The results of Paich and Sterman (1993) support the misperception of feedback hypothesis put forward by Sterman (1989*b*) that 'the stronger the feedback process in the environment the worse people do relative to potential'. Paich and Sterman (1993) conclude that in order for participants to learn when dynamic complexity is high the students needed to become modellers not just players in a simulation game.

2.3.3 Conclusions on Model Use and Learning

In summary, the literature on learning from the use of simulation models is mixed. It would seem that there is some benefit from interactively using a simulation model to improve on decision makers initial solutions. However, the search and results reviewing process that a decision maker uses can radically affect this outcome. Indeed, some studies illustrate that learning can be quite poor relative to potential.

One possible problem with drawing conclusions from this literature review is the lack of empirical studies using DES models and the prevalence of SD models. Some confidence can be drawn from the Tako and Robinson (2009) study. This demonstrated that users' perceptions of using a specific SD and DES model is largely the same. One issue may be that participants in the Tako and Robinson study indicated that the DES results were more difficult to interpret. Hence learning difficulties may be worse in a complex DES study.

Further to what has been discussed, an important aspect of learning from modelling and model use is the credibility of the study, model and results with the individuals or group who will actually make the decisions. This is the topic of the final section of this chapter.

2.4 Models, Learning and Credibility

Chapter 1 introduced the concept of simulation model reuse and the issues reuse may cause with model credibility. Clearly the confidence that a decision maker has in a simulation study is important to understand learning since without confidence the decision maker is unlikely to use the model as an aid to decision making. This section provides a review of how the concept of credibility has been defined, studied and operationalised in several areas: DES, Human Computer Interaction, Psychology, and Media and Marketing. Using this review some extensions are suggested to the current treatment of credibility in the simulation literature and conclusions and

assumptions are listed for the definitions of credibility and its measurement in this research.

2.4.1 Discrete-Event Simulation and Credibility

When building a simulation model a modeller is concerned with the concept of validity: an assessment of if the model is sufficiently accurate for purpose (Robinson 2002, 2004; Pidd, 2004; Law and Kelton, 2000; Law and McComas, 2001; Banks et al., 1996; Robinson et al, 2004; Tako and Robinson, 2008; Carson, 1986). Model credibility is similar, but taken from the simulation clients' perspective. In the DES literature it is defined as the confidence that a client has in using a simulation model and its results for decision making (Tako and Robinson, 2009; Robinson, 2004, 2002). That is, although a modeller may consider a model sufficiently accurate for purpose, i.e. valid, a client may not find it credible. For example, there may be a detail the client believes is important to the model that is not currently included. Although the model is technically valid from the modeller's perspective it lacks credibility on the client's part, as it is felt to not be fully representative or give realistic results, thus affecting how it is used in the decision making process.

Within the simulation literature there are several studies where credibility has been measured (Tako and Robinson, 2009; Robinson, 2002, 1998) and many that discuss the methods that can be employed to improve credibility (e.g. Balci and Nance, 1985; Balci, 1994; Law, 2007; Robinson, 2008). Tako and Robinson (2009) study the differences between users' perceptions of the use of System Dynamics and Discrete-Event Simulation models of the same problem. Credibility is operationalised as a multi-dimensional construct using five point Likert scales for user attitudes towards the model's representativeness of the case study, realism of model outputs and confidence in using the model for decision making. Robinson (1998, 2002) tackles credibility from a service quality perspective. The SimQual instru-

ment, developed in the research, measures all aspects of service quality including confidence in the model and credibility of the simulation provider. Robinson applies the SimQual instrument within three case studies; credibility ratings are neutral, negative and positive respectively.

2.4.2 Human Computer Interaction and Credibility

The human computer interaction (HCI) credibility literature concentrates on user perceptions of information technology (Tseng and Fogg, 1999; Waern and Ramberg, 1996; Andrews and Gutkin, 1991). This is similar to the DES literature as it considers credibility issues to arise when computers report measurements or when indeed they run simulations, albeit in the most general sense. Tseng and Fogg (1999) review HCI credibility research and propose that a good synonym for credibility is 'believability'. Furthermore they propose that credibility can be of different types: presumed, reputed, surface and experienced.

Presumed credibility refers to the general assumptions in the perceiver's mind (Tseng and Fogg, 1999). For example, a manufacturing manager new to DES may be considerably impressed by the prospect of building a visual computer model of their production line and generally assume that computers are 'clever'; hence it automatically has high credibility with him or her. Alternatively the manager may have been involved in unsuccessful modelling projects or modelling projects with undesirable outcomes in the past. Hence he or she may assume that similar outcomes will occur in future projects.

Reputed credibility refers to the how much the perceiver believes something based on what third parties have reported (Tseng and Fogg, 1999). For example, key decision makers may not be directly involved in a simulation study and listen to statements about credibility from other stakeholders involved or the simulation modeller(s).

Surface credibility refers to how much a perceiver believes something based on simple inspection (Tseng and Fogg, 1999). For example, in the context of simulation research a VIS package could easily be made to look like a process flow diagram and its output may even look correct relative to actual data. Based on this simple inspection the model and its results are believable. However, this is not an indication of whether the dynamic behaviour of the model is correct. This is credibility based on a simple Turing test.

Lastly, *experienced credibility* refers to how believable something is based on first-hand experience. In the context of a simulation study, for example, this could be based on involvement in the construction, interaction and testing of the dynamic behaviour of the model over the course of the study. This last category seems to be akin to the confidence building process recommended by many DES authors (e.g. Balci, 1994; Robinson et al., 2004).

Measurement has been similar to Tako and Robinson (2009). For example, in their study of participant responses to computer and human generated reports Andrews and Gutkin (1991) measure credibility, diagnostic interpretations and confidence in their judgement based on the report using likert scales. The latter variables are strongly correlated to results and confidence in using a model for decision making. However, the meaning of the word *credibility* will likely be different depending on who you ask and is possibly a weakness in the approach.

2.4.3 Media, Marketing and Credibility

Credibility has also been studied extensively in media research. This typically takes the form of comparing the credibility of different types of media and an analysis of the factors influencing credibility (e.g. Cassidy, 2007; Abdulla et al., 2004; Johnson and Kaye, 1998, 2002; Kiousis, 2001; Meyer, 1988; Gaziano and McGrath, 1986). Credibility is defined as a multi-dimensional construct; although it is operationalised

differently depending on the study. Versions of Gaziano and McGrath's (1986) credibility index are used by several other studies (Meyer, 1988, Johnson and Kaye, 1998, 2002; Cassidy, 2007). Four measures are usually used: believability, lack of bias, accuracy and comprehensiveness (or depth of information). These are again operationalised as Likert scales. A credibility index is calculated as the summation of the dimensions. Items such as comprehensiveness may measure the same dimension as representativeness as found in the simulation literature.

2.4.4 Persuasion Theories and Credibility

Before discussing the research in this area, it is worth pointing out that the DES definition of credibility - an attitude of the client - and its distinction from validity is inherently linked to persuasion. Building a decision maker's confidence in a modeller, a model and it results for decision making through verification, validation and discussion is persuasion. Classical persuasion models such as the Elabouration Liklihood Model (Petty and Cacioppo, 1986) and the Heuristic-Systematic Model (Chaiken et al., 1989) suggest that individuals change their attitudes through two routes:

1. Central/Systemic - messages are considered carefully and information contained is tested and validated.

2. Peripheral/Heuristic - simple heuristics are used e.g. I presume that simulation will be correct or I did not like the way the modeller spoke to me.

In a SD study of group model building Rouwette (2003) assumes that attitude change operates primarily through the systematic route. This also assumes that the motivation and ability to process arguments is high. Rouwette found that motivation and ability was indeed high thus the central route was used. However, he suggests that the heuristic route may at times play a significant role in attitude change:

> *"This respondant remarks that she has changed her opinion not so much on the basis of the content of the discussion, but because of its general negative tone."* Rouwette (2003: 236)

The heuristic route and Rouwette's empirical work demonstrates that experienced credibility is derived not just from the model and its results, but also how persuasive the modeller and the discussion around the model is. This is a factor that has not been overlooked in the DES (Robinson, 2002) and general OR (Churchman and Schainblatt, 1965) literature, although this has not been empirically measured.

The persuasion literature has been concerned with perceived *source credibility*; defined as a message source's perceived expertise and trustworthiness (Kelman and Hovland, 1953). Research has looked at the impact of credibility on attitudes (Tormala et al., 2006; Priester and Petty, 1995; Chaiken and Maheswaran, 1994) and how different mechanisms result in different credibility attribution and use of information (c.f. Reinhard and Sporer, 2008). Contemporary research has also made an important contribution to distinguishing between cognitive and metacognitive (i.e. thinking about thinking) processes (see Petty et al., 2002, for a discussion). For example, an individual may be asked to provide an answer to a question on the accuracy of the model. A separate dimension to the problem is the individual's confidence in the validity of their own answer e.g. they answer that they feel the model is accurate, but are only moderately sure that this answer is correct. The latter part of the last statement is the metacognitive process.

Reinhard and Sporer (2008) provide empirical evidence that only high task involvement and high cognitive capacity leads to the systematic route of processing of arguments when judging credibility. That is, the more a capable individual is involved in a 'task' (for example, building a simulation model) the more likely they are to use more of the available information to decide source credibility (for example, a model and its results) than individuals with less involvement. In fact individuals

with less involvement in a task are more likely to use the heuristic route to decide on source credibility. Measurement is operationalised on a nine-point scale[1] and summed into a credibility index (Reinhard and Sporer, 2008) and as an attitude towards a source of information using averaged nine-point Semantic differential scales (Tormala et al., 2006).

2.4.5 Conclusions on models, learning and credibility

It is, perhaps, unsurprising that there is correspondence between the definitions and measurement found in the HCI and DES literature; decision support using DES could be classified as a subset HCI. However, the HCI as well as the media, and persuasion literature can make several theoretical contributions to the understanding of credibility in the DES literature.

1. Specifically it is experienced credibility simulation of the model, its results and its author(s) to which the DES literature refers.

2. Given a valid model, the route a client takes in processing information from the model or modeller can affect resultant credibility.

3. The level of involvement that a client has in developing a model may affect the route of processing when judging credibility. Specifically low involvement can lead to heuristic judgement.

4. If the simulation study is considered as a source of information then measurement of credibility can be operationalized using a scale of believability, lack of bias, accuracy and comprehensiveness.

5. There are two dimensions to measurement of credibility. An assessment of model credibility and the confidence an individual has in that statement.

[1] It is not indicated if a Likert, Semantic differential or any other type of scale is used

The second and third points have potential impacts on learning when reusing models. If a decision maker does not consider that they have high involvement in building the model then they may only use limited information and heuristics to reach neutral or negative judgements of credibility. This is in line with the central hypothesis of this research: model building and reuse studies result in different learning outcomes and processes for clients.

2.5 Summary

This chapter has reviewed the OR and simulation literature related to learning from model building and model use. This literature argues that involvement in model building improves transparency, by which is meant understanding, of computer and mental models. Mental models refer to the images that limit us to familiar ways of thinking and acting (Senge, 1990). This is important as understanding of both the computer and mental models aids attitude change and deep learning in the decision maker. This deeper learning is often referred to as transferable learning. That is, it is learning that can be transferred to future decision making in the same or an analogous system.

Although some empirical studies have been conducted to explore learning from modelling, it can be difficult to pull out the relative contribution of model building and use. One study does explicitly compare the two (Shields, 2001); however, this only considers the benefits of involvement in building qualitative models. Moreover, none of these studies employ a discrete-event simulation model.

In fact, much of the research about learning from simulation has been driven from a use perspective. However, while it is felt by many that the use of simulation models is highly beneficial for learning, empirical results have been mixed. Users could both gain insights, but could also struggle to learn anything. Explanations for these results have ranged from users failing to realise underlying assumptions in

their behaviour to the requirement for users to become model builders in order to learn effectively.

Lastly, it was argued that the credibility of the study is an important mechanism for learning from model building and use. Methods for building credibility are numerous in the DES literature, but comparatively few have explicitly measured it. The topic also comes up frequently in HCI, media and psychology fields. These fields provide useful insights into the mechanisms and measurement techniques for credibility that can be applied in a DES context.

The next chapter uses the results of this literature review to describe a conceptual framework for measuring learning from model building and model use. This brings together the concepts of attitude change and transfer of learning under a general framework of double and single-loop learning.

Chapter 3

Learning Framework

3.1 Introduction

The previous chapter reviewed the simulation and learning literature. It concluded that learning outcomes for decision makers involved in model building and use can be divided into two categories: attitude change and transferable knowledge. This chapter discusses a combined framework for attitude change and transfer of learning using Argyris and Schon's theory-of-action framework. This equates attitude change to the term single-loop learning and achievement of transferable knowledge through double-loop learning.

The chapter begins with an overview of Argyris and Schon's theory-of-action and learning framework (Argyris and Schön, 1996). The results of the Bakken et al. (1994) study are used to aid understanding of the framework. This is followed by sections reviewing relevant attitude and transfer measurement theory from psychology. Firstly, attitudes and their link to behaviour are discussed in the context of the Theory of Planned Behaviour (Ajzen, 1991). This includes a review of the persuasion literature and its link to attitudes about credibility. Secondly, the transfer of learning literature is reviewed. The chapter ends with a summary of the main

points.

3.2 A Theory-of-Action Framework

As an example of a theory-of-action consider Figure 3.1: an example of a learning process that a manager of several regional call centres may go through during a DES study. In his or her day to day management of the call centres, the manager might reason that small groups of call operators can be specialised to increase the speed of call handling and, at the same time, maximise their utilisation. These objectives are an example of the governing variables (Argyris and Schön, 1996) that make up the theory (Argyris and Schön, 1996) or image (Senge, 1990) that limits and controls the managerial decisions made by the individual. Long term decision making will attempt to satisfy these variables. Thus, the manager's long term strategy uses separate regional call centres for handling 'customers'.

Assume now that the manager is then involved in a DES study to reduce customer average wait time on the phone by x number of minutes. The study may show that combining all or even some of the call centres helps meet the modelling objective i.e. queues are reduced. This, of course, also leads to larger, less specialised, call operator teams. The manager may even be surprised to find out that overall the average utilisation of the call operators has increased slightly.

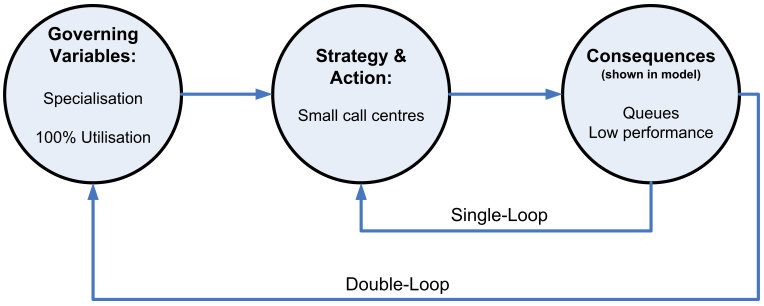


Figure 3.1: An illustration of single and double-loop learning

One outcome from the study may be that the call centre manager has a positive

attitude towards combining the call centres and the larger call operator groups for the business. This may lead to implementation of the study results and an improvement in long term service level. This is defined as single-loop learning: a correction of errors in the management of the specific business problem (Argyris and Schön, 1996). Figure 3.1 illustrates this concept through the feedback from consequences of action, in this case highlighted by the model, to actions and strategy.

Single-loop learning is distinct from the manager's understanding of why the combination option outperforms his or her theory of small efficient call centres. A second level of error correction is necessary to achieve this understanding (Argyris, 1992; Argyris and Schön, 1996). This is defined as double-loop learning: a reflection of the difference between the manager's theory and the performance of the simulation model. The outcome of a double-loop learning process is a change in the managers' long term decision making behaviour.

As an example, consider now that same manager oversees a back office process dealing with customer applications and record checking. This is also a queuing process subject to variation in arrival rate and activity time that is split up by region. If the manager has reflected on the results of the call centre simulation model and the reasoning behind their own management decisions then they may realise that a relationship exists between resource utilisation and performance. Thus they may *consider* the option of combining the back office resources to speed up application processing. If, however, they have only undergone single-loop learning then they would automatically apply the same governing variables as before and not consider combining resources.

If the framework is now used to consider the Bakken et al. (1994) result, discussed in Section 2.3.2, it can be seen that it is the *theory-of-action that decision makers use to learn from simulation that is key* to double-loop learning. Figure 3.2 illustrates the learning systems used by the management and student participants respectively.

The managers used what Argyris and Schön (1996) call a 'win not lose' governing variable. The managers may have felt that they had much more face to save in the experiment than the students i.e. they had experience in the domain and going bankrupt frequently might be perceived as quite embarrassing. Thus, in general, the management participants were not open to testing or reflecting on their theories. They 'knew how the system worked' and did not attempt to refute this knowledge. This resulted in only minor attitude change about what to do in the first simulation model and low transfer of learning to the second model.

On the other hand the students may have the advantage of feeling that they had less to lose; hence they were open to failure and reflected on why they performed poorly. This enhanced their learning in the first simulation model and demonstrated their understanding in the transfer model. Figure 3.2 illustrates this with the feedback from the consequences to problem governing variables.

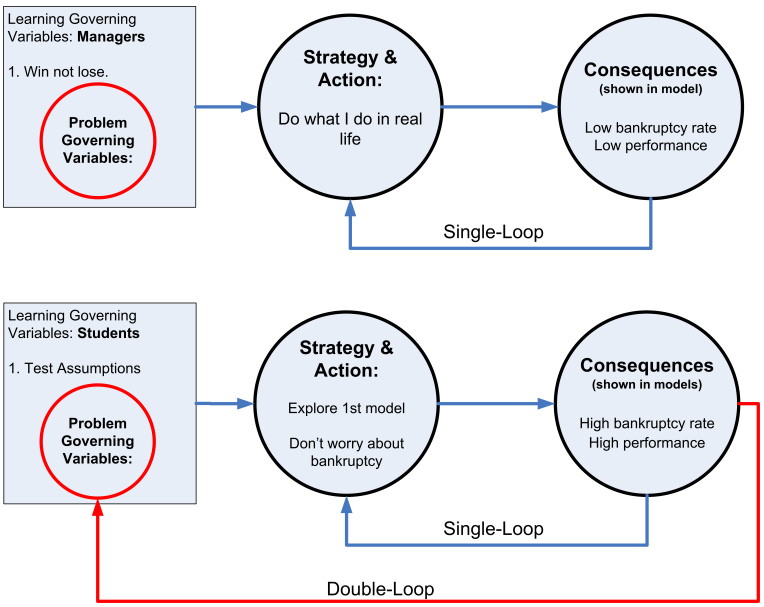


Figure 3.2: Learning systems used by management and students participants in Bakken et al. (1994)

The majority of the research into double and single-loop learning has focused

on methods to encourage double-loop learning (see Argyris, 1992), but little has been done to measure it. The most relevant study, Bakken et al. (1994), measures transfer of learning as an indicator of double-loop learning; however, it does not consider single-loop learning and the potential benefits that can bring for decision making.

The remainder of this chapter reviews the literature relevant for measuring single and double-loop learning - namely attitude change and transfer of learning theory.

3.3 Single-Loop Learning: Attitude Change

3.3.1 Attitude Theory

Attitudes are general evaluations people hold in regard to themselves, other people, objects and issues (Petty and Cacioppo, 1986). As an example, consider a simulation project that is investigating several options for improving the speed at which patients are treated in an accident and emergency department. Assume one of these options is to employ an additional nurse at peak times. The attitude towards the additional nurse would be the degree to which a person has a favourable or unfavourable evaluation or appraisal (Ajzen, 1991) of employing an additional nurse.

Attitudes are formed from the *salient beliefs*, the small number of beliefs that an individual can access at a given moment (Ajzen, 1988), about the object of the attitude. The beliefs about the improvement option, in the example above, may have been constructed from 'direct observation, self-generated by inference processes, or formed directly by accepting information from outside sources such as friends or media' (Ajzen, 1988, :7). For example, a manager within the A&E department may believe that:

- The current nurses could work harder at a higher utilisation and improve performance to the same extent as employing an additional nurse.

The measurement of attitudes follows an expectancy-value model (Ajzen, 1991). Attitudes are the product of the subjective likelihood that a belief is true and an evaluation of outcome desirability. In the example above the manager may feel it is very likely that working the nurses harder will improve performance and also finds it very desirable to have both improved performance and higher value-for-money from resources.

Figure 3.3 illustrates that the expectancy-value model takes account of multiple beliefs. The measure of attitude is calculated as simply the summation of the n beliefs (Ajzen, 1991). The only requirement for this calculation is that the beliefs are a compatible sample. The usual way to achieve this is to sample beliefs for a specific time period. For example, a simulation study may be concentrating on improving performance of an A&E department over a six month period. Thus all measurement instruments (typically questionnaires) would focus on eliciting beliefs relevant to this time period.

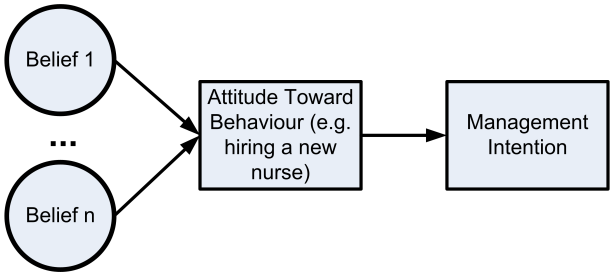


Figure 3.3: Expectancy Value Model

Figure 3.3 also illustrates that attitude is a significant predictor of (management) intentions. Large literature reviews have substantiated this claim (Ajzen, 1991, 2001; Rouwette, 2003). In fact, this construct is typically referred to as an attitude towards behaviour (Ajzen, 1991). This is relevant in a simulation study as the behaviour we are interested in predicting is implementation action (Rouwette et al., 2010; Rouwette, 2003).

As a side note it is worth pointing out that the construct of attitude towards

behaviour is taken from a larger psychological theory of behaviour called the Theory of Planned Behaviour (TPB) (Ajzen, 1991). In the TPB there are two other constructs that predict intention: perceived social pressure to perform or not to perform the behaviour (subjective norm) and perceived ease or difficulty of performing the behaviour (perceived behavioural control). These are both relevant in a real simulation studies involving multiple stakeholders (Rouwette et al., 2010). However, as the current study does not require this level of detail and will not involve multiple decision makers, the discussion of these variables will not be taken further at this point. Full details can be found in Ajzen (1991) and Rouwette (2003) and are discussed in the limitations of the experiment in Section 11.5.5.

3.3.2 Persuasion Theory

Chapter 1 highlighted credibility as an issue within model reuse studies. That is, decision makers (and modellers) may not have sufficient confidence to use a model as an aid to decision making if it has not been built in-house. Chapter 2 reviewed the DES, HCI, media and psychology credibility literature. As part of the psychology literature review the concept persuasion through systematic and heuristic processing of information was discussed. Systematic processing is a comprehensive, analytic orientation in which individuals scrutinise all available information for its relevance to their judgemental task (Chaiken et al., 1989). Heuristic processing lies at the other end of the spectrum, requiring much less effort, and is theory driven (e.g. trust experts or consensus implies correctness) (Chaiken et al., 1989).

The most relevant aspect of persuasion theory for credibility in model building or reuse studies is the Sufficiency Principle taken from the Heuristic-Systematic Model (Chaiken et al., 1989). The principle has two important assumptions. Firstly, it assumes that people are economy minded and prefer less effortful to more effortful modes of information processing in order to attain the validity of an attitude.

Secondly, it assumes that in general people cannot know if their attitudes are completely correct. For example, it is well accepted in DES literature that there is no such things as general validity of a simulation model; instead validation should be viewed as a confidence building exercise (Robinson, 2004).

Given these assumptions, the sufficiency principle introduces the concept of a sufficiency (confidence) threshold. Figure 3.4 illustrates the principle using a hypothetical scale of confidence in a simulation model. For a given simulation project individuals may have different confidence level thresholds that they deem sufficient before the model can be used. In Figure 3.4 there are two individuals with thresholds ST1 and ST2 respectively. Clearly what is sufficient for person one is not sufficient for person two. This difference is important as it affects the type of processing and effort the individuals will give. For example, as person two's threshold is relatively high compared to person one it is more difficult to achieve. Person two is much more likely to need to undertake systematic processing to achieve his or her threshold than person one. Person one will only expend as much effort as necessary to achieve his or her threshold. As his or her threshold is lower it is more likely that heuristic assessment will be enough (e.g the model visually looks correct).

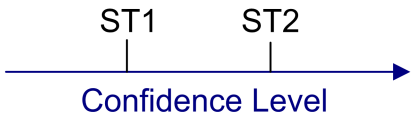


Figure 3.4: Example of Sufficiency Thresholds

To illustrate the link to simulation studies and model reuse a dynamic hypothesis is depicted using a causal loop diagram in Figure 3.5. The direction of the arrow head in causal loop diagrams indicates the direction of causality while the sign of the arrowhead indicates the effect of the causality (Pidd, 2004). The small loops containing positive and negative labels indicate the reinforcing or balancing feedback of the loops respectively. The diagram illustrates a decision makers attitude to

Verification and Validation (V&V) of a model using sufficiency thresholds.

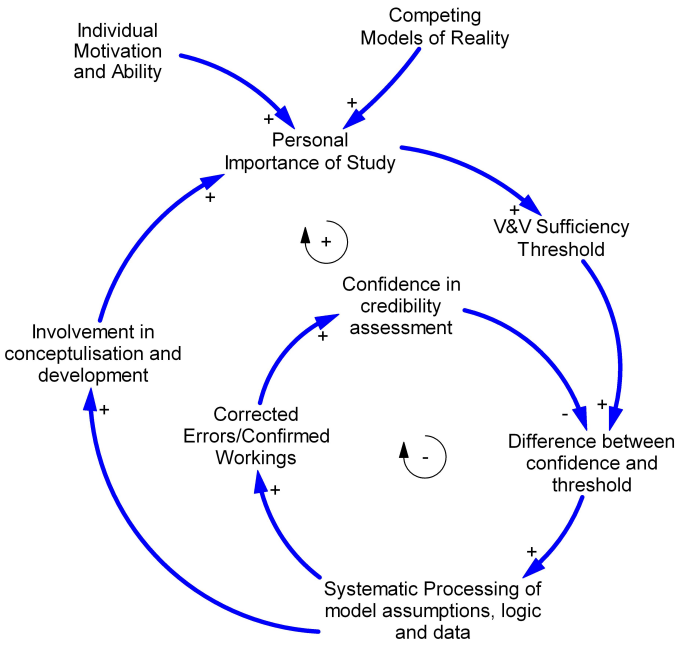


Figure 3.5: Sufficiency Thresholds for Verification and Validation

A key variable affecting sufficiency thresholds is the *personal importance of the simulation study* to the decision maker (Chaiken and Maheswaran, 1994). This will be affected by an individual's natural motivation to engage and possibly any competing models of reality (e.g. I think the system works in this manner) that can also explain system performance. High involvement in the conceptualisation and building of the model increases personal importance of the study above what it would have been. This in turn increases the threshold for where V&V of the model is viewed as sufficient. Importantly a higher V&V sufficiency threshold increases the systematic processing of information (Chaiken et al., 1989): model assumptions, logic and data. This reinforces involvement in the study as re-conceptualisation may be required. Balancing this loop is confidence in the validity of credibility assessments. The higher the systematic processing of information the more 'errors' or inconsitencies with the real system in the model are exposed, fixed or confirmed. The more clients

participate in this way the higher their 'thought confidence' becomes (Petty et al., 2002). This continues until the V&V sufficiency threshold is reached. Thus it is predicted that clients with high involvement in building will apply more scrutiny to the model and have high confidence in the validity of their credibility assessments. It is also predicted that participants with low involvement in development will use more heuristics in forming their attitude of credibility (e.g. the model has been built by an expert) as their sufficiency threshold for V&V is lower. In other words, although both model building and model reuse clients may rate the model as credible, model reuse clients may have low confidence in the validity of their assessment of credibility - a sort of heuristic or surface credibility (Tseng and Fogg, 1999). Model building clients may hold a more systematic high confidence attitude to the credibility of the model and its results as they have scrutinised the model to a greater extent.

This depiction agrees within one conclusion from Rouwette (2003): 'if the issue [being modelled] is not sufficiently important [to the decision maker], learning effects may be absent and thus implementation of modelling conclusions hampered'. That is, systematic processing often confers greater attitude persistence than heuristic processing (Chaiken et al., 1989). Personal importance aids V&V and the chances of implementation.

3.4 Double-Loop learning: Transfer

The final aspect of learning theory to examine deals with the transfer of learning. This is relevant to double-loop learning as it used as a method to elicit an individuals reasoning on a subject. A typical social psychology experiment of transfer consists of analogous training and transfer problems. A participant solves the first problem, is given feedback and then attempts to solve the transfer problem (Bassok, 2003). For example, an often cited study, Gick and Holyoak (1980), used the Tower of Hanoi problem in the training task. Participants play the role of a general that

must use the army to capture the Tower of Hanoi from a rival general. There are four routes to the tower. If the participant attempts to capture the tower using any single route the army is defeated. However, if the participant divides their army and uses all four routes at once they can overpower the enemy forces and capture the tower. In the transfer problem the participant must decide how to kill a tumour in a patient using x-rays. The problem is that the required dosage of x-ray will damage the tissue it passes through on the way to the tumour. The transfer concept is divide and conquer i.e. apply lower intensity x-rays from different sides of the body simultaneously. Note that this is the same basic approach as used in the Bakken et al. (1994) study of transfer between SD models.

Figure 3.6 illustrates two important factors - structural and perceived surface similarity - and their impact on the likelihood of transfer success. *Structural similarity* refers to the underlying mechanics of the problem being studied. For example, Bakken et al. (1994) study the transfer performance between two system dynamics models with the same underlying feedback structure, but different surface components: a housing market and an oil tanker market. Similarly, as Figure 3.6 illustrates, in the context of DES both an A&E department and a call centre have an element of structural similarity. They can be considered as queuing systems subject to process variation with much of the variability being driven by arrivals.

Perceived surface similarity is an important cue for initiating a transfer attempt. If an individual does not see any similarity to the training task in the transfer task then they will not attempt to transfer the knowledge. If, however, they do see the similarity then transfer will be attempted (Bassok, 2003). Transfer success is most likely when an individual perceives that the new problem is highly similar to one they have tackled before and, in addition, the original and new problems are structurally similar. Figure 3.6 illustrates this by the positioning of the NHS and call centre icons. If it is assumed that an A&E context is the 'training task', people may find many

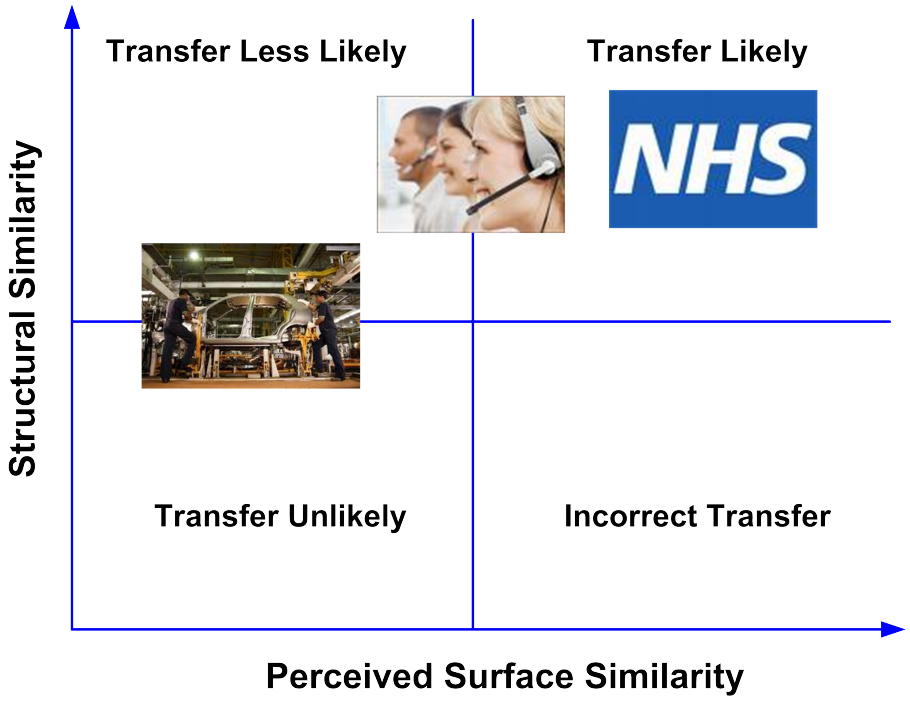


Figure 3.6: Transfer of Learning

aspects similar between the two situations e.g. both concern people being served and serving and both have highly unpredictable inter-arrival times. However, as shown by Figure 3.6, transfer is less likely from an A&E context to a manufacturing. This is, firstly, because perceived surface similarity may be low, due the difference of people in a process versus widgets on a production line. Secondly, structurally the problems are slightly different: in the manufacturing domain variation is driven from cycle times and machine breakdowns as opposed to arrivals (although one might consider machine breakdowns as arrivals to the system). Lastly, it is worth noting that incorrect transfer can occur. This is when an individual perceives a problem to be highly similar to one they have tackled before, but structurally they are very different.

The difference in perceived surface similarity can be thought of as close and far transfer. Empirical studies support the view that far transfer is generally more

difficult (Barnett and Ceci, 2002). Although this appears to be a fairly straightforward point to grasp it is actually quite difficult to define 'far' as perceived similarity is highly dependent on contextual and individual factors (Barnett and Ceci, 2002). For example, 'far' may refer to different contexts (such as manufacturing and healthcare), the time lag between transfer attempts or the location where transfer is attempted (e.g. a classroom versus a work environment). Nonetheless a lot of research has been conducted into the area (Gick and Holyoak, 1987). One method that seems to have a positive effect on transfer is to provide multiple examples of the problem in the training task. Transfer performance appears to improve if both close and far examples are given during training (Gick and Holyoak, 1983). The explanation appears to be that the differences in surface similarity in the examples improve the individual's ability to abstract the structure of the problem - giving them a deeper understanding that is transferable.

A final aspect to consider is spontaneous versus informed transfer (Bassok, 2003; Gick and Holyoak, 1987). Consider a manager who is involved in a simulation study of a manufacturing production line. In the study the manager learns that he or she cannot run production line machines at greater than 90% utilisation and still achieve lead time targets. Furthermore the manager finds that there are huge differences in lead times if machines are run at 85% and 95% utilisation. Sometime later the manager moves organisation and is put in charge of a new production line. If the manager considers making similar decisions here, i.e. considers lower average utilisation to improve lead times, then the transfer was spontaneous. The manager recognised the similarity between the problems, accessed the relevant knowledge and transferred it successfully.

The opposite of spontaneous transfer is where the manager, or whomever, is told to transfer what they have learnt in 'training'. Informed transfer is an artificial procedure only available in a laboratory setting. Its use is in differentiating between

problems in accessing relevant knowledge and problems in learning. Many transfer of learning studies employ a design where one group is given a transfer hint (i.e. now use what you have learnt to solve the following problem) and another group is simply presented with the new problem (Bassok, 2003). If the hint group outperforms the no-hint group then it can be concluded that learning is present, but there is an access problem i.e. the no-hint group does not recognise the similarity between the training and transfer problems. If neither group perform well then there would appear to be a learning problem.

In summary, to achieve transfer an individual must not only possess the relevant knowledge, but also perceive the new problem to be similar to a previous problem. Of course, for transfer to be correct the new problem must also be structurally similar to the previous. Perceived similarity can be subjectively divided into close and far; the latter making transfer success less likely. Improvement of transfer likelihood to far domains comes with increased exposure to analogous problems in different domains. As an analogy consider an experienced DES modeller who has worked in similar projects across manufacturing, healthcare, the public sector and other domains during their career. They are much more likely to think in terms of the structure of the queuing problem than an individual who has had involvement in only a single simulation study. Lastly, in daily life, transfer, if it occurs, is spontaneous; there is no hint to application. However, in the artificial world of the laboratory hints can be given to analyse access issues and aid the focus of training improvements.

3.5 Summary

This chapter provides the conceptual framework used for analysing learning in this research. The two learning categories of attitude change and transfer are combined under a theory-of-action framework. Attitude change is used to infer single-loop

learning: short term learning used to solve the immediate business problem. Transfer of learning is used to infer double-loop learning: a change in the governing variables of an individual resulting in longer term learning that can be applied elsewhere.

Attitude change has been studied extensively using the Theory of Planned Behaviour and theories of persuasion. Attitudes are conceptualised using an expectancy value model. That is, the set of salient beliefs behind an attitude can be measured and summated in order to calculate an indicator for overall attitudes. A substantial amount of literature has shown that these measures are good predictors of intention - perhaps to implement a specific option following a simulation study.

The heuristic-systematic model of persuasion provides two important results for understanding attitudes about credibility. Firstly, the Principle of Sufficiency introduces the concept of confidence thresholds. The higher this threshold the more effort is needed in order to generate the required amount of confidence. Hence decision makers with high thresholds in a simulation study will have to perform some systematic processing of model logic, data and assumptions during verification and validation. Secondly, evidence appears to suggest that the higher an individual rates task importance the higher this confidence sufficiency threshold becomes. For example, in a simulation study of an accident and emergency ward a manager may be under substantial pressure to reduce waiting times. Hence he or she is much more likely to engage in systematic processing when assessing the credibility of the model as they must be very certain their decisions improve performance.

Transfer of learning is most likely to occur when an individual perceives a new problem to be similar to one previously encountered. Of course, for successful transfer to occur the old and new problems must be structurally similar. As the surface similarity of the problems decreases, for example, if the analogous problem is in a new domain, the likelihood of transfer decreases. This is termed 'far' transfer. Individuals who are able to achieve far transfer appear to have a 'deeper' understanding

of the structural aspects of the problem in addition to the surface aspects. This is possibly because they have seen more examples or been given a way to abstract the underlying structure from the specifics of their example.

It should be acknowledged that any framework designed to measure learning can only provide indications. Thus there may be subtleties of learning outcomes and mechanisms that are missed. Nevertheless the framework should provide enough information to give an indication of differences between conditions within an experiment. Section 11.5.5 discusses potential limitations of the framework in greater detail.

The next chapter describes a methodology to measure double and single-loop learning using attitude and transfer theory. This is an experimental methodology set in a learning laboratory where participants either reuse a DES model or are involved in building it.

Chapter 4

Experimental Design and Predictions

4.1 Introduction

This thesis adopts a novel methodology for OR used to infer and compare learning within the theory-of-action framework. Learning is studied at the individual level within a laboratory experiment setting using case study and questionnaire materials. The learning studied is about a decision maker's approach to managing resource utilisation and process variation within a queuing system. These two behaviours/concepts are, amongst others, fundamental to the management of queuing systems and used by DES modellers when they approach a problem.

This chapter details an experimental design to test the hypotheses related to the commonly held belief that involvement of decision makers in the building of a simulation model aids learning. Learning is defined using the framework outlined in chapter 3. Single-loop learning is operationalised using the attitude towards behaviour construct taken from the Theory of Planned Behaviour (Ajzen, 1991). Changes in theories of action resulting from a double-loop learning process are in-

ferred using transfer of learning to new analogous problem solving scenarios and correlation between attitude change and transfer. To measure these two areas of learning, a participant is exposed to four separate stages in the experiment:

1. A pre-test questionnaire to measure initial attitudes
2. Involvement in a discrete-event simulation case study
3. A post-test questionnaire to measure attitudes after the simulation case study
4. A post-test questionnaire to measure reasoning about queuing problems

The first section of this chapter provides an overview of the independent and dependent variables and predictions for single and double-loop learning. In the experiment participants - undergraduate business students - are involved in a DES project to improve performance at a case study A&E department in England. So that the experiment can be recreated, if desired, the chapter details the pilot work, participants profile, case study, independent variables, dependent variables, predictions, experimental materials (i.e. simulation model and questionnaires) and procedure. Full details of the case study and model documentation can be found in Appendix A.

4.2 Pilot Experiments

In order to develop the procedure and questionnaires used in the experiment, it was necessary to run a series of pilot experiments. The pilot had six aims:

1. Develop the procedure to involve participants in model building;
2. Test if participants understood the written case study;
3. Test the order and clarity of instructions given to participants;
4. Test and develop the attitude questionnaire;

5. Test and develop the transfer scenarios;

6. Observe any areas worthy of further investigation.

Participants

In total 16 volunteers took part; these were from the Warwick Business School's MSc in Management Science and Operational Research, MSc in Business Analytics and Consultancy, and 1st and 3rd year undergraduate programme. Three pilot participants were also taken from the Warwick Mathematics Institute's 3rd year undergraduate programme.

Model Building Procedure

The early sessions focussed on how to build the model with participant involvement. The first approach was to build each stage of the model directly in front of the participant via a set route. In an attempt to make this feasible several Simul8 components were created that automatically loaded Visual Logic (Simul8's bespoke scripting language) into the model once included. After the first three sessions and feedback from the participants it became obvious that:

- Participants became frustrated if they did not have a choice on the order at which level of detail in the model is increased.

- After watching two stages of simple model building, participants could visualise how one model transitioned to another. That is, once an understanding of how models are built was gained, participants only wanted to see the end product as opposed to an analyst using Simul8 to create it for them.

The first observation led to a revised approach to involve participants in model building in line with the approach used in real world System Dynamic's studies by Lane (1994). Participants had freedom to choose how the model was built and the

modelling exercise became an approach to 'play back' participants assumptions to them.

The second observation meant that the model building procedure could be shortened. Instead of building every stage of the model in front of the participant using components, small sub-models were developed that incorporated the aspects of the problem modelled at different levels of detail. See Section 4.5.2 for the final procedure.

Case Study and Questionnaires

The questionnaire was continually revised throughout the pilot experiments. In fact, given the limited time of the volunteers, some sessions were entirely devoted to the case study and questionnaire. The procedure was as follows:

- Record participants beliefs about the advantages and disadvantages of 100% resource utilisation;
- Record participants beliefs about the advantages and disadvantages of process variation;
- Gain feedback from participants about the clarity of the case study (e.g what did they struggle to understand);
- Gain feedback from participants about the clarity of the questions within the questionnaire;

Transfer Scenarios

The transfer scenarios were possibly the most difficult area to set up. The approach has not been tried before hence, firstly, appropriate topics needed selection: these were found in an extensive search of papers available from the Winter Simulation Conference and the Journal of the Operational Research Society archives as well as numerous books on process modelling and improvement.

Secondly, the problems discussed in the conference papers and articles needed to be highly simplified. The objective of the scenarios was to trigger transfer of learning gained from the experiment. Thus the problems needed to also include cues that participants could detect. A debrief after each pilot was used to try and help improve the scenarios. This process resulted in a number of the scenarios being discarded altogether. This was due to the level of difficulty being too high. For example, a computer network queuing problem involving prioritisation proved to be too difficult for all pilot participants.

4.3 Design of Experiment

The experiment is based around a short case study problem set in an A&E department in a fictional hospital. It was designed to be suitable for DES analysis. This section provides an overview of the case study along with details of the independent and dependent variables.

4.3.1 Participants

Sixty four business undergraduates, with no simulation experience, volunteered to take part in the research (first year undergraduates = 41[64%], second year undergraduates = 23[36%], male = 35[55%] female=29[45%], age range = 18-22). All students were registered on Warwick Business School (WBS) modules; Table 4.1 illustrates the distribution of students accross degree programmes.

To encourage paticipation the students were paid a small fee for their time (£10 for the two hour conditions or £15 for the three hour condition). To improve participant motivation an additional cash prize was available for the best performance (£25). The research was advertised through WBS's online student portal 'my.WBS' (used for downloading lecture notes and reading WBS announcements), e-mail and announcements in lectures. Due to the length of time to collect data, procurement

Table 4.1: Participant Degree Courses

Degree	Count	% of Total
Accounting and Finance	27	42%
International Business	8	13%
Law and Business Studies	7	11%
Management	7	11%
Maths, OR, Stats and Econ.	5	8%
Computer and Business Studies	2	3%
General Engineering	1	2%
German and Business Studies	2	3%
Business Administriation	1	2%
Chemistry with Management	1	2%
Maths and Business Studies	1	2%
Philosophy, Politics and Economics	1	2%
Sociology	1	2%

of volunteers happened in two rounds: in the spring term of the 2008/09 and the autumn term of 2009/10 academic years. Sign-up to the research was through a free online meeting service (www.agreeadate.com). The mini-site contained instructions to participants including eligibility criteria (no simulation experience) and free slots.

4.3.2 Independent Variables

The experiment has one independent variable (IV) - the simulation study process - that takes three levels. Participants were randomly allocated to conditions to avoid potential systematic bias. Once all participants had signed up names were entered into Microsoft Excel 2007 (in the order they signed up). A list of random numbers U (0, 1) were then generated using Excel's built in data analysis pack - one for each participant. Using this number the participants were then placed into ascending order. The first participant was placed into level one of the IV, the second into level two, and the third into level three. The cycle is then repeated itself for all remaining participants.

Participants were either involved in model building and then extensive use (MB), involved in model building and then limited use (MBL) or the reuse of a simulation model (MR). Figure 4.1 illustrates the time participants spend on the simulation task depending on the level of the IV. Each condition manipulates the time spent participants spend involved in experimentation and model building. The following sections give an overview of each condition.

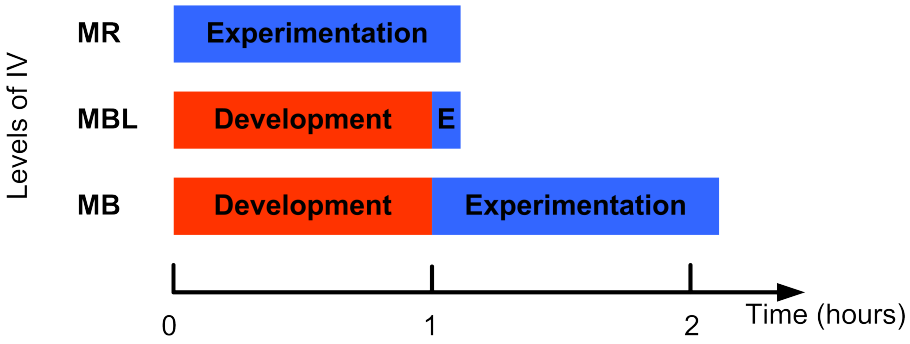


Figure 4.1: Levels of Independent Variable

Model Building (MB)

This condition lasts for a total of three hours and fifteen minutes. Participants spend two hours and fifteen minutes involved in model building and experimentation (See Figure 4.1) and a total of one hour for completing the questionnaires. Participants do not build the model directly, but instead take the role of a domain expert: they review the model and if there are issues with it decide what should be changed (see Section 4.5 for full procedure). Experimentation is part prescribed and part free choice. The case study (Section 4.3.3 and Appendix A.1) details six scenarios that must be investigated by participants. They also have the ability to select their own scenarios for simulation within the time limit. Volunteers of the MB condition are paid £15 for participation in the experiment.

Model Building with Limited Experimentation (MBL)

This condition lasts for approximately one hour and fifteen minutes. The time can only be approximate as the participants must act as domain expert during model building and complete all prescribed experimentation. Thus if participants have struggled with the building stage they may overrun slightly. The model building section of the experiment is identical to MB. Experimentation is manipulated so that participants only have time to explore three scenarios. Volunteers of the MBL condition are paid £10 for participation in the experiment.

Model Reuse (MR)

This condition lasts for one hour and fifteen minutes. The process is manipulated by removing model building. Instead participants are given a completed simulation model. Hence their task is to assess the credibility of the model and then perform the same experimentation as in MB. Although Figure 4.1 indicates that the full time is taken up by experimentation, some of the task time will be spent understanding the model (possibly through experimentation). Volunteers of the MR condition are paid £10 for participation in the experiment.

4.3.3 Case Study

The experiment is based around a case study problem. This takes the form of a fictional Accident and Emergency (A&E) department - St. Specific's - a simplified version of the generic models described in Fletcher et al. (2007) and Günal and Pidd (2006, 2009). Figure 4.2 depicts the full simulation model. The main simplification from models in the simulation literature is the limiting of diagnostic tests to x-ray only (i.e. the omission of additional diagnostic processes such as blood tests). This decision was made during piloting in order to limit the time spent on model building to a maximum of an hour and fifteen minutes. The simplified problem was deemed

to be suitably complex for the participants.

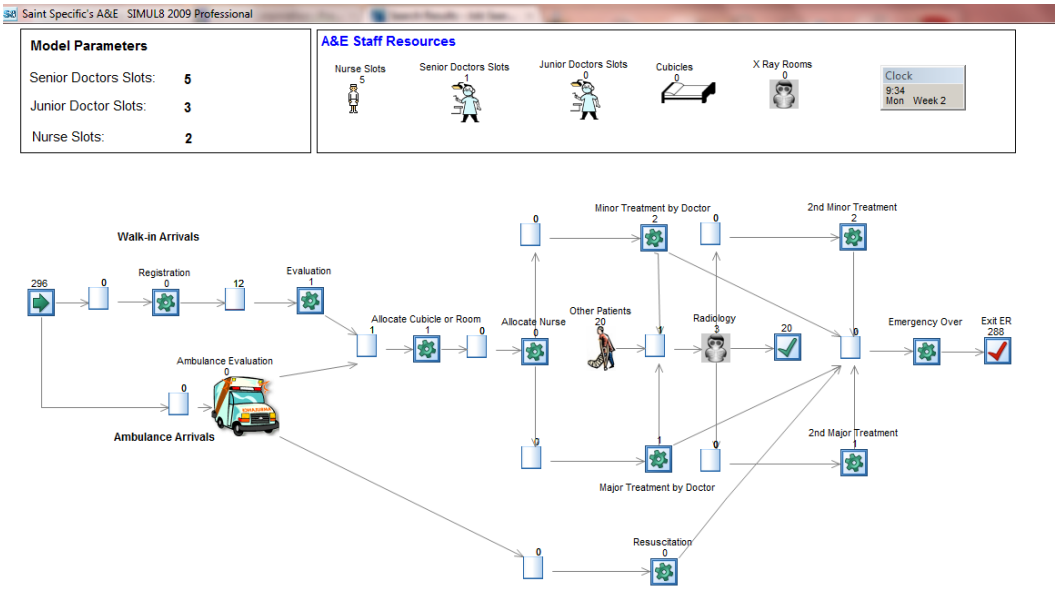


Figure 4.2: Case Study Model Visuals

The objective is to reduce the percentage of patients that are in A&E for longer than four hours over a six month time horizon. Specifically the participants are informed about the NHS target that 98% of patients must not spend longer than four hours in A&E. However, in the last six months only 85% of patients spent less than this time in the case study A&E.

Two main factors interact with each other to cause these issues in the case study A&E: process variation and resource utilisation. Current resource utilisation may seem quite satisfactory at first glance - the human resources seem reasonably busy (75% - 85%). However, the high variation in patient arrivals and A&E treatment makes it exceptionally difficult for the majority of patients to travel through the system in less than four hours. In fact the 98% target is very difficult to achieve.

Depending on the condition, the case study details a number of scenarios that the participant must investigate. These are split into three types: resource reallocation between shifts, increasing the numbers of resources and reducing process variation

in the radiology department. By examining these scenarios using the simulation model participants have the opportunity to learn about three concepts:

- Maximising resource utilisation of doctors and nurses over a six month period leads to a large number of performance target breaches.
- There is always a trade-off between the mean time patients spend in the A&E and the utilisation of A&E resources.
- Eliminating the variable arrival of non-A&E patients to radiology increases the performance of the system over a six month period.

Participants in the MB and MR conditions can also suggest new scenarios which may investigate the default scenarios further or something else entirely. For example, a participant may investigate the effect of patient prioritisation on performance.

The full case study that participants read can be found in Appendix A.1; variations on task descriptions for MR and MBL can be found in Appendices A.1.5 and A.1.6 respectively; details fo the simulation modl can be found in Appendix A.2.

4.3.4 Dependent Variables

Table 4.2 summarises the dependent variables included in the experiment. These are classified as either single-loop or double-loop.

Table 4.2: Summary of Dependent Variables

Single-Loop	Double-Loop
1. Attitude change	1. Transfer of learning
2. Credibility assessment	2. Correlation: attitude and transfer
	3. Confidence in reasoning

Single-loop learning is investigated through participants' attitude change. This

allows understanding of how the exposure to the simulation process affects a participant's intentions to tackle an instrumental problem - how can performance be improved at the A&E department over the next six months? Section 4.3.3 identifies three important learning areas within the case study; hence three attitude change variables are included (see Section 5.1.1 for a detailed description of each measure).

In order for single-loop learning to take place participants must assess and form an opinion about the credibility of the simulation model. There are two credibility measures included to test predictions. The first asks the participant to report how credible they consider the model and results ($CredAssess$). The second asks the participant to report how much self confidence they have that their assessment of the model is sufficient and correct ($SelfConf$).

As discussed, attitude measurement helps understand the change in a participant's intentions to tackle an instrumental problem. However, this fails to give information on whether the participant understands why a particular scenario or management approach helps performance and why another does not. When a participant has reflected on how they define effective performance/management for the system they have undergone double-loop learning. Indeed if participants have undergone a degree of double-loop learning in the experiment then they should be able to transfer this to new analogous problem solving scenarios. Thus the first double-loop variable listed in Table 4.2 explores the transfer of learning. Any degree of double-loop learning also involves single-loop (Argyris and Schön, 1996); hence a correlation between attitude change and transfer of learning is included as the second double-loop variable. The final double-loop variable considers the confidence participants have in their reasoning. This can be used to look for over and under confidence in the condition; for example do any of the conditions make more high confidence errors than the others?

4.3.5 Predictions

Single-Loop Predictions

The main hypothesis that this thesis tests is that decision makers learn about their problem when they are involved in model building. The first type of learning explored relates to single-loop learning: attitude change in order to solve an instrumental problem within the case study A&E department. The single-loop learning comparison tests four predictions; these are summarised in Table 4.3. In a real simulation study, one might argue that a decision maker is neither right nor wrong, but instead makes a decision based on his or her own worldview. Hence attitude change is neither incorrect nor correct. In the simplified world of the laboratory experiment, it is possible to determine the direction of attitude change that improves performance of the A&E system and the direction of attitude change that does not improve performance. For the sake of clarity and understanding these directions will be labelled as 'the correct direction' (hypothesis s.1) and the 'incorrect direction' (hypothesis s.2) of attitude change.

If it is true that decision makers involved in model building are learning more than decision makers involved in model reuse then it is expected that they are more likely to form intentions to act on their attitude. In other words it is predicted that participants involved in the model building conditions (MB and MBL) will report a larger change of attitude in the 'correct direction' than the model reuse condition (MR).

It is naive to assume, however, that all participants will change their attitudes in the 'correct direction'. In simulation studies, some decision makers will not accept results; have a strong opposite view; fail to understand the message or have other reasons for why some other action should be taken. In a simplified laboratory setting it is assumed that participants may also experience attitude change in the 'incorrect direction', but that political agendas for not accepting results are minimised. How-

Table 4.3: Single-Loop Learning Hypotheses

Hyp	Measure(s)	Prediction
s.1	Attitude change in the correct direction	MB & MBL > MR
s.2	Attitude change in the incorrect direction	MB & MBL < MR
s.3	Credibility assessment score	All three conditions rate the model similar
s.4	Self confidence in assessment of model	MB & MBL > MR

ever, if involvement in model building does aid the learning process then we might expect it to restrict attitude change in the 'incorrect direction' relative to the model reuse condition.

Hypotheses s.3 and s.4 concern the credibility rating participants give to the model and the results. Based on pilot observations and a review of the persuasion literature (Chaiken et al., 1989) it was predicted that participants would rate the model equally credible, but the type of credibility held would be different. It is expected that participants involved in building the model will have applied more scrutiny to the model logic, assumptions and simplifications - as part of verification and validation (V&V). Thus any assessment of credibility they hold will have evidence (memories) to back up their judgement. This is named *experienced credibility* and model builders will have high self confidence in their assessment. Conversely it is expected that model reusers will have applied less scrutiny to this part of V&V and possibly used a heuristic approach to assessment (e.g. the PhD student is an expert). These participants will hold less confidence in their overall assessment of the model.

Double-Loop Predictions

The second set of predictions concern double-loop learning: a change in both the attitude and understanding of the participant. The double-loop learning comparison tests five predictions; these are summarised in Table 4.4. Understanding is

operationalized as firstly the prediction of performance in the case study and then transfer of learning to analogous problem scenarios. Close transfer refers to problem domains that have a high surface and structural similarity to the simulation model (i.e. set in a healthcare domain). Far transfer refers to problem domains that have a low surface similarity to the simulation model (i.e. set in a call centre and manufacturing domains).

Table 4.4: Double-Loop Learning Hypotheses

Hyp	**Measure(s)**	**Prediction**
d.1	Prediction Success	MB & MBL > MR
d.2	Total Transfer Success	MB & MBL > MR
d.3	Close Transfer Success	MB & MBL > MR
d.4	Far Transfer Success	MB & MBL > MR
d.5	Correlation: attitude change and transfer	Correlation present in all conditions, but MB & MBL > MR

In Table 4.4 hypotheses d.1 to d.3 list predictions that participants involved in MB and MBL will achieve higher transfer success than MR at a total, close and far level. The final hypothesis acknowledges double-loop learning cannot be achieved without its single-loop component. Thus it is expected that a correlation will be present in all conditions, but that the correlation coefficient will be larger in MB and MBL.

4.4 Questionairre Design

Research on single and double-loop learning has primarily focussed on methods that encourage double-loop. In fact a literature review revealed only one study to incorporate measurement of both single and double-loop learning. Lahteenmaki et al. (2001) developed a questionnaire for double-loop learning to be administered within an organisation. One problem with the approach is that it assumes that what

people report is not subject to face saving mechanisms and uses direct questions. For example, when an individual is asked 'What is the organisation's commitment to the change process?' or 'what is your commitment to change processes/open communication?' the reported answer is likely to be consistent with an espoused theory-of-action (for example, 'of course I have high commitment to change') as opposed to one that is actually in use. The assumption being that these are potentially embarrassing questions: automatic and highly skilled defensive routines would control responses Argyris (1992).

Due to the lack of studies with reusable scales and approaches, the materials used to measure learning in this research are novel. In addition to the case study described in Section 4.3.3, a questionnaire is administered pre and post test to measure participant's attitude change and views on study credibility. Following completion of the post-test questionnaire participants answer nine reasoning questions about the case study and transfer scenarios. This section discusses these materials in detail. The full questionnaire can be found in Appendix A.3.

4.4.1 Attitude Questionnaire

The questionnaire has two parts: attitude measurement and credibility measurement. The first part, attitude measurement, is administered both pre-test and post-test.

The measurement of each attitude is based on the assumptions of the Theory of Planned Behaviour (TPB) (Ajzen, 1991). This assumes that an attitude is formed from a number of *salient* beliefs about behaviour. For example, when considering service level and utilisation a participant may hold the belief that maximising the utilisation of nurses increases stress levels. Note that there may be many beliefs some of which conflict with others. Attitude measurement is operationalised as follows. A participant provides a subjective probability (b) of the likelihood that a

belief is true and, secondly, an evaluation if the outcome of that belief is desirable (e). As an example consider the following belief:

- Maximising resource utilisation at St Specific's will reduce queues and improve performance over the next 6 months.

A participant in the experiment will firstly rate the likelihood that this belief is true (b) on a seven point scale (1= Extremely Unlikely, 7 = Extremely Likely). They will then rate how desirable the outcome of this belief is (e) on a bipolar scale (-3 = Extremely Bad, +3 = Extremely Good). As there may be a number of beliefs about one particular behaviour, the measure of attitude A_i is constructed from the summation of the j products of the subjective probability and outcome evaluations. This is summarised in (4.1).

$$A_i = \sum_{j=1}^{k} b_j \times e_j \tag{4.1}$$

Where attitude i is made up of k beliefs. Under TPB assumptions high attitude scores are good predictors of behavioural intentions.

4.4.2 Credibility Questionnaire

The credibility questions are administered to participants in the post-test questionnaire. The first measure $CredAssess$ is based on Gaziano and McGrath's (1986) credibility index. Five measures are used: believability, lack of bias, accuracy, representativeness and reliability. These are all operationalised as seven point Likert Scales (1 = completely disagree to 7 = completely agree). A scale of seven was chosen to be consistent with the seven point scales used in the attitude questions. All five measures are summated into a credibility index score (out of 35).

The second measure $SelfConf$ is based on the scales used by (Petty et al., 2002) to investigate the self confidence participants held in their answers to a questionnaire.

This is operationalised as five nine-point scales (1 = not at all confident to 9 = extremely confident). Each scale is related to the confidence participants have in their assessment of the believability, lack of bias, accuracy, representativeness or reliability of the model and its results. Answers are summated into a self confidence index (max 45).

4.4.3 Reasoning Questionnaire (Transfer of Learning)

The reasoning questions are split into three types: case study specific, close transfer scenarios and far transfer scenarios.

Case Study Specific Reasoning

The first question participants answer is related to the A&E case study. In the case study resources such as cubicles and doctors are shared between all patients (pooled resources). This pooling of resources helps speed up a process subject to variation and also increases resource utilisation. A simple explanation of this is that resources, if available, can go to where demand is waiting. Participants are asked to predict what would happen if A&E resources were split up and dedicated to specific emergency streams (i.e. whether performance increases or decreases) and to state why this change in performance occurs. Answers are analysed qualitatively. Firstly, on whether the participant predicts a decrease in performance and secondly whether the reasoning behind the prediction is correct.

Transfer Scenarios

The transfer scenarios are divided into two groupings: close and far. All four close transfer scenarios are set in a healthcare context. This closeness is therefore the *surface similarity* of the scenario to the case study. In contrast the far transfer scenarios have less surface similarity to the case study, but are still *structurally sim-*

ilar. The far transfer scenarios are either set in call centres or a food manufacturing plant. Each scenario details a problem, provides transfer cues, lists multiple choice answers and provides space for a qualitative answer. A scenario will look for the transfer of one out of two concepts from the simulation experiment:

1. Recognition of the relationship between resource utilisation and the speed at which an entity can travel through a process;
2. Recognition that eliminating variation from a process can increase the speed at which an entity can travel through a process.

The case study contains numerous examples of these concepts in action. For example, participants run experiments analysing the allocation of nurses to shifts and the use of resources to learn about resource utilisation and performance. If a degree of double-loop learning has occurred the participants should be able to transfer these concepts beyond the case study to analogous scenarios. Table 4.5 lists the context of the scenarios and the concept that is tested within each. Refer to Appendix A.3 for details of the scenarios.

Table 4.5: Transfer Scenarios

Scenario	Context	Reasoning required for transfer success
$S1$	GP's Surgery	Process Variation linked to Performance;
$S2$	A&E department	Resource Utilisation and Performance;
$S3$	Operating Theatre	Resource Utilisation and Performance;
$S4$	NHS walkin Centre	Process Variation linked to Performance;
$S5$	Pie Factory	Resource Utilisation and Performance;
$S6$	Police Call Centre	Process Variation linked to Performance;
$S7$	Pie Factory	Process Variation linked to Performance;
$S8$	Call Centre	Resource Utilisation and Performance;

There are two ways to analyse the answers participants provide to the transfer scenarios.

1. Assume that the reasoning behind the multiple choice answers is known beforehand.

2. For those correct answers, assume that a significant proportion of the participant's reasoning is captured in the qualitative information provided and that it is accurate. Interpret if the basic description of reasoning provided by the participant is correct.

The first of these approaches is problematic as it may be that participants have chosen the correct answer by chance. The second approach requires multiple judges; differences in interpretation of answers must also be resolved.

Multiple judges are not available for the interpretation of the reasoning answer. Instead the data are coded multiple times by the same judge with a time lag in between each coding. Two codings of the reasoning answers will take place with at least three months in-between. For acceptability these are expected to have high reliability score - measured using Cronbach's α and intra-class correlation (Field, 2009). In addition to minimise any judge bias (as the experimental hypothesis is known) all participant details are hidden from view and the order of answers randomised in each coding.

Transfer of learning is coded as a binary variable: zero for correct choice and incorrect reasoning (transfer failure) and one for correct choice and correct reasoning (transfer success). As an example of the coding procedure consider the information/cues provided to a participant by scenario two:

1. A hospital is going to experience an increase in demand (amount unspecified).

2. Current average utilisation of human resources is approximately 75%.

3. Current average utilisation of cubicles is 60%.

4. Similar to the hospital in the simulation the hospital must meet a performance target (98% of patients spend less than four hours in A&E).

The performance target and percentages provide the cues to recall the performance of the simulation model. The participants are asked if they agree that more staff should be introduced. All participants are asked to give the reasons why they think their choice would improve the performance of the system. An example answer provided by a participant is

> 'As the workload is going to increase, it is important to increase the number of staff. Depending on increase in workload the number of cubicles may be necessary to increase, but is unlikely as their utilisation is currently lower than staff utilisation.'

The second sentence provides the most detail on the participants reasoning. The participant is thinking in terms of maximum capacity. If the resources are working at less than 100% then they can cope. If 100% is exceeded then more are necessary to serve them. Whilst this seems sensible it fails to transfer any learning from the experiment. An example of the transfer we are looking for would discuss performance levels and the relationship to utilisation: even if utilisation is increased to 90% there will be a change in the performance of the system. Hence this answer is coded as a zero - correct choice, but failure to transfer learning.

Confidence in Reasoning

Measurement of the confidence participants hold in their answers to the transfer scenarios is, again, based on the scales used by (Petty et al., 2002). At the bottom of each transfer scenario participants are asked to rate the confidence they have in their answer on a nine-point scale (1 = not at all confident to 9 = extremely confident).

4.5 Experiment Procedure

All experiments took part in rooms supplied by WBS. Unfortunately the same room could not always be used due to scheduling conflicts with other WBS functions. During model building and use of the simulation models participants are sitting with the researcher in front of a desktop Personal Computer (PC) loaded with Simul8 2009 education edition (Simul8, 2009) and Microsoft Excel 2007. The researcher modifies the simulation model, during building and experimentation, at the request of the participant. Results from the model are automatically loaded into a Microsoft Excel Spreadsheet for participants to review as they wish.

This section details this procedure. The running of the experiment is broken down into the following sections: introduction to simulation, model building, model building with limited experimentation, and model reuse.

4.5.1 Introduction to Simulation

All participants read a case study based around reducing the length of stay in the St.Specific's A&E department and they are told that they must take the role of management. They read that the A&E department is performing poorly against a government target and that a recent consultation with staff has suggested a number of operational improvements. The task is to use simulation to evaluate the options and how A&E performance against the target can be improved.

Introduction to Simul8

All participants are then introduced to Simul8 (Simul8, 2009) - the software used for the study - and informed that it is commercial software used in industrial and healthcare domains. The visuals of Simul8 and basic principles of DES are explained through the construction of a simple - one queue one resource - model representing a very high level view of the A&E department. This is illustrated in Figure 4.3

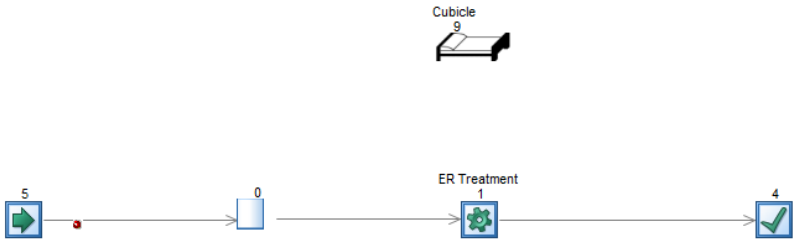


Figure 4.3: Introductory Model

The model is firstly setup as deterministic; participants watch patients (animated using Simul8's default small red icons) arriving at equal intervals, taking a cubicle resource, always staying the same length of time in A&E and then leaving. This queuing system is in equilibrium as a patient always leaves just before a new patient arrives. Thus no queues form and the time a patient spends in the system is constant. This allows the participant to be easily introduced to the performance measures generated by Simul8 for resource utilisation and percentage of patients breaching the four hour target.

Variability and Replications

The deterministic simplification is then removed and sampling from distributions is introduced. The visuals are again watched; now showing small queues building up and falling. Performance measures show an increase in breaches and resource utilisation (due to the process variation).

Participants are asked to remember the values of the performance measures and then compare these to two additional runs of the simulation model - each run uses different streams of random numbers (Simul8 provides a menu item for quick access to this function). To explain the discrepancy between the performance measures of each run participants are told that the data provided by the case study is a sample

and it is not possible to predict the exact result from period to period. Replications are then introduced (referred to as 'Trials' in the experiment to correspond to Simul8 terminology) as a method for studying the average performance of a system along with the distribution of a performance measure (although a long run is a more efficient method of data collection for a non-terminating system, it was decided that the concept of multiple replications is quicker to explain and understand).

The majority of participants are familiar with the concept of a confidence interval from the quantitative course on WBS's undergraduate program. However, this is briefly explained again and linked to black box validation of the simulated output (Robinson, 2004). Further validation techniques are introduced at a later point.

Introduction to Experimentation

Participants are told that scenarios can either be quantitative or process based. Firstly, the predefined scenarios within the case study relate to resource reallocation or increases are noted as examples of quantitative changes. An example scenario is quickly run for the simple model: an extra cubicle increases performance. Here participants are reminded of the objective to increase performance, but also told they must weigh up any performance increase against the number of extra resources required (no specific costs of resources are given). Secondly, the predefined scenario within the case related to radiology is noted as an example of a change in process. Participants are told that sometimes these changes may not carry monetary costs, but may have organisational issues.

4.5.2 Model Building

Participants involved in the MB condition then go through five additional stages of model building and validation. At each stage participants are presented with a visual model they can watch or step through, a results screen (from replications)

and a single A4 page detailing the change to the model, simplifications included, and assumptions. Participants review this information and must decide if the model is fit for purpose or suggest how it should be changed.

Stage One

In the first stage the researcher always increases the detail of the model in the same way. The tricky concept of modelling doctor and nurse multi-tasking is handled here (See Günal and Pidd, 2006). Participants are prompted to ask for a repeat explanation if they are not clear. The final visual is illustrated by Figure 4.4.

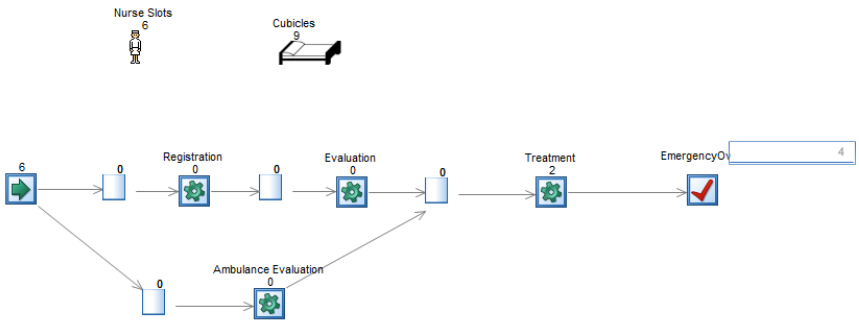


Figure 4.4: Model Building: Stage One Model

Ongoing Verification and Validation

Following the completion of the stage one model, the results and validation dashboard is introduced. Figure 4.5 provides a screen shot of the batch run results from the final model; results vary depending on the simplifications and assumptions included in the model. The results screen details the following areas:

- Performance target breach estimates (overall and by emergency type);
- A&E Resource utilisation estimates;
- A time series and frequency distributions of target breaches by emergency type;

- A breakdown of nurse resource utilisation by shift;
- A frequency distribution comparison between simulated and actual data.

The results screen allows the magnification of any of the charts through clicking on the relevant chart.

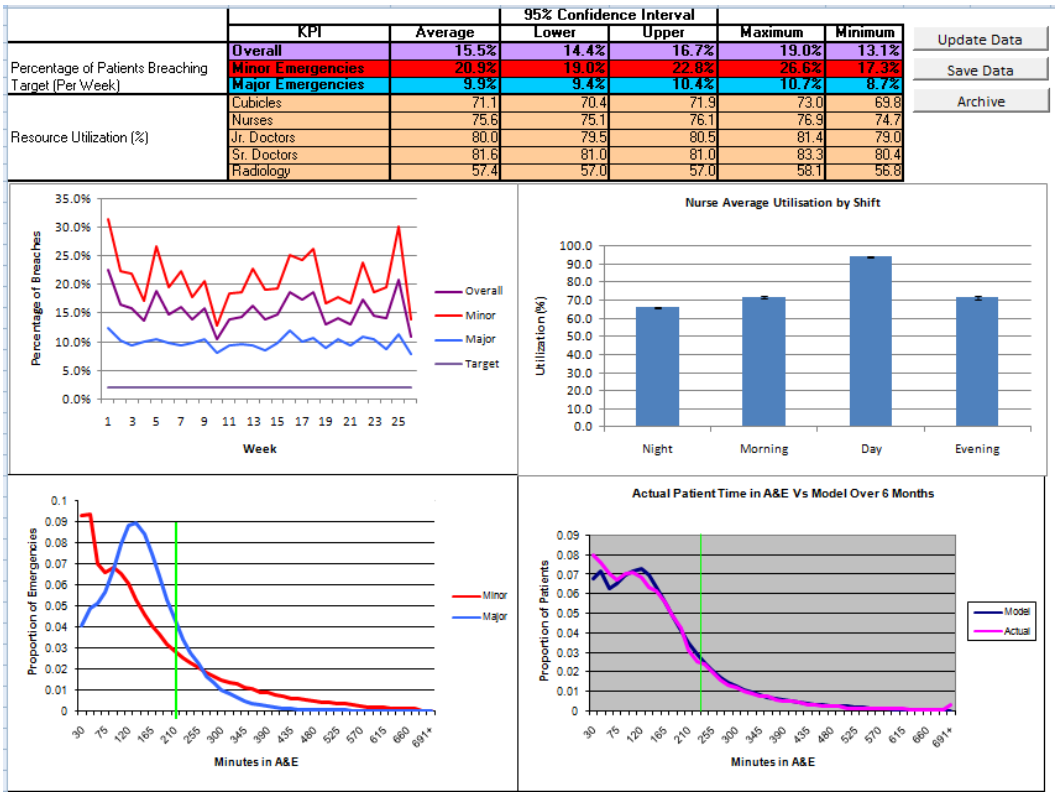


Figure 4.5: Results Dashboard Screenshot

Participants are told that the results screen provides help with assessing if the model is fit for purpose. An objective validation criterion is provided by a comparison of a sample of performance from the case study to a confidence interval for the performance target produced by the model. Subjective validation is operationalised by comparing frequency distributions of simulation output to sample data from the case study (Sargent, 1996). Participants can also look at the other outputs and decide if they appear sensible.

In addition to getting batch run results, subjective validation is opertionalised by conceptual model review (Law, 2007); participants are informed that they can ask about model logic and workings, watch the model running and question any simplifications and assumptions included in the model. Simplifcations and assumptions are provided to the participant on a single page of A4. Table 4.6 provides an example of this at stage one.

Table 4.6: Exert from Stage One Conceptual Model

Simplifications
• Nurse multi-tasking is not modelled in detail. For example, if a nurse can do two things at once then two nurse slots are included in the model;
• An average number of nurses are included. Thus the staff shifts are not explicitly modelled;
Assumptions
• Nurses always correctly evaluate the level of emergency a patient represents.

Participants can ask to remove any of the simplifications or clarify their meaning. The first simplification in Table 4.6 is an example of a simplification that cannot be removed. If a participant asks for an explanation of the simplification or for it to be removed the researcher provides a reason why it is included. In the case of the first simplification listed above, no data is available to model the movement of nurses in greater detail and it is difficult to make any sensible assumptions. Participants can ask for repeats of this explanation at any stage in the experiment. The second simplification is an example of a where the model is oversimplified. Participants need to ask to remove it, in order to reach the final model.

Assumptions are made either when there are uncertainties or beliefs about the real world being modelled (Robinson, 2008). Thus participants questioning the assumption in Table 4.6 are asked to refer back to the case study and told that any assumptions they are unhappy with can be explored during experimentation (if desired).

Level of Detail

At stage one the model produces inaccurate output due to its oversimplification. The oversimplification also means that the model cannot simulate the scenarios specified in the case study. Participants are told this and asked - as the domain expert - to advise on what they think should be modelled in greater detail or changed. The choices participants can make are fairly limited and thus anticipated; although it may not appear that way to the participant as they are not given a choice of what to change. They must consider the information given to them, the current model and details in the case study. At stage one, for example, the participant could request to increase the detail on

1. Inter-arrival times of patients and corresponding nurse availability;
2. Doctor treatment of different types of emergency patients;
3. Routing of patients through the radiology department.

Figure 4.6 illustrates this procedure. Once the participant has chosen to increase the detail in one area the corresponding model is selected from a set of existing sub-models. The review/modify cycle then begins again until the participant is satisfied the model is fit for purpose. The results look extremely close at stage five; however, participants may continue to consider what can be altered and test simplification and assumptions. Full details all of choices participants can make at each stage of the model building process can be found in Figure A.1 of Appendix A.2.4.

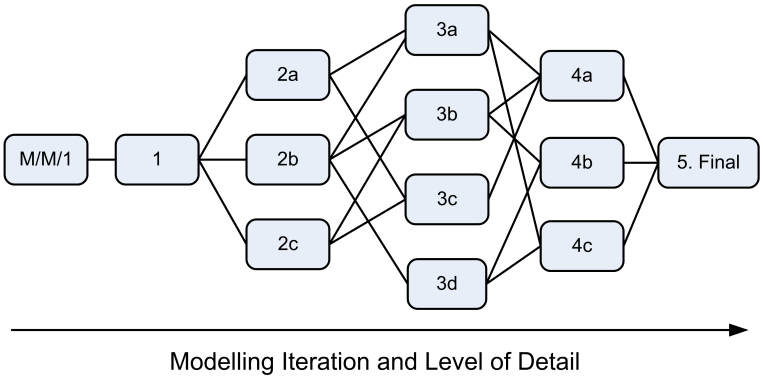


Figure 4.6: Model Building Procedure

Experimentation

Model building participants must investigate six scenarios predefined in the case study (these are listed in Appendix A.1.4). These relate directly to the three attitudes measured pre and post-test. They are also free to suggest new scenarios based on their own ideas. These new scenarios may be directly relevant to the measures of the experiment - i.e. further exploration of resource reallocation, resource increases, the radiology process or a combination of them - or may be something new entirely. After each scenario is run it is again loaded into Microsoft Excel, but this time the batch run results are reviewed in a scenario comparison worksheet. Figure 4.7 is an example screen shot. If a scenario is statistically different from the base case the scenario name is highlighted in red (reduced performance) or green (increased performance). Participants can review a confidence interval for mean differences or drill further into the data by clicking on the charts and buttons included in Figure 4.7. Each chart and statistic is explained in the same way to all participants.

Before analysis of every scenario, participants are reminded that - if desired - they can watch the model's visual display as well as review batch run results. Additionally, once results are loaded into the scenario comparison spreadsheet participants are reminded that they can drill further into the data if desired.

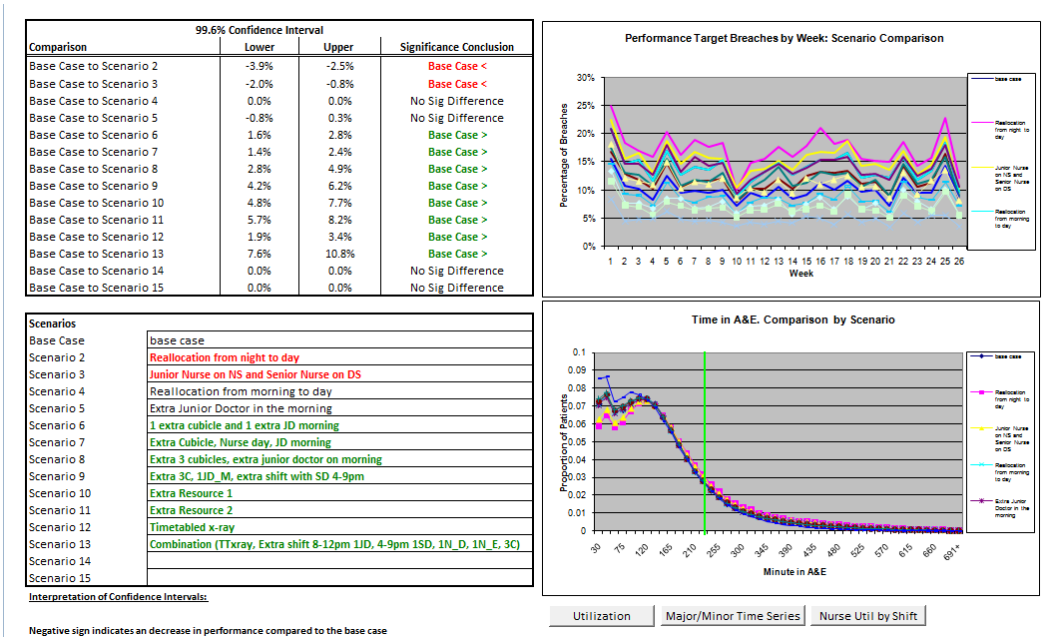


Figure 4.7: Model Building Procedure

4.5.3 Model Building with Limited Experimentation

Participants follow exactly the same procedure for simulation education, building the model and V&V as MB. Once building is complete participants are only permitted to perform the three scenarios defined in the case study (listed in Appendix A.1.6). These are a subset of the six predefined scenarios performed by MB. This means that MBL participants still have the opportunity to review results related to reallocation of resources, extra resources and scheduling radiology sharing, but not to as great an extent as MB (or MR). The results of these scenarios are again reviewed using the spreadsheet illustrated in Figure 4.6.

4.5.4 Model Reuse

Participants in MR undergo the same simulation education as MB and MBL. However, MR participants are informed that a model was developed for a different hospital with a similar process and objective. They are then presented with the full

computer and conceptual models as well as the batch run results. The following procedure is then followed:

1. The model is run in visual interactive mode and participants are talked through the process; for example arrivals, priority queuing, emergency treatment and visiting radiology;

2. The explanation for the result screen (Figure 4.5) given to MB and MBL at stage one of model building is repeated;

3. Participants are presented with a list of assumptions and simplifications in the model.

An alternative procedure could present the list of assumptions and simplifications to the participants prior to the Simul8 model. This would be equally valid, but experience in the pilot suggested that participants were not really sure what to do with them until after they had viewed the computer model. The order detailed above appeared to help participants see the relevance of the assumptions and simplifications and reflect on the simulation model.

Participants are reminded that the model was developed for another similar hospital. Thus they must assess the models fitness for purpose. They are allowed to ask questions about the models logic, results and simplifications and assumptions. Once participants are satisfied that suitable V&V has been completed, the same experimentation procedure is undertaken that was used in MB.

4.6 Summary

This chapter provides an overview of an experiment to test the hypotheses that involvement of decision makers in model building aids learning - both single and double-loop. Figure 4.8 provides a high level view of the experiment. Attitudes

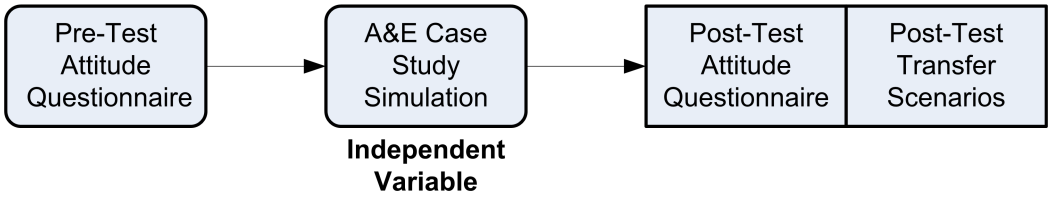


Figure 4.8: High level view of experiment

towards the management of three aspects of an A&E queuing system are measured before and after the use of a simulation model to analyse the problem. The simulation study process that participants are involved in is manipulated depending on the condition: participants are either involved in development of a model followed by extensive use (MB), involved in development of a model followed by limited use (MBL) or involved in the reuse of a model (MR). It is predicted that the MB and MBL should aid attitude change in the 'correct direction' and restrict attitude change in the 'incorrect direction' more so than MR.

Following the post-test attitude questionnaire, participants answer reasoning questions that present analogous queuing problems to the A&E department. These transfer scenarios are classed as either close, i.e. within a healthcare context, or far, i.e. within call centre and manufacturing contexts. It is predicted that the MB and MBL conditions will achieve higher transfer success relative to MR. It is further predicted that all conditions will show some extent of double-loop learning, i.e. a correlation between attitude change and transfer, but that the relationship will be strongest in the model building conditions.

The procedure developed for model building requires the participant to take the role of a domain expert rather than a modeller. This means that the participant is involved in conceptualising and validating the model, but not direct building. Participants are able to make choices about the level of detail within the model and review the results of their suggestions. Although the choices participants can make are limited they can make them in an order of their choosing or apparent

obviousness.

The next chapter provides a description of single-loop learning for participants in each condition. This is followed by a formal comparison of the conditions and test of predictions.

Chapter 5

Single-Loop Learning Results

5.1 Introduction

A theory-of-action perspective on learning assumes that individuals have a definition of effective performance for a system. For example, an individual may define effective performance of an A&E system as 'very high utilisation of resources and quick turnaround of admission, treatment and discharge of patients'. When working on an instrumental problem, such as improving performance of an A&E department against a target, individuals will strive to meet the objectives of this definition of effective performance - perhaps without realising any relationships between objectives or factors that may be missing. If this happens then an individual might be surprised when the results of their actions do not fit with their expectations of performance. Under theory-of-action assumptions an individual is more likely to try to find solutions that fit with their definition of effective performance than examine the definition itself. In other words attitudes in what action to take may change, but deeper understanding and objectives *may* remain the same. This is called single-loop learning and is the focus of this results chapter.

The chapter is split into four sections. This introductory section provides a re-

view of attitude and supporting variables used in the analysis. The second section discusses the analysis methodology used in the research and the alternatives. The third section is organised by experimental condition and presents descriptive statistics of the three attitude change variables followed by exploration of within group differences using the process variables. The final section provides a summary of the key findings of the single-loop results. Chapter 6 then builds on these descriptive results with a comparison between each condition and a discussion of the support for predictions.

5.1.1 Attitude Measures

Throughout this chapter and chapter 6 three attitude measures are of interest. These all relate to management of the A&E queuing system. Table 5.1 details the meaning and interpretation of these variables. The first two of these variables $MaxUtil$ and $TradeUtil$ relate to a participants attitude towards an aspect of managing resource utilisation within the A&E department over the next six months. The third attitude, $ElimVar$, relates to a participants attitude towards reducing the variation in radiology resource availability over the next six months.

5.1.2 Supporting Measures

In addition to the three attitude variables that measure learning outcomes, seven process variables are analysed for each condition. No specific hypotheses are tested within this chapter. Instead the results are used as a possible source of explanation for different outcomes within a condition. Specifically the seven variables are used to explore differences between correct and incorrect directions of attitude change within a condition. For example, MB and MR participants experienced strong correct and incorrect attitude change on $TradeUtil$. The seven process variables are explored for differences between these two groupings within the MB and MR conditions.

Table 5.1: Attitude Measures

Attitude	Description
$MaxUtil$	The change in attitude towards pushing A&E resource utilisation to its maximum. A negative change represents beneficial attitude change resulting from the simulation, as very high utilisation is detrimental to system performance;
$TradeUtil$	The change in attitude towards trading off some resource utilisation to achieve higher system performance. A positive change represents beneficial attitude change resulting from the simulation, as this indicates that participants recognise that utilisation has a relationship to system time;
$ElimVar$	The change in attitude towards reducing the variation in the availability of radiology resources. A positive change represents beneficial attitude change resulting from the simulation, as lower variation in the availability of radiology improves long term system time.

The process variables are divided into two groups. The first of these are credibility measures.

- Median credibility assessment score;
- Median self confidence in the credibility assessment.

The second group of process variables give information on how participants searched the solution space.

- Percentage of scenarios including resource reallocation;
- Percentage of scenarios including extra resource;
- Percentage of scenarios including reduced variability in the radiology department;
- Percentage of scenarios including other variables;
- Mean number of scenarios simulated.

5.2 Analysis Considerations

Before proceeding to the results there are three issues that must be considered with the analysis. Firstly, a sensible approach to identify and exclude outliers is required. Given the small sample size of the experiment and the number of variables measured, it was decided to adopt a multivariate approach. Secondly, a particular statistical issue arises when working with pre-test post-test designs called regression to the mean. This phenomenon is described along with reasons why the subgroup analysis of conducted in this research warrants a specialised analysis of the data. Lastly, there are several approaches available for dealing with regression to the mean. These are briefly described along with the relative advantages and disadvantages.

5.2.1 Outlier Analysis

One problem with an experiment where multiple variables are measured is that there are more chances that univariate outliers (outliers on single variables) will occur. Given the small sample size of the experiment, it was deemed appropriate that cases would be excluded only if they consisted of a unique combination across variables (i.e. not extreme on an individual variables, but have unique multivariate profiles) (Hair et al., 2006).

The outlier analysis was run in an iterative manner, i.e. as outliers can mask other outliers (Wilcox, 2005) the analysis was repeated after outliers had been removed to verify no further outliers were present. The outlier analysis consisted of three stages. Firstly, univariate checks for outliers across all variables using box-plots, histograms and z-scores (scores standardised to the normal distribution so that extreme cases are more obvious). Secondly, bivariate checks were made using scatter plots. Finally a multivariate profile check was run using Mahalanobis depth (Hair et al., 2006; Wilcox, 2005).

The analysis revealed three cases that justified exclusion. In all cases the results

were quite different from others cases in the experiment. For example, participant MB4 appeared to provide a highly extreme reverse of the expected learning and gave an exceptionally low credibility assessment score. This meant that initially MB4 believed (in fact, his or her score was the highest) that pushing for 100% resource utilisation would reduce performance and that there is always a trade-off between resource utilisation and performance. By the end of the experiment the participant reversed these views (the change was extremely large). Reasons for this odd case were investigated by reviewing the tape recording of the experiment. However, there was no indication why the participant may have reported these values. In fact, the participant appeared to cope quite well with the experiment and in fact learn the opposite of what they reported. Similar results were found for the remaining two cases identified as outliers.

5.2.2 Regression to the Mean

One threat to the internal validity of pre-test post-test designs is the so called regression to the mean (RTM) effect (Rocconi and Ethington, 2009; Kelly and Price, 2005; Field and Hole, 2003; Bonate, 2000; Chaung-Stein and Tong, 1997; Rogers, 1980). In many experiments it has been observed that participants with extreme scores on a first measurement tend to report scores closer to the mean for that measure on a second measurement. This is because any quantitative measure is the sum of two components: the true value and the error of measurement (Rocconi and Ethington, 2009).

For example, consider the $TradeUtil$ measure in this experiment (pre-test mean = 22). A participant reported a pre-test score of -9 and a post-test of 18. On the surface it appears the participant has learnt a great deal given exposure to the simulation model. It is also clear that initially the participant does not favour a $TradeUtil$ behaviour. However, if the error component of the measure of the initial

attitude is inflated for some reason (e.g. due to the participant being particularly over enthusiastic and wishing to demonstrate their knowledge, misunderstanding the questionnaire, not engaging in the problem, or believing that this is what the researcher wished to hear) then this inflates the apparent importance of the condition. This change has nothing to do with an experimental effect and can mask effects between conditions or introduce one where none exists. This is particularly problematic if participants are selected for an experiment based on their pre-test score. That is, those selected may have extreme scores simply due to a large error component and the measurement of learning - or whatever is measured - may be subject to a RTM effect.

A common method believed to get round the problem of RTM is to conduct a controlled experiment - or at least one where the relative differences between conditions are compared - using random allocation of participants. This is based on the belief that the size of RTM will be similar across all conditions and the relative difference should still be valid. However, some warn that typical baseline scores must be the same for the approach to work (Chaung-Stein and Tong, 1997). The pre-test attitude change results were tested and no differences were found between conditions (Kruskall Wallis $p > .1$). One problem with this study, however, is that the analysis of attitude change focuses on subgroups of attitude change. That is, participants are selected based on their direction of attitude change which is strongly negatively correlated with their pre-test attitude. This leads to the possibility of extreme scorers influencing the results of any inference procedure. Furthermore, descriptive statistics and omnibus tests indicated that pre-test scores were not equal across subgroups for the $ElimVar$ measure ($p < .1$): indicating that the spread of the more extreme attitudes was not equal across subgroups.

In fact some, authors believe that the RTM effect must be taken into account whenever a pre-test score is negatively correlated with the change score (Rogosa,

1988, is often quoted). Analysis supports that this is the case with the attitude data for this research: $MaxUtil$ ($r = -.436$), $TradeUtil$ ($r = -.468$) and $ElimVar$ ($r = -.520$) were all significant ($p < .01$). Given this correlation and a potential imbalance across one of the subgroups, a comparison of the observed data, to data adjusted to account for the possibility of regression to the mean is advisable.

5.2.3 Procedures to deal with regression to the mean

A common post-hoc approach is adjustment of the change scores assuming that the effect of the independent variable is constant (Rogers, 1980; Chaung-Stein and Tong, 1997; Bonate, 2000; Rocconi and Ethington, 2009). Chaung-Stein and Tong (1997) show that in a pre-test post-test design the amount of RTM is given by $(x - \mu)(\rho - 1)$; where x is the pre-test score, μ is the mean of the pre-test scores and ρ is the test-retest correlation ($0 < \rho < 1$) (note that Rogers (1980) argument is similar). A simple interpretation of this quantity is that the regression effect is larger as the deviation of x from μ increases (i.e. the regression effect is worse for more extreme pre-test scores). The RTM effect is also worse the when ρ is low.

If the RTM effect is assumed to be additive then the expected observed treatment effect (the expected difference between two conditions Δ') can be considered as the sum of two components: the true treatment effect (Δ) and the regression effect. This is illustrated by (5.1).

$$\Delta' = \Delta + \sum (x - \mu)(\rho - 1) \tag{5.1}$$

The simple procedure is to firstly adjust all differences by $(x - \mu)(\rho - 1)$ and then take the average of the differences (Chaung-Stein and Tong, 1997; Bonate, 2000). Observed and adjusted change scores will be highly correlated; however, average change scores of subgroups will be slightly different. Furthermore, some authors prefer to think of this as an adjustment to the the pre-test score (Rogers,

1980) or the post-test score (Bonate, 2000). This aids understanding of the approach as users of the procedure can easily understand that pre-test scores that were above the (pre-test) mean are adjusted to be slightly lower than the observed score and vice versa.

It is acknowledged that this approach is only approximate - it could be nothing else - and can never remove all of the RTM effect and has the limitation of assuming that an experimental effect is constant no matter what the pre-test score (as opposed to a multiplicative effect). However, when used in a comparative nature it may help illustrate where regression to the mean may be masking or introducing experimental effects into the results and aid interpretation. The approaches taken in this analysis are:

- As inference tests only support a difference in the ElimVar distributions of subgroup pre-test scores, the observed ElimVar measures are accompanied by equivalent statistics adjusted for regression to the mean.
- All inference tests performed on the observed in section 6.4 are repeated for the adjusted data. Appendix B.3 provides a comparison of results between the observed and adjusted data. As an extra check these tests are repeated for all measures, not just ElimVar. The main text in section 6.4 provides a summary of the main differences in results.

It is worth noting that the other approach often recommended to deal with regression to the mean is Analysis of Covariance (ANCOVA). ANCOVA is also an adjustment procedure: it uses the covariate (i.e. pre-test score) to adjust the dependent variable (post-test score) and 'removes' the RTM effect (Bonate, 2000); so similar to a the procedures defined by others ANCOVA results include adjusted estimates of mean differences.

The papers that discuss ANCOVA for dealing with regression to the mean generally view it as better than post-hoc adjustment. However, as with many of the

more complicated statistical techniques the assumption of the approach and sample size can restrict its application. Firstly, the attitude data break some assumptions necessary for application. Here the data violates the assumption of homogeneity of regression slopes (a common problem in ANCOVA (Miller and Chapman, 2001)). Secondly, sample size is still a problem - perhaps to a greater extent in the more complicated analysis - as multiple comparisons of conditions and subgroups are needed. To rectify this problem non-parametric approaches are available based on rank transformations (Quade, 1967; Conover and Iman, 1982), multi-level linear models (Field, 2009) or resampling techniques (Wilcox, 2005). However, there seems to be three reasons why a simpler adjustment procedure is a more suitable choice:

a.) No practical examples could be found of the non-parametric ANCOVA procedures in use;

b.) The bootstrap approach for ANCOVA is more complicated relative to the simple bootstrapping of mean differences;

c.) The multi-level linear modelling approach requires an infeasible sample size for the current research and is substantially more complicated.

Lastly, it should also be noted that procedures other than adjustment and ANCOVA are available for dealing with RTM. However, these are considerably more complex and involved procedures; for example Krause and Pinheiro (2007) illustrate how the p-values can be adjusted in the presence of regression to the mean using a combination of modelling, simulation and bootstrapping. The simpler procedure outlined above was chosen as it was felt to be more transparent and allowed for both the adjusted and originals to be easily contrasted.

5.3 Descriptive Results

This section details the single-loop results by condition. Treatment of specific hypothesis for differences between conditions is reserved for Chapter 6. Description of each condition is split into two subsections. The first subsection presents the findings for learning outcomes - attitude change - within the condition. The second subsection explores the process variables for possible explanations of differences within groups.

5.3.1 Model Building

Table 5.2 presents the quartiles of the change in attitudes experienced by the MB participants. The second column details the direction of attitude change that is in the correct direction for managing the case study. The third to fifth columns in Table 5.2 detail the lower, median and upper quartiles of the measures. All three measures have similar variability ($IQR = 29$ to 34); this range always just includes zero indicating that there was some inconsistency in the direction of attitude change.

Table 5.2: MB Participant Difference Scores

		Quartiles		
Measure	**Correct**	Q_1	Mdn	Q_3
MaxUtil	−	-28.5	-7.0	0.5
TradeUtil	+	-4.8	7.5	28.0
ElimVar	+	-1.0	15.0	32.5

Correct = the direction of correct attitude change

Table 5.3 divides the distributions into positive (+) and negative (−) changes in attitude. For each attitude measure the direction of correct attitude change is highlighted in grey. For example, Table 5.3 shows that 70% of MB participants had a reduction in attitude $MaxUtil$ - this was a beneficial change in the case study.

Table 5.3: MB Participant Difference Scores by Polarity

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
MaxUtil	30%	70%	4.0	-21.5
TradeUtil	70%	30%	23.0	-16.0
ElimVar	70%	30%	25.0	-3.0

The direction of correct attitude change is highlighted in gray

Table 5.3 also reports the median change in attitude by polarity. For example, we can see that, although, 30% of MB participants experienced an incorrect increase in attitude $MaxUtil$ the median of this change appears to quite weak compared to that of the median correct change.

This difference is noticeably less so for the $TradeUtil$ measure. Incorrect attitude change (negative polarity) is of a similar size to correct attitude change (positive polarity). This result is surprising: incorrect attitude change was expected to be minimised in the MB condition. The large incorrect change in $TradeUtil$ indicates that this prediction backfired 30% of the time.

Regression to the mean

Table 5.4: ElimVar before and after adjustment for RTM: MB Subgroups

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
ElimVar	70%	30%	25.0	-3.0
ElimVar_Adj	80%	20%	15.6	-3.7

The direction of correct attitude change is highlighted in gray

ElimVar_Adj is adjusted to account for regression to the mean

Table 5.4 compares $ElimVar$ to the same variable after the data is adjusted to account for RTM - $ElimVar\_Adj$. It can be seen that there is a decrease in the

positive subgroups median value (from 25.0 to 15.6). This decrease is because the majority of the pre-test scores in the subgroup were below the overall pre-test mean for $ElimVar$. Thus there may be a tendency for these scores to increase simply because they were initially low.

Process Variables

Figure 5.1 is a spider diagram illustrating the process variables for the MB participants grouped by correct and incorrect directions of change in $TradeUtil$. The spider diagram allows for a multivariate plot of process variables; illustrating point estimates only. Large differences may indicate explanatory factors for attitude change in the correct and incorrect directions. In Figure 5.1 correct and incorrect groupings are represented with green and red lines respectively.

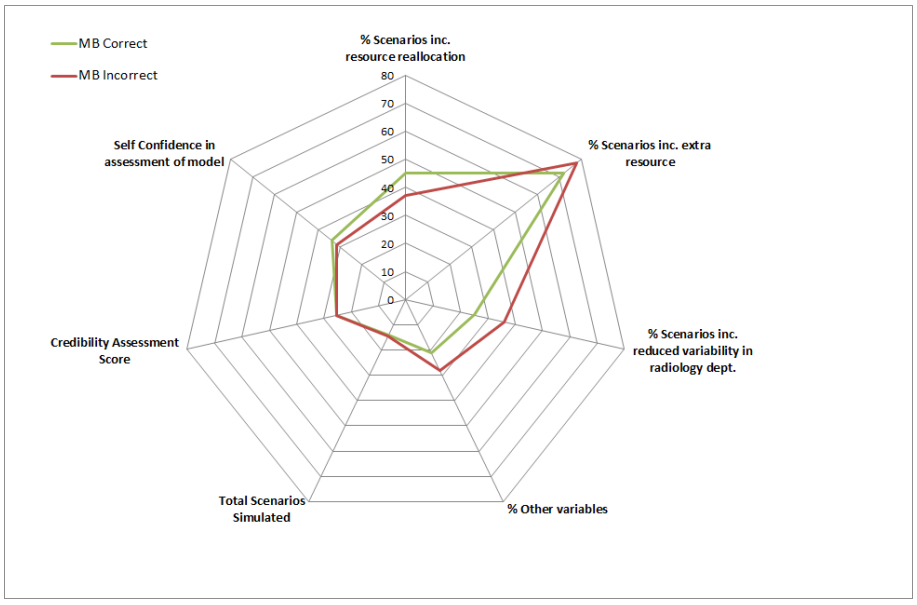


Figure 5.1: Process Variables Grouped by Correct and Incorrect $TradeUtil$

Figure 5.1 suggests that the two groupings were fairly consistent across the measures. The largest differences occur between the percentage of scenarios including resource reallocation and the percentage including reduced variability in radiology

department availability. The correct grouping gave more attention to the resource reallocation (45% to 37%) while the incorrect grouping gave more attention to radiology (36% to 25%). As resource reallocation within the simulation model will be testing the participants understanding of resource utilisation and performance then perhaps this extra attention was beneficial; although in practical terms we are only talking about one or two scenarios extra. There was also no evidence that those in the incorrect grouping or the amount of attention paid to radiology was related to change of $ElimVar$ ($r_s = .226$, $p > .1$).

No practical differences are found between the credibility measures of the groupings. Both typically rated the model and results as highly credible (25.5 out of 35) and typically had high self confidence in their assessment (34 and 31 out of 40 for correct and incorrect subgroups respectively). However, participants experiencing correct change were more variable in their self-confidence rating ($IQR = 8.8$) compared to the incorrect grouping (IQR = 3.5).

5.3.2 Model Building with Limited Experimentation

Table 5.5 presents the quartiles for the change in attitudes experienced by the MBL participants. The third to fifth columns in Table 5.5 detail the lower, median and upper quartiles of the measures, respectively. The $MaxUtil$ and $TradeUtil$ measures have similar variability (IQR's of 16.0 and 17.0 respectively) which is slightly less than $ElimVar$ ($IQR = 21.0$). The majority of the variation in $ElimVar$ comes from the bottom 25% of changes (i.e the median is closer to Q3). Notably, the upper and lower quartiles for $ElimVar$ are positive: this is a sign of consistency of attitude change in MBL.

Table 5.6 divides the distributions into positive (+) and negative (−) changes in attitude. For each attitude measure the direction of correct attitude change is highlighted in gray. For example, Table 5.6 shows that 81% of MBL participants had

Table 5.5: MBL Participant Difference Scores

		Quartiles		
Measure	**Correct**	Q_1	Mdn	Q_3
MaxUtil	−	-12.0	-6.0	4.0
TradeUtil	+	0.0	10.0	17.0
ElimVar	+	2.0	18.0	23.0

Correct = the direction of correct attitude change

Table 5.6: MBL Participant Difference Scores by Polarity

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
MaxUtil	33%	67%	6.0	-10.0
TradeUtil	81%	19%	15.0	-14.0
ElimVar	76%	24%	21.0	-9.0

The direction of correct attitude change is highlighted in gray

an increase in attitude $TradeUtil$ - this was a beneficial change in the case study. The majority of MBL participants experienced correct attitude change across the three measures. The highest proportion of correct attitude change occurs in the $TradeUtil$ and $ElimVar$ measures.

Table 5.6 also reports the median change in attitude by polarity. For example, we can see that 67% of MBL participants experienced a median reduction in attitude $MaxUtil$ of 10.0 while the remaining 33% experienced a median increase in attitude of 6.0.

Perhaps the most noticeable result for MBL participants is that the median size of positive and negative attitude change is similar *within* two of the variables. Attitudes $MaxUtil$ and $TradeUtil$ are both of small to medium size attitude change (with similar variation in scores as well) while $ElimVar$ is of a large size in the positive (correct) direction, but smaller in negative (incorrect) direction. Although

the proportions of correct and incorrect attitude change favour the correct side, in two cases attitude change is quite small and in all cases incorrect attitude change can be as strong as correct.

Regression to the mean

Comparing $ElimVar$ to $ElimVar\_Adj$ in Table 5.7 it can be seen that similar to MB there is a decrease in the positive subgroups median value (from 21.0 to 16.6). This again was due to a large number of pre-test scores in the subgroup falling below the overall pre-test mean for $ElimVar$.

Table 5.7: ElimVar before and after adjustment for RTM: MBL Subgroups

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
ElimVar	76%	24%	21.0	-9.0
ElimVar_Adj	76%	24%	16.6	-7.2

The direction of correct attitude change is highlighted in gray

ElimVar_Adj is adjusted to account for regression to the mean

Process Variables

The only relevant process variables for MBL participants are the credibility measures. Table 5.8 details the median scores and inter-quartile range (IQR) for the credibility assessment score and the self-confidence participants have in their assessment. The median credibility assessment score is reasonably high at 26 (scale maximum 35) and appears to have been fairly consistent (IQR = 2.5). The median self-confidence is also reasonably high at 31 (scale maximum 40); however, this was slightly more variable (IQR = 8.0).

Table 5.8: MBL Credibility Measures

Measure	Mdn	IQR
Assess Cred	26.0	2.5
Conf Assess	31.0	8.0

5.3.3 Model Reuse

Table 5.9 presents the quartiles of the change in attitudes experienced by the MR participants. The third to fifth columns in Table 5.9 detail the lower, median and upper quartiles of the measures, respectively. Again like MB and MBL the $MaxUtil$ measure is the least variable ($IQR = 15.8$). The range between the upper and lower quartiles incorporates zero for all three measures. However, note that the upper and lower quartiles for attitude $ElimVar$ are very close to zero MR participants favoured positive attitude change.

Table 5.9: MR Participant Difference Scores

		Quartiles		
Measure	**Correct**	Q_1	Mdn	Q_3
MaxUtil	−	-13.5	-9.5	2.3
TradeUtil	+	-2.3	11.0	27.5
ElimVar	+	-0.8	9.5	22.3

Correct = the direction of correct attitude change

ElimVar_Adj is adjusted to account for regression to the mean

Table 5.10 divides the distributions into positive (+) and negative (−) changes in attitude. For each attitude measure the direction of correct attitude change is highlighted in gray. For example, Table 5.10 shows that 75% of MR participants had an increase in the strength of attitude $ElimVar$ - this was a correct change in the case study. Like the other conditions, the majority of MR participants experienced correct attitude change across the three measures.

Table 5.10: MR Participant Difference Scores by Polarity

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
MaxUtil	35%	65%	7.0	-13.0
TradeUtil	70%	30%	22.5	-27.5
ElimVar	75%	25%	18.0	-5.0

The direction of correct attitude change is highlighted in gray

The last two columns of Table 5.10 report the median change in attitude by polarity. Perhaps the most interesting result to note is that, although the majority of MR experienced a correct attitude change on $TradeUtil$, the median incorrect attitude change was larger than the correct. This is a large change in the attitude of the participant. Indicating that a small proportion of MR participants were highly unlikely to trade off A&E resource utilisation in order to improve the service level of patients.

Regression to the mean

Table 5.11: ElimVar before and after adjustment for RTM: MR Subgroups

	Polarity of Change			
Measure	+	−	$Mdn+$	$Mdn-$
ElimVar	75%	25%	18.0	-5.0
ElimVar_Adj	75%	25%	18.5	-7.5

The direction of correct attitude change is highlighted in gray

ElimVar_Adj is adjusted to account for regression to the mean

Process Variables

Figure 5.2 is a spider diagram illustrating the process variables for the MR participants grouped by beneficial and non-beneficial change of $TradeUtil$. Correct and

incorrect groupings are represented with green and red lines respectively.

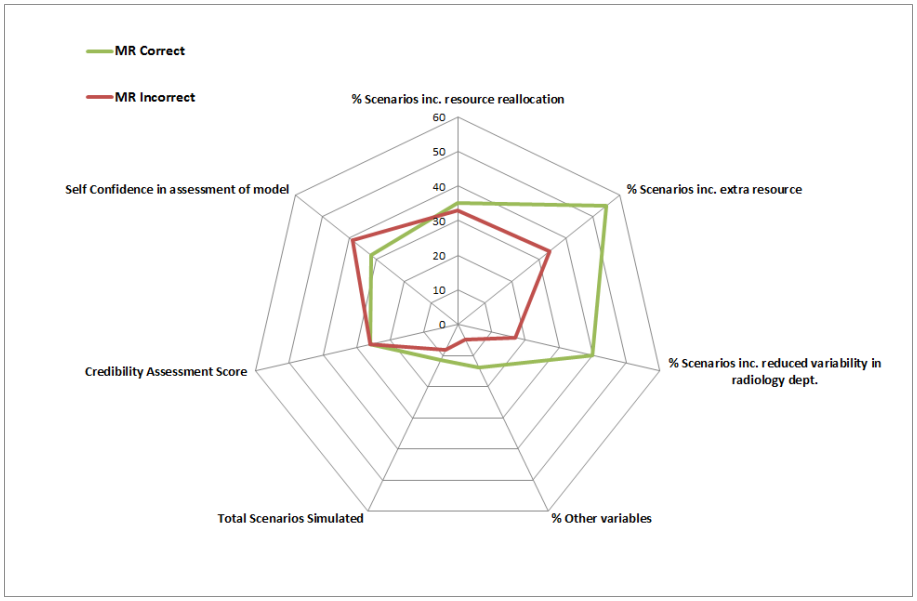


Figure 5.2: Process Variables Grouped by Correct and Incorrect *TradeUtil*

Figure 5.2 indicates some differences in the experimentation approach of the groupings. Firstly, the correct grouping performed more experimentation on average (11.6 to 8.1 scenarios). This perhaps is related to the amount of attention that a participant paid to scenario results or the motivation that they had to participate.

Secondly, there is a difference in the search of the solution space by the groupings. The incorrect grouping tended to look at experimental factors in isolation - illustrated by the low percentage of inclusion for each factor. The extra scenarios that the correct grouping investigated used these variables in combination and results in the high percentages for the experimental factors.

Both correct and incorrect groupings rated the model and results as highly credible (26 out of 35). However, the incorrect grouping reported more self confidence in their assessment of credibility ($mdn = 37$ out of 40) than the correct grouping ($mdn = 32$). Similar to the MB condition the correct grouping's self confidence was also more variable ($IQR = 11.5$) than the incorrect grouping ($IQR = 3.0$).

5.3.4 Summary

Table 5.12 details the main findings of the single-loop analysis by condition. Findings are grouped into five areas. The first group lists findings about the consistency of the attitude change. This details, for example, the most consistent outcomes of the experiment. The second grouping lists findings about the strength and polarity of attitude change. This details, for example, conditions where typical attitude change was particularly large or small. The third grouping provides a health warning about regression to the mean in the data and summarises the impact of score adjustment. The fourth group lists findings for how participants searched the solution space. For example, how did a participant who experienced correct attitude change on one measure differ in search behaviour to a participant who experienced incorrect attitude change. Finally, the last group lists findings about the confidence participants placed in the model, results and their own evaluation of the simulation model.

Table 5.12: Summary of Single-Loop Result by Condition

Consistency of Attitude Change
• The most consistent change across the three conditions was $MaxUtil$;
• The majority of participants change their attitude in a correct direction;
Strength and Polarity of Attitude Change
• Typical MBL change in attitude in $TradeUtil$ was of a small to medium size in either direction;
• Attitude change in the incorrect and correct directions could be large for MB and MR on the $TradeUtil$ measure;

Table 5.12 – continued from previous page

Regression to the mean

- Tests indicated a difference in typical pre-test scores for ElimVar across subgroups;
- The adjusted ElimVar measure provides a reduced estimate of change for MB and MBL;
- The adjusted ElimVar measure provides a slightly increased estimate of change for MR;

Searching the solution space

- MR participants who experienced correct attitude change for $TradeUtil$ combined experimental factors in scenarios more than the incorrect grouping;
- MB participants were fairly consistent in their experimentation approach no matter the polarity of attitude change on $TradeUtil$;

Credibility and Confidence

- The self confidence in their assessment of the model was more variable for participants experiencing correct attitude change for $TradeUtil$;
- MR participants who experienced attitude change in the correct direction for $TradeUtil$ reported slightly less self-confidence in their assessment of the model than the incorrect grouping.

Chapter 6

Single-Loop Learning Comparison

6.1 Introduction

The previous chapter presented the result that the majority of participants experienced attitude change in the correct direction, given the involvement in the building and/or use of the simulation model. Attitude measurement is operationalised using the assumptions of the Theory of Planned Behaviour (see Section 4.4.1). Attitudes are positively correlated with managerial intentions: the stronger the attitude the stronger the intention to act/behave in that manner. Thus if the typical attitude change of an experimental condition is greater than that of another condition then we conclude that the first condition has had a larger effect on single-loop learning. This is the focus of this chapter: a comparison of single-loop learning across the three conditions using descriptive and inferential methods.

As a reminder, the single-loop learning comparison tests four predictions; these are summarised in Table 6.1. Predicions s.1. and s.2 concern the direction of attitude change. It is predicted that participants involved in model building will

experience larger attitude change in the correct direction relative to the model reuse participants. It is also predicted that involvement in model building will limit then extent of attitude change in the incorrect direction.

Table 6.1: Summary of Single-Loop Predictions

Hyp	**Measure(s)**	**Prediction**
s.1	Attitude change in the correct direction	MB & MBL > MR
s.2	Attitude change in the incorrect direction	MB & MBL < MR
s.3	Credibility assessment score	All three conditions rate the model similar
s.4	Self confidence in assessment of model	MB & MBL > MR

Predictions s.3 and s.4 concern the credibility rating participants give the model and the results. It is predicted that participants will rate the model equally credible, but the type of credibility assessment and thus confidence in the model will be different. It is expected that participants involved in building the model will have applied more scrutiny to the model logic, assumptions and simplifications - as part of verification and validation (V&V). Thus any assessment of credibility they provide will have evidence (memories) to back up their judgement. This is named *experienced credibility* and model builders will have high self confidence in their assessment. Conversely it is expected that model reusers will have applied less scrutiny to this part of V&V and possibly used a heuristic approach to assessment (e.g. the PhD student is an expert). These participants will have less confidence in their overall assessment of the model.

The chapter is split into four sections. The first section discusses the analysis methodology adopted and explains the use of statistics in explaining the size of differences between conditions. The second section presents a graphical analysis of differences between conditions. This is done using Quantile-Quantile (Q-Q) plots and box plots. The two types of chart give a good overview of the differences in distribution between conditions. The third section presents the inferential results.

Lastly, a discussion of the results including if and how these transfer to real world simulation studies is presented.

6.2 Analysis Considerations

Due to financial constraints, each condition in the experiment has a sample size of approximately 20 participants. A concern with this sample size is that the inferential statistics, for example, non-parametric tests, will lack sufficient power to detect a difference between the conditions where one exists. The opposite of this argument is also possible: small sample size may lead to apparent difference by chance alone; for example, simply due to the assignment of participants to groups.

To gain a better handle on these issues the analysis approach adopted makes use of three concepts. Firstly, using a standard approach from psychology, the size of the effect of the independent variable on the dependent variable is standardised. This terminology for this statistic is an *effect size* and provides information about if the effect of the independent variable on dependent is trivial (useful when assessing the practical significance of a result) or substantial (useful for non-significant results due to high variability and low sample size). An effect size is also standardised allowing comparison of measures with different scales within and across studies (section 6.2.1).

The second approach to tackling the sample size issue is to make use of computer intensive techniques for statistical inference. Here bootstrapping, i.e. resampling with replacement, is used as it is often cited as providing greater statistical power than classical parametric and non-parametric techniques for inference (Wilcox, 2005) (section 6.2.2).

Lastly, as bootstrapping may be unfamiliar to some Operational Researchers, traditional non-parametric test (Mann-Whitney) results are provided alongside the bootstrap result. Test results that disagree are highlighted and explored using effect

sizes.

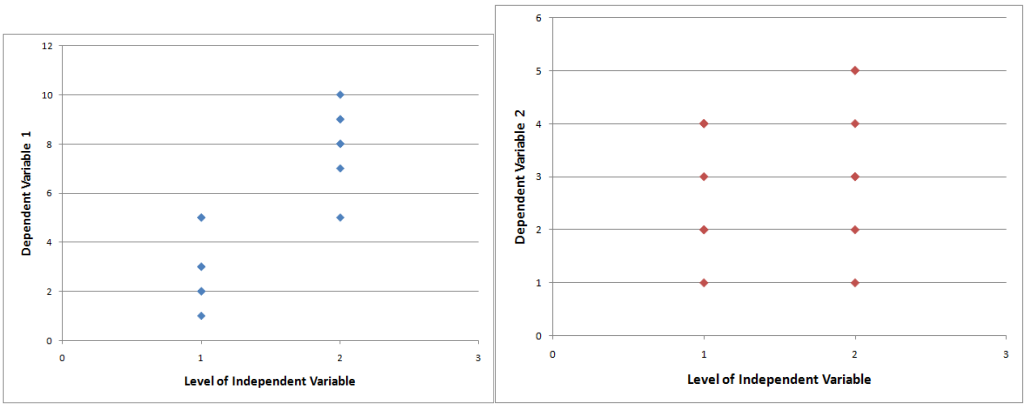
6.2.1 Standardising the Effect of an Independent Variable

Psychologists have long advocated the use of a statistic called an effect size to quantify the practical significance/influence of an independent variable on a dependent (Cohen, 1990; Field, 2009). This position has been adopted as they recognise that the result of a significance test is a function of sample size, variability and the specified probability of a type 1 error. All samples give different results, but larger samples are more likely to give statistically significant results that have little practical significance (Field, 2009). The opposite is also true: practically significant results may be statistically non-significant due to a small sample size. Thus in addition to significance results all psychologists who wish to publish their results must also quote an effect size to substantiate their claims (Field, 2009).

One approach for effect sizes is to use confidence intervals for mean differences. However, the bespoke scales and measures used in psychology research require a method that allows comparison within and across studies. This has led to the use of standardised effect sizes. Often this takes the form of the correlation between independent and dependent variables (Field, 2009).

As an example, consider Figure 6.1 illustrating two fictitious relationships between independent and dependent variables. In each chart the horizontal axis represents the two levels that the independent variable can take. In Figure 6.1a it appears that the independent variable is positively correlated to the dependent. The Pearson correlation, r, serves as the measure of standardised effect size here. ($r = .86$). Thus we would conclude that when the level of independent variable is two then there is a large positive effect on the dependent variable relative to level one. The size of this effect can then be compared to the impact on dependent variable two (Figure 6.1b) measured on a different scale. Here the standardised effect size is notably

smaller ($r = .17$), although it could also be significant given a large enough sample size. If for example, Figure 6.1b came from a study working in a similar area then the researcher(s) could interpret the effect size they found in the context of already existing results (Cohen, 1990); perhaps concluding that their independent variable has a small effect on attitude change relative to previous findings.



(a) Example Experiment 1 **(b)** Example Experment 2

Figure 6.1: Interpretting Effect Sizes

No study of learning from simulation that has been found has used effect sizes this way to date. When no relevant existing studies with effect sizes can be found, psychologists use a rule of thumb for effect sizes (Cohen, 1990). Table 6.2 provides details of these taken from Field (2009).

Table 6.2: Standard Interpretation of Effect Sizes

Effect Size	**Pearson** r
Large	$r \geq .50$
Medium	$r \geq .30$
Small	$r \geq .10$

The results of a non-parametric test are easily converted to an effect size in the

above form. Appendix B.1.1 details the procedure used.

6.2.2 Bootstrap Inference

A rationale for the bootstrap technique starts by considering how a researcher would gain access to the real world distribution of sample means for a given variable (Lunneborg, 2000). This would either be through repeated random sampling from a population or, more conveniently, given a suitable sample size, by use of the Central Limit Theorem to estimate the standard error of the mean. When asymptotic normality assumptions do not hold then the researcher does not have to repeatedly, and expensively, resample from the population, but rather from their best estimate of it - the original sample taken (Lunneborg, 2000).

A typical critique of bootstrapping states that the resampling may be taking values from a biased distribution. Hence no amount of resampling will help. This critique is usually answered in two ways by proponents of the bootstrap technique. The first argument is that if the sample is unrepresentative of the population of interest then all of the inference procedures, classical or computer intensive, are invalid. The second argument is that a large number of simulation studies have shown the bootstrap to have greater statistical power to classical parametric and non-parametric tests of inference - especially when the sample size is small (Wilcox, 2005).

This extra power is useful specifically because the analysis should also take into account that 18 comparisons are made across the conditions. If care is not taken in the analysis then the probability of incorrectly detecting that an effect is present (difference between conditions) is inflated with each comparison made (Field, 2009). In other words inference procedures must be stricter. The extra power from bootstrapping means that multiple comparison control is now practical - introducing extra rigour into the analysis. Details of the general multiple comparison problem

and procedures to correct for it can be found in Appendix B.1.4. All bootstrapped tests were subject to multiple comparison control.

It is emphasised that the bootstrap technique should not be confused with simply increasing the sample size of the distribution by repeatedly resampling with replacement. The point of the bootstrap is to construct the sample distribution of a test statistic, for example the difference between the means of two populations, so that inferences can be made about that difference. Appendix B.1.2 details the full procedure.

6.2.3 Distribution of Participant Ability

An early version of this research was published at the 2009 Winter Simulation Conference (see Monks et al., 2009). Feedback from one reviewer suggested that the relative statistical ability of each participant across the conditions should be considered. To an extent this was handled by the randomisation approach for allocation to conditions and it was it was too late to incorporate this point into the main experimental design; however, the feedback was taken seriously and a *retrospective* analysis was undertaken using exam marks for a business school module that it was believed many of the participants had taken. Participant permission to perform the analysis was also obtained retrospectively. The full analysis can be seen in Appendix B.2.

The sample size obtained for the retrospective analysis is extremely small as it was not possible to obtain suitable exam marks for all participants and all results should be considered with this limitation in mind. Nevertheless the analysis found no substantial evidence for differences in exam performance between the conditions. When combined with the randomisation approach this should give some confidence in the distribution of ability across each condition.

6.3 Graphical Results

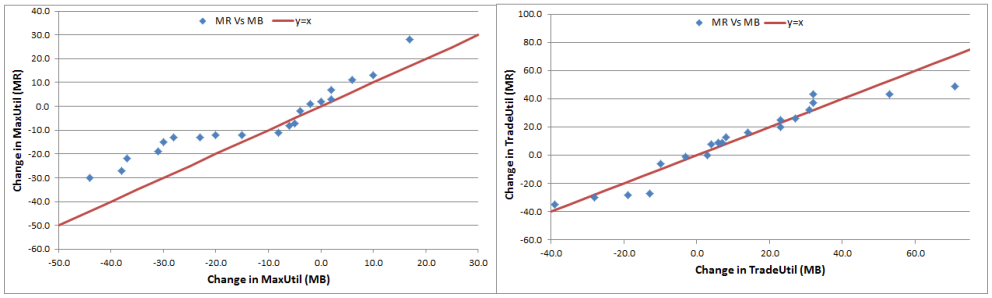
Before proceeding to the inferential results, this section provides a graphical description of potential differences between the conditions. There are three comparisons in total: MR versus MB, MR versus MBL and MB versus MB. A summary of if and how the findings support the predictions made about learning is included at the end of each comparison.

6.3.1 MR Versus MB

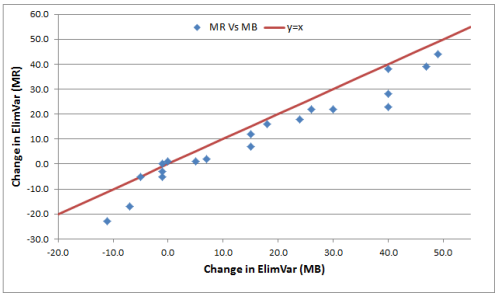
Attitude Measures

Figure 6.2 compares the distributions of attitude change for $MaxUtil$, $TradeUtil$ and $ElimVar$ (see Table 5.1 in Section 5.1.1 for definitions) across the MR and MB distributions using Quantile-Quantile (Q-Q) plots. Each chart is comprised of two data series: a bivariate plot of the quantiles of attitude change and the line $y = x$. If the distributions of MR and MB are the same then they will approximately lie on the line $y = x$. If the points mainly lie above the line $y = x$ this is an indication that quantiles of the condition on the x axis (here MB) are lower than that the quantiles of the condition on the y axis (MR). Similarly if the points largely lie below the line $y = x$ then it is an indication that the quantiles of the condition on the x axis (MB) are higher than the y axis condition (MR).

For $MaxUtil$, the line $y = x$ lies above the majority of the points. This indicates that, firstly, if attitude change occurred in the correct direction (negative) then this was somewhat greater in the MB condition. MB participants, in line with predictions (s.1), appear to be less likely than MR to push A&E resource utilisation to 100%, given exposure to the simulation model. Secondly, also in line with prediction (s.2), if attitude change in the incorrect direction occurred then this appears to be slightly larger for MR participants; although the magnitude of this difference is notably less so than the negative changes.



(a) MaxUtil **(b)** TradeUtil



(c) ElimVar

Figure 6.2: Q-Q Plots of Attitude Change

For $TradeUtil$, the points that are above zero on either axis indicate attitude change in the correct direction. These approximately lie on the line $y = x$; indicating that MR and MB gave similar results in correct learning which does not support the prediction (s.1). A similar result is found for the incorrect direction of attitude change: the plots indicates a similar distribution which again does not support the prediction (s.2).

Finally for $ElimVar$, it appears that the majority of the points lie below the line $y = x$. In fact, point could be considered in three groups. Firstly, in line with predictions, the high quantiles fall below the line; indicating that when attitudes changed in the correct direction for $ElimVar$ then this was somewhat greater for MB than MR (s.1). Secondly, the middle quantiles (clustered around zero) are close together and fall approximately on the line; indicating that a number of participants

in both conditions experienced little to no attitude change. Lastly, the lowest two quantiles are below the line; indicating that MR attitude change in the incorrect direction could be relatively strong compared to MB (s.2).

Of course, as Chapter 5 indicated, when the subgroups of correct and incorrect attitude change are analysed results could be influenced by a regression to the mean effect. In fact, this does affect interpretation of inference tests of the difference in correct attitude change in ElimVar. Section 6.4.1 discusses this in more detail.

Experimentation with new variables

In the experiment, participants in the in MB and MR conditions are given six scenarios that they must simulate. In addition they are allowed to make their own choices about new scenarios. Participants may focus on variations and combinations of the variables already outlined in the predefined scenarios. However, they may also identify other problem relevant variables for experimentation; for example, what would happen to overall performance if the prioritisation system was removed.

Figure 6.3 illustrates the distribution of the number of new variables identified in the MB and MR conditions. This shows that over 50% of the MR condition did not identify and run any experiments containing new variables. On the other hand it appears that the majority of MB participants identify and run experiment on new variables (the exact numbers of participants were 9 and 18 for MR and MB respectively).

One problem with this analysis is that in general MB participants typically simulate 3.5 scenarios more than MR ($Mdn_{MR} = 10$, $Mdn_{MB} = 13.5$, $p < .1$), as MR use some of the time to get a handle on using a simulation model; so it may just be that MB use these additional scenarios for analysing new variables. To check the comparability of the conditions Figure 6.4 illustrates the scenarios where MR and MB participants first identify a new variable for experimentation. For example,

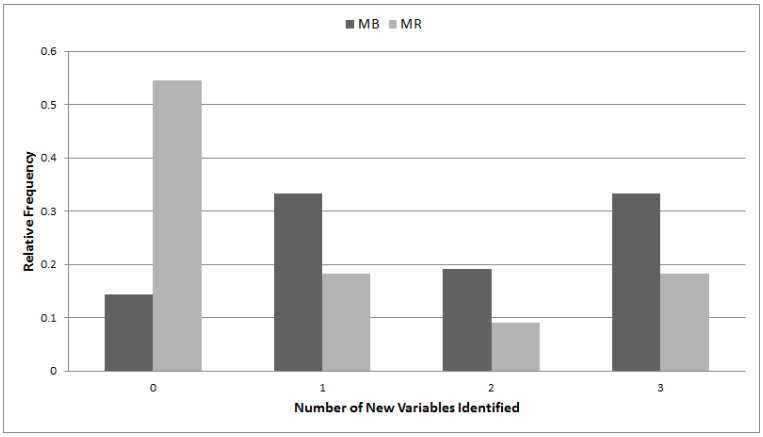


Figure 6.3: Histogram of the number of new variables identified in experimentation

it can be seen that six MB participants and four MR participants identify a new variable in the first scenario. The figure illustrates that MB participants typically identify a new variable and run a scenario containing it within the first seven scenarios, i.e. below the median number of scenarios simulated by MR ($Mdn_{MR} = 10$); so the exploration of scenarios was not biased to the extra scenarios. One explanation is that the MB participants had thought through the problem during model building and are keen to test ideas early in experimentation. An alternative explanation is that as MB participants already had a good understanding of the case study they did not feel the same time pressure as MR.

Figure 6.4 also shows that scenario one contains the highest number of new variables explored. A possible explanation for this is that the scenario runs are 'validation type' scenarios. Table 6.3 lists the new variables explored in scenario one. Three of these variables are listed as validation scenarios as they concerned the testing of assumptions and simplifications included in the model; for example, a number of participants were concerned that the nurses did not make any mistakes when triaging patients. The two remaining scenarios are labelled as 'performance' as they are believed to improve performance. One observation is that MB participants only run validation scenarios.

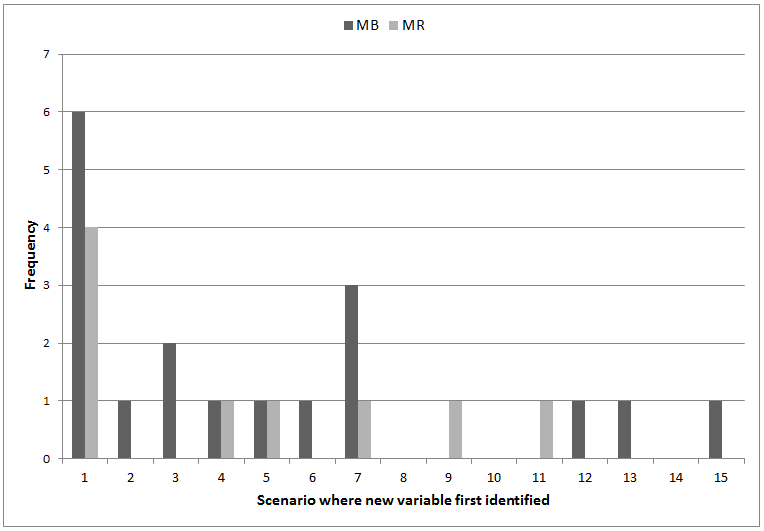


Figure 6.4: Histogram of the scenario numbers for first new variable. The x-axis represents the scenario number where participants first identify a new variable. A new variable is defined as a variable related the case study problem, but different from the pre-defined scenarios given to participants. For example, six MB participants identified a new variable in the first scenario.

Table 6.3: New variables identified in first scenario.

		Number of Scenarios		
Variable	**Type**	**MR**	**MB**	**Total**
Multi-tasking*.	Validation	1	3	4
Alternate shift patterns	Performance	1	0	1
Misdiagnosis of patients in triage	Validation	1	1	2
Removal of prioritisation system	Validation	0	2	2
Faster JD doctors	Performance	1	0	1

*E.g. major patients take more than one doctor slot

Table 6.3 explores the differences further and lists all of the scenarios run by participants in the experiment along with the total number of participants that ran them. One difference that stands out from the rest is the number of participants that ran scenarios concerning the multi-tasking of resources in the model (i.e. doctors and nurses treating multiple patients at once. See Appendix A.2.2 for a description of implementation in the model; alternatively see Günal and Pidd (2006)). As discussed this was largely a validation exercise (participants were sometimes concerned about major patients taking only one slot or mini-doctor), but participants also looked at what would happen if, for example, doctors could cope with more patients.

Table 6.4: New Variables and Use in Experimentation

	No. of Scenarios		No. Participants	
Variable	**MR**	**MB**	**MR**	**MB**
Multi-tasking	1	11	1	10
Alternate shift patterns	7	8	5	7
Misdiagnosis of patients in triage	1	2	1	1
Removal of prioritisation system	0	5	0	4
Treatment times	2	4	1	3
Dedicated Resources	0	4	0	3
X-ray rooms resources	3	3	3	3
Extra evaluation room resources	0	1	0	1
Pooling resources	0	1	0	1
Registration Desks	0	1	0	1
Better x-ray decision making	3	0	3	0
Cubicle sharing	1	0	1	0
Less patients brought to hosptial	2	0	1	0

There are other interesting differences in Table 6.3; although it must be noted that these are much smaller than the multi-tasking difference. One such difference is that no MR participants explored the prioritisation system while four MB par-

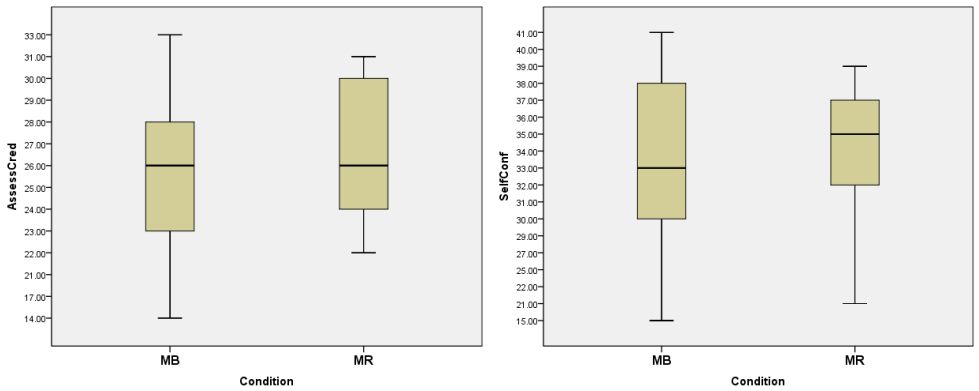
ticipants did. Although this number is small, and hence any interpretation should be taken with caution, it is anecdotally noted that these MB participants simulated these scenarios in response to noticing changes of model behaviour and results during model building. Furthermore, a number of other MB participants noticed the same behaviour (MB21 is one such example), but were unable to pinpoint the prioritisation system as possibly cause of the behaviour. Section 7.1.2 discusses this point in more detail along with a possible discovery and novelty mechanism in the experiment.

In summary an explanation for these results would seem to be that MB participants have built up a number of concerns about the validity of the model as well as an idea about how to improve performance during model building and wish to test them. MR participants may need to use the model to learn more about it and the problem before they can think of new ideas.

Credibility Measures

Figure 6.5 compares the distributions of $CredAssess$ and $SelfConf$ accross the MR and MB conditions using box plots. The top and bottom of the shaded box represent the upper and lower quartiles of the distributions respectively, while the middle line represents the median. The lengths of the lines (whiskers) extending from the shaded box are either one and a half times the height of the box or if no values are present in that range to the maximum and minimum values in the distribution. Any dots that appear below or above the whiskers are univariate outliers (there are none in this case).

For $CredAssess$, the median value of the distributions, in line with expectations, are practically the same ($mdn_{MR} = 26.0$, $mdn_{MB} = 25.5$). The maximum rating available is 35; so typically MR and MB participants rated the model and results as quite credible. The variability of results is also similar ($IQR_{MR} = 5.8$, $IQR_{MB} =$



(a) CredAssess **(b)** SelfConf

Figure 6.5: Box Plots of Credibility Measures

5.0). The MB condition also has wider tails for the credibility rating: the odd participant was either very convinced or very sceptical.

For $SelfConf$, it was expected that the extra involvement in model development of MB participants would increase their confidence in the assessment they made of model credibility. However, Figure 6.5b indicates that MR participants typically had similar self confidence in their rating to the MB participants ($mdn_{MR} = 34.5$, $mdn_{MB} = 33.0$). The maximum self confidence rating possible was 45; indicating that both conditions typically had high confidence in their assessment.

There is some difference in the variability of the conditions. Figure 6.5b indicates that MR participants were slightly less variable in their report of self confidence ($IQR = 5.0$) than MB participants ($IQR = 8.0$); although, the bottom 25% of MR self confidence scores is quite variable. Indeed, some MB participants reported low confidence in their assessment (ratings of < 20 out of 45).

Summary

It was predicted that involving the participants in building the simulation model (MB) would allow them to learn more about the system than if they were excluded

and reused the model (MR). Given an important aspect of managing the A&E queuing system over the six month time horizon - resource utilisation and process variation - it was predicted that MB participants would experience greater attitude change that was beneficial for system performance. If attitude change occurred that was non-beneficial it was predicted that this would be less severe in MB participants than MR.

Table 6.5 summarises the initial graphical analysis of the three attitudes undertaken in this section. The analysis indicates support of the correct direction hypothesis in two cases: beneficial attitude change appears to be greater for $MaxUtil$ and $ElimVar$ (although note that ElimVar may be subject to a regression to the mean effect). MB participants appear to have been more likely than MR to stop their policy of 100% resource utilisation and to seek to eliminate variability in the availability of radiology resources. However, the analysis only indicates support for the incorrect direction hypothesis in one case: incorrect attitude change appears to be slightly less for $MaxUtil$. Out of those participants who experienced attitude change in the incorrect direction MR participants were the most likely to act on their intentions and push resource utilisation to its maximum.

Table 6.5: Summary of MR versus MB attitude change hypothesis support

	Support For Predictions	
	Correct Direction	**Incorrect Direction**
MaxUtil	$MR < MB$	Weak Evidence $MR > MB$
TradeUtil	No Difference	No Difference
ElimVar	$MB > MR$	No Difference

The second set of predictions made concern the credibility rating participants hold for the model and the results. Based on pilot observations it was predicted that participants would rate the model equally credible, but the type of credibility held would be different. Participants involved in building the model (MB) will

report *experienced credibility* and will have high self confidence in their assessment of credibility. Participants who were not involved in the building of the model will perform more of a heuristic assessment of the model and be relatively less confident in their own assessment of credibility.

Table 6.6: Summary of MR versus MB credibility hypothesis support

	Support For Predictions
CredAssess	Prediction Supported - Evidence to suggest that both conditions typically accepted the model and results
SelfConf	Prediction not supported. Some evidence to suggest that MR was more consistent in outcome.

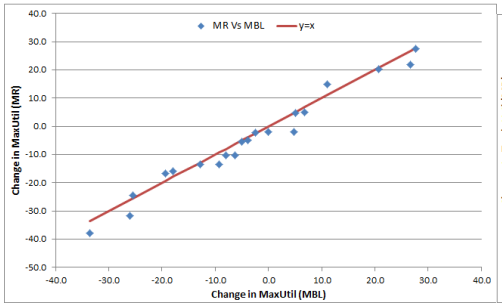
Table 6.6 summarises the initial graphical analysis of the two credibility measures. The prediction that MB and MR participants would rate the model as equally credible is supported. However, the predicted difference in self confidence is not supported. Some weak evidence exists that the MR participants were slightly less variable (i.e. a lower IQR) in their self confidence than MB.

6.3.2 MR Versus MBL

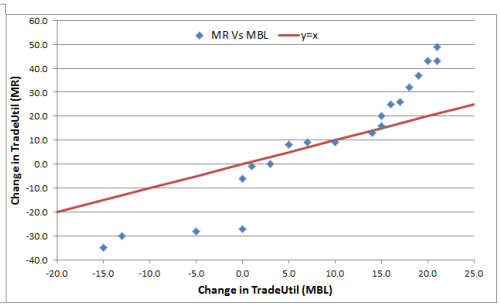
Attitude Measures

Figure 6.6 compares the distributions of attitude change for $MaxUtil$, $TradeUtil$ and $ElimVar$ across the MR and MBL distributions using Quantile-Quantile (Q-Q) plots.

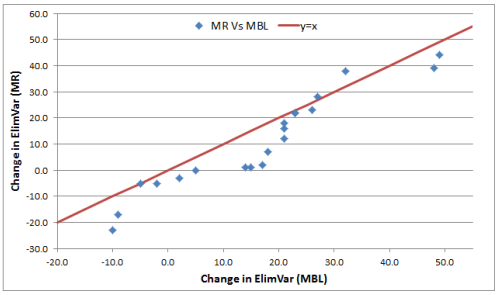
For $MaxUtil$, in Figure 6.6a, the majority of the points lie on the line $y = x$. Thus counter to predictions (s.1. and s.2) attitude change in the incorrect direction (positive) and the correct direction (positive) was similar across MR and MBL participants. The similar distributions indicate that both conditions had a similar effect on participants' attitude towards pushing A&E resource utilisation to



(a) MaxUtil



(b) TradeUtil



(c) ElimVar

Figure 6.6: Q-Q Plots of Attitude Change

its maximum.

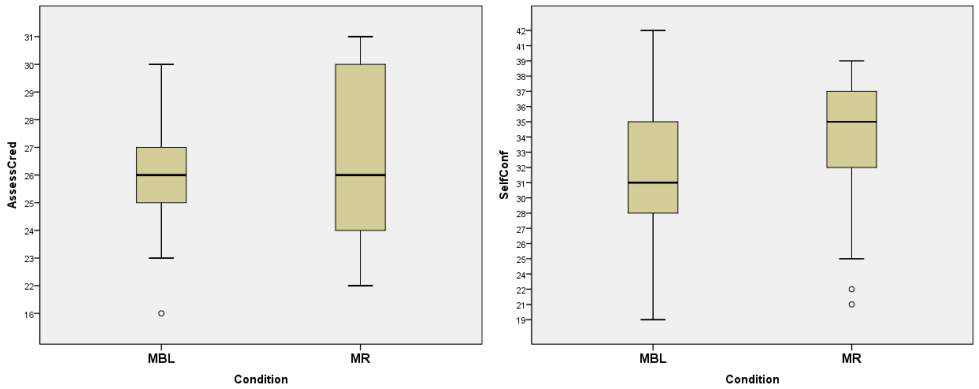
For $TradeUtil$, Figure 6.6b illustrates that the quantiles form an approximate s-shape around the line. Counter to predictions (s.1), if attitude change occurred in the correct direction (positive) then this appears to be strongest in the MR condition. MR participants appear the most likely to form and act on the intention to trade off resource utilisation in order to achieve higher system performance. However, of those participants who experienced attitude change in the incorrect direction, Figure 6.6b indicates that, in line with predictions (s.2), MR also experience the strongest change. MBL then appears to have produced a much more consistent outcome that MR: a maximum range of -15 to +20 for attitude change MBL compared to -35 to 50 in MR.

For $ElimVar$, Figure 6.6c illustrates that there is a particular difference between

the low positive quantiles of the MR and MBL distributions ranging from 0 to +25 on the x axis (i.e. the lower quantiles of MBL are larger than MR); however, there is little difference in the upper positive quantiles. This may suggest that MBL participants are more likely to learn 'something' about or notice the importance of ElimVar during model building than MR. However, when MR participants do notice the importance of ElimVar then correct attitude change is very similar. Of course, this result should again be considered with the impact of regression to the mean in the subgroup. Section 6.4.2 discusses this in more detail.

Credibility Measures

Figure 6.7 compares the distributions of $CredAssess$ and $SelfConf$ accross the MR and MBL conditions using box plots. In this case there is one univariate outlier in MBL that rated the credibility of the model quite low and two univariate outliers in MR that reported low self confidence in their assessment of the model.



(a) AssessCred **(b)** SelfConf

Figure 6.7: Box Plots of Credibility Measures

For $CredAssess$, the median value of the distributions, in line with expectations, are the same ($mdn_{MR} = 26.0$, $mdn_{MBL} = 26.0$). The maximum rating available is 35; so typically MR and MBL participants rated the model and results

as quite credible. The variability of results is quite different though ($IQR_{MBL} = 2.5$, $IQR_{MR} = 5.8$): MBL produced the most consistent rating.

For $SelfConf$, it was expected that the extra involvement in model development of MBL participants would increase their confidence in the assessment they made of model credibility. However, Figure 6.7b provides some weak evidence that the reverse occurred. MR participants typically report slightly higher self confidence in their rating than the MBL participants ($mdn_{MR} = 34.5$, $mdn_{MBL} = 31.0$). Although, as the maximum self confidence rating was 45, both conditions typically had high confidence in their assessment.

Again there is some difference in the variability of the conditions. Figure 6.7b indicates that MR participants were slightly less variable in their report of self confidence ($IQR = 5.0$) than MBL participants ($IQR = 8.0$). Similar to MB participants, some MBL participants reported low confidence in their assessment (ratings of < 20 out of 45).

Summary

Table 6.7 summarises the initial graphical analysis of the three attitudes undertaken in this section. The analysis only supports the hypothesis about the correct direction of attitude change (s.1) in one case: as found in the comparison of MR to MB the lower quantiles of the correct direction attitude change appears to be greater for $ElimVar$. MBL participants appear to have been more likely to learn 'something' about the importance of eliminating variability in usage of radiology than MR.

The analysis only supports the prediction about the incorrect direction of attitude change in one case: attitude change in the incorrect direction appears to be less for $TradeUtil$. Out of those participants who experienced the incorrect direction of change MR participants were the least likely to trade off resource utilisation to improve performance.

Table 6.7: Summary of MR versus MBL attitude change hypothesis support

	Support For Predictions	
	Correct Direction	**Incorrect**
MaxUtil	No Difference	No Difference
TradeUtil	$MR > MBL$	$MBL < MR$
ElimVar	Weak Evidence $MR < MBL$	No Difference

Table 6.8 summarises the initial graphical analysis of the two credibility measures. The prediction that MBL and MR participants would rate the model as equally credible is supported. However, the predicted difference in self confidence is not supported. Some weak evidence exists that the MR participants held slightly more self confidence than MBL.

Table 6.8: Summary of MR versus MBL credibility hypothesis support

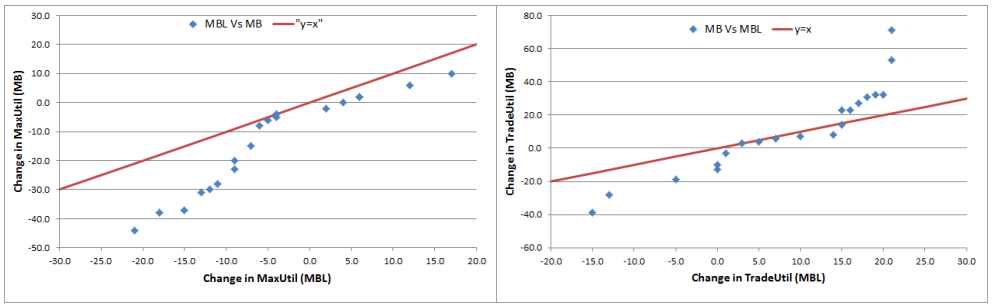
	Support For Predictions
CredAssess	Prediction supported - Evidence to suggest that both conditions typically accepted the model and results. MBL produced most consistent rating;
SelfConf	Not supported. Some weak evidence to suggest that MR participants had greater self confidence.

6.3.3 MB Versus MBL

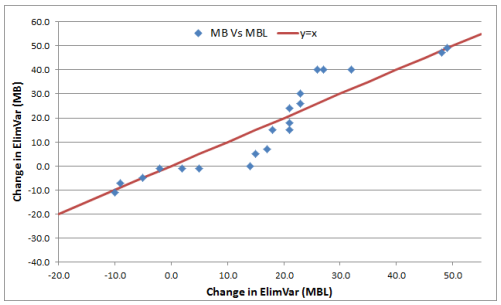
Attitude Measures

Figure 6.8 compares the distributions of attitude change for $MaxUtil$, $TradeUtil$ and $ElimVar$ across the MB and MBL distributions using Quantile-Quantile (Q-Q) plots.

For $MaxUtil$, in Figure 6.8a, the majority of the points lie below the line $y = x$. This indicates that the MBL quantiles are typically higher than the MB quantiles.



(a) MaxUtil **(b)** TradeUtil



(c) ElimVar

Figure 6.8: Q-Q Plots of Attitude Change

This indicates two differences. The first difference is that MBL participants experience slightly less attitude change in the correct direction (negative) of $MaxUtil$; hence are less likely to drop the intention to push A&E resource utilisation to its maximum. The second difference is that if attitude change in the correct direction (positive) occurs then MBL participants are most likely to push A&E resource utilisation to its maximum.

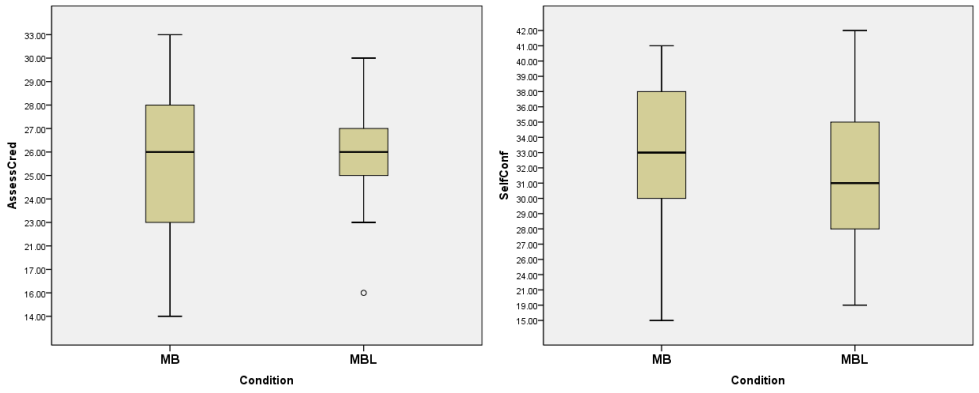
For $TradeUtil$, Figure 6.8b illustrates that the quantiles are below the line when they are negative in value and above the line when they are positive in value. This indicates a similar result as found in the comparison of MR and MBL. When attitude change is in the correct direction (positive) MB participants are more likely to intend to use the trade-off relationship between resource utilisation and performance than MBL participants. When attitude change is negative MBL participants are the most

likely to drop intentions to use the relationship.

For $ElimVar$, Figure 6.8c illustrates that the quantiles are mainly positive in value; indicating that both conditions generally favour the correct direction of attitude change. The positive changes can be split into two groups. The lower positive quantiles are greater in MBL and the higher positive quantiles are larger in MB. This result could be due to attitude change for $ElimVar$ being inherently variable (descriptive results from Chapter 5.3 indicated it is the most variable measure). The two distributions may be very similar, but results are affected by the small sample size. An alternative explanation is that model building aids the discovery of information about ElimVar, hence attitudes are shifted in the correct direction, and experimentation further aids attitude refinement and reinforcement. Section 7.1.2 discusses this possible mechanism for learning further.

Credibility Measures

Figure 6.9 compares the distributions of $CredAssess$ and $SelfConf$ accross the MB and MBL conditions using box plots.



(a) AssessCred **(b)** SelfConf

Figure 6.9: Box Plots of Credibility Measures

For $CredAssess$, the median value of the distributions, in line with expectations,

are practically the same ($mdn_{MR} = 26.0$, $mdn_{MBL} = 26.0$). The maximum rating available is 35; so typically MB and MBL participants rated the model and results as quite credible. The variability of results is quite different though ($IQR_{MBL} = 2.5$, $IQR_{MB} = 5.0$): MBL produced the most consistent rating.

For $SelfConf$, Figure 6.9b illustrates that MBL credibility ratings are slightly less than MB ($mdn_{MR} = 33.0$, $mdn_{MBL} = 31.0$), although this small difference in the medians may be due to random variation in the ratings.

There is some difference in the variability of the conditions. The longer bottom whisker of MB in Figure 6.9b indicates negative skew (only two out of 20 participants rated their self confidence below 29) whilst the MBL distribution is fairly symmetrical. This possibly indicates that MB was more consistent in outcome than MBL. Although note that both inter-quartile ranges are equal ($IQR = 8$).

Summary

Table 6.9 summarises the initial graphical analysis of the three attitudes undertaken in this section. Although there are no specific predictions, differences between MBL and MB may indicate areas where experimentation is important for learning; especially if the MB result agrees with MR.

Table 6.9: Summary of MB versus MBL attitude change hypothesis support

	Differences	
	Correct Direction	**Incorrect**
MaxUtil	$MB > MBL$	Weak evidence $MB < MBL$
TradeUtil	$MB > MBL$	$MB > MBL$
ElimVar	No Difference	No Difference

Table 6.10 summarises the initial graphical analysis of the two credibility measures. Like the other comparisons MB and MBL both typically rate the model as credible. However, the descriptive statistics indicate that MBL participants are the

most consistent in their rating. Boxplots of $SelfConf$ indicated that the median for MBL may be slightly lower than MB; however, this could be due to random variation in the data and requires further analysis.

Table 6.10: Summary of MB versus MBL credibility differences

	Results of Graphical Comparison
CredAssess	Evidence to suggest that both conditions typically accepted the model and results. MBL produced most consistent rating;
SelfConf	Inconclusive. Some weak evidence to suggest differences in variability.

6.4 Inferential Results

This section builds on the graphical analysis by providing formal inferential results for the predictions made about learning (see Table 6.1 in the introduction of this chapter or Section 4.3.5). The information presented here is in summary form only; detailing support, lack of support or contradictions of predictions given by the bootstrap and non-parametric tests. Where disagreement exists between the inference tests, typically in the form of the non-parametric test failing to find a difference where the bootstrap does, discussion of the graphical results and effect sizes are used in an attempt to explain. One issue with the analysis may be the masking or introduction of effects due to regression to the mean (see section 5.2.2, for an explanation). The subgroups for $ElimVar$ are particularly suspect, as section 5.2.2 indicated an imbalance in the pre-test scores and section 5.3 indicated some difference to typical MB and MBL $ElimVar$ change scores after adjustment. To assess these problems additional analyses were undertaken on the adjusted scores for all measures; differences in interpretation from the raw scores are highlighted and discussed using Q-Q plots and effect sizes. A full comparison of the inferential results for the raw and adjusted scores is presented in Appendix B.3.

Given the descriptive results, of the previous section, show that the median credibility assessment score made by participants is equal across the three conditions, no formal inference tests of $CredAssess$ are undertaken. However, the descriptive results did indicate MBL participants may have lower self confidence in their assessment than MR or MB participants. These test results are reported along issues caused by some influential cases.

6.4.1 MR versus MB

Attitude Measures

As a reminder, it is predicted that involvement in model building will influence participant attitude change in the correct direction more than if the client was excluded from development. It is also predicted that if attitude change is in the incorrect direction - resulting in intentions that will reduce system performance - then this should be less extreme for participants involved in model building.

Table 6.11: Summary of Inferential Results for MR versus MB attitude change

	Correct Direction		**Incorrect Direction**	
	Conclusion	Effect Size (r)	Conclusion	Effect Size (r)
MaxUtil	$MB > MR^{*}$	.20	-	.07
TradeUtil	-	.14	-	.07
ElimVar	$MB > MR^{*}$	.28	$MB < MR^{**}$	.39

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key: Agreement | Disagreement

* $p < .1$, but no longer significant after multiple comparison control. FDR cut-off = .011

** $p < .05$, but no longer significant after multiple comparison control

Table 6.11 summarises the results of the non-parametric and bootstrap tests. There are three cases where significant results are found; only in one case do the non-parametric and bootstrap analyses agree. For correct change in $ElimVar$, the

bootstrap concludes that there is evidence to suggest the attitude change is larger in MB ($r = .21$). This difference between conditions was to some extent expected, given the graphical analysis in Section 6.3.1.

The bootstrap also concludes that the difference in the incorrect change of $ElimVar$ is not due to chance alone. The non-parametric result cannot find enough evidence to provide such a conclusion, but suggests that a medium effect size is present ($r = .39$). Thus differences in interpretation may be largely due to the increased power of the bootstrap.

There is also a disagreement between the bootstrap and non-parametric tests about the difference in the correct change of $MaxUtil$. The bootstrap concludes that a difference exists in the correct change of $MaxUtil$ (agreeing with the graphical analysis). However, the non-parametric test cannot conclude there is a difference, but suggests a medium effect is present ($r = .28$). Thus differences in interpretation may be largely due to the increased power of the bootstrap.

Table 6.11 also illustrates that if multiple comparison control is applied to the results then none of the results are significant. Hence if a strict statistical position is taken, then it cannot be concluded that any differences between attitude change in MB and MR is due to anything, but random variation.

Differences from Adjusted Results

In addition to the results outlined above the inference tests were run for the ElimVar subgroup results adjusted for regression to the mean. After this adjustment only the incorrect change in ElimVar remained 'significant' (both non-parametric and bootstrap tests agreed that a large effect was present, $r = .49$). The correct direction of attitude change now only reported an extremely small effect size ($r = .02$). Thus there may have been very little difference between the correct directions of attitude change in the conditions. Appendix B.3 discusses this in more detail.

The other difference to note is that the difference in $MaxUtil$ in the correct subgroup is no longer significant after adjustment. This result is not too worrying as inference tests report a similar effect size ($r = .24$) and Q-Q plots also indicate a difference in the distributions of the conditions. A prudent conclusion is therefore that there is some effect on attitude change between the conditions, but that it is only small. Given only the single-loop results, it is difficult to determine if such a small difference in attitude is important. One consequence might be that the extra change in attitude produced by MB makes the difference in implementing an action or not. Alternatively such a small difference may have no practical impact at all. Furthermore, the slight difference in attitude change may be an indicator of double-loop learning, i.e. the MB condition may understand the relationship between resource utilisation and performance to a greater extent than MR. The double-loop results in chapters 8 and 9 explore this possibility further.

6.4.2 MR versus MBL

Attitude Measures

The same predictions are made for the MBL comparison to MR as made for MB participants. Attitude change in the correct direction will be greater than MR and when incorrect attitude change occurs this will be substantially less in MBL than MR.

Table 6.12 summarises the results of the non-parametric and bootstrap tests. There is one case where a significant result is found: for the correct direction of change in $TradeUtil$; although the non-parametric and bootstrap analyses disagree about this result. The non-parametric result, although non-significant, indicates a medium effect size ($r = .40$) that disagrees with prediction (s.1). The graphical results in Section 6.3.2 also indicated that a difference existed. Thus the disagreement of the inference tests is likely to be owing to the increased statistical power of the

bootstrap test.

Table 6.12: Summary of Inferential Results for MR versus MBL attitude change

	Correct Direction		**Incorrect Direction**	
	Conclusion	Effect Size (r)	Conclusion	Effect Size (r)
MaxUtil	-	.22	-	.14
TradeUtil	$MR > MBL$	.40	-	.23
ElimVar	-	.11	-	.10

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key: Agreement | Disagreement

Multiple comparison control: FDR cut-off = .011

The graphical analyses also indicated a difference in the incorrect direction of attitude change for $TradeUtil$. This difference is not supported by significance tests, however, a small to medium effect size is suggested ($r = .23$). Although no statistical difference can be supported, it is interesting to note that even when participants were involved in MBL their attitude change was less extreme in either direction (a similar result is was found for MB versus MBL). An explanation for this difference in extremity may be related to the participants approach to experimentation and choice of scenarios. Sections 7.2 and 7.1.1 discusses mechanisms related to this point in more detail.

The result for correct change in $ElimVar$ is expected as the graphical analysis (Section 6.3.2) suggested that a difference may be present in the lower quantiles of the distributions (the upper quantiles are similar). A speculative mechanism for learning in the experiment may then be that model building condition somehow draws attention to the object of the ElimVar attitude and that this is less obvious in the MR condition. Section 7.1.2 discusses some observations from the experiment in relation to participants discovering novel factors influencing performance in the A&E department and how this might affect attitude change; for example,

seeing performance of the model become closer to reality when the level of detail is increased. Section 8.2.3 takes this further by discussing how the order of choices participants make in the model building, or the obviousness of choices, may influence their reflection on the problem and transfer success.

Differences from Adjusted Data

In addition to the results outlined above the tests were also run for the adjusted $ElimVar$ data. Both tests agreed with the original data with similar effect sizes (correct change $r = .11$ and incorrect change $r = .21$). The Q-Q plot of the adjusted data also suggested that the lower quantiles of the distribution were higher in MBL than MR; however, it should be noted this difference is less so in than in the original data.

In addition to the significant difference found in the original data the adjusted analyses concludes that two additional differences exist. The first conclusion is that MR resulted in more attitude change in the correct direction for $MaxUtil$. Furthermore, the data supports that this was a medium effect size ($r = .31$). This agrees with the graphical analysis of the difference in section 6.3.2. The second conclusion is that MR resulted in more incorrect attitude change than MBL for $TradeUtil$ ($r = .21$). This again agrees with the indications of the graphical analyses in section 6.3.2. Appendix B.3 discusses both points in more detail.

Lastly, it should be pointed out that when multiple comparison control is applied to the tests the only result that remains significant is the conclusion that MR has more influence than MBL on the correct change of $TradeUtil$. Thus if a strict statistical view of the data is taken the adjusted and original results are the same.

Credibility Measures

It was predicted that participants involved in model building would gain more self confidence in their assessment of the simulation model. The graphical analysis indicated that, counter to this prediction, MBL participants held slightly less self confidence in their assessment of the simulation model than MR (and possibly MB).

Two influential cases, MR6 and MR19, affect the result of any test of significance. These cases can be seen as dots in Figure 6.7b in Section 6.3.2. If these are excluded both the non-parametric and bootstrap (for difference in median) tests are significant ($p < .1$) with a medium effect size ($r = .30$). When they are included the effect size reported is only small ($r = .20$); indicating that only a weak difference exists between the typical self confidence in the conditions.

6.4.3 MB versus MBL

Attitude Measures

No specific predictions were made regarding differences between the MB and MBL conditions. However, one plausible argument is that if the participants are changing their attitudes during model development similar learning outcomes might be expected.

Table 6.13: Summary of Inferential Results for MB versus MBL attitude change

	Correct Direction		**Incorrect Direction**	
	Conclusion	Effect Size (r)	Conclusion	Effect Size (r)
MaxUtil	-	.27	-	.19
TradeUtil	MB > MBL	.28	-	.21
ElimVar	-	.16	-	.42

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key: Agreement | Disagreement

Multiple comparison control: FDR cut-off = .011

Table 6.13 summarises the results of the non-parametric and bootstrap tests for MB versus MBL. There is one case where a significant result is found. The non-parametric and bootstrap analyses agree that there is a difference in the correct direction of change in $TradeUtil$ ($r = .28$). That is, MB participants are more likely to act on their intentions to trade off A&E resource utilisation in order to improve service level. A similar conclusion was reached in the MR versus MBL comparison; however, the effect size was slightly higher ($r = .40$).

The graphical analysis of Section 6.3.3 indicated that the there was also a difference in the incorrect direction of attitude change for $TradeUtil$. However, like the MR versus MBL comparison the results are not statistically significant. One reason why this conclusion cannot be reached is that no predictions are made about the differences between the conditions; hence a less powerful two-sided test is conducted. However, the results do suggest that a small effect is present in the data - estimated to be slightly smaller than the result for the correct direction. This would tend to agree with the graphical analysis. Thus again an interesting point about the results is that the longer model building condition could, in a similar manner to MR, produce substantial attitude change in the incorrect direction. A possible explanation for this may be related to the approach to experimentation in the case study. In particular the avoidance of scenarios that disconfirmed knowledge was observed. Section 7.2.1 discusses this in more detail.

There are two further points worthy of mention. Firstly, the graphical analysis indicated a difference in the correct direction of change in $MaxUtil$. Although this is non-significant, the data suggest a medium effect size is present. Secondly, the data also suggest that a medium effect size is present in the incorrect direction of change for $ElimVar$, but that there is insufficient power to detect it.

Differences from Adjusted Data

The inference results for the adjusted ElimVar subgroups are also non-significant. In fact, the estimated effect sizes and Q-Q plot are similar to the original data. This is unsurprising as the mean differences between pre-test scores of the MB and MBL subgroups is relatively small compared to the differences from MR.

An important difference to note is that the adjusted data only indicates a small difference in the correct change of $TradeUtil$ (and is no longer significant). This result is interesting as the results show (even when multiple comparison control is applied) that MR affected the correct change in $TradeUtil$ more than MBL.

Credibility Measures

The graphical analysis indicated that, counter to predictions, MBL participants held slightly less self confidence in their assessment of the simulation model than MB; although, this appears to be less so than between MR and MBL . Both the non-parametric and bootstrap (of a difference in median) found no evidence of a difference ($p > .1$, $r = .15$).

6.5 Summary

Table 6.14 provides a summary of the findings that supported the learning and credibility hypotheses as well as other important findings. The following chapter discusses possible explanations for these results in terms of learning mechanisms and considers if the findings transfer to the real world.

Table 6.14: Summary of Single-Loop Findings

Hyp	Measure(s)	Prediction	Findings
s.1	Correct direction	MB, MBL > MR	MB > MR for $ElimVar$; MBL > MR but only for lower quantiles of $ElimVar$; $ElimVar$ difference between MB and MR may be due to a RTME; Counter to prediction $MR > MBL$ for $TradeUtil$; $TradeUtil$ difference between MB and MBL may be affected by a RTME;
s.2	Incorrect direction	MB,MBL < MR	Graphical evidence that MBL < MR,MB for $TradeUtil$
s.3	Credibility score	MB = MBL = MR	MB = MBL = MR Weak, graphical evidence, that MBL has the least variation in credibility assesmment score & MR has the most.
s.4	Self confidence in assessment of model	MB, MBL > MR	Counter to prediction weak evidence MR > MBL
-	Experimentation	-	MB thought of more new variables than MR; Limited evidence that MB used early scenarios for validation; Limited evidence that MB discovers more novel aspects of the problem during model building and investigates in experimentation;

Notes: RTME = regression to the mean effect.

Chapter 7

Discussion of Single-Loop Results

The results reported in the previous two chapters provide some support to the model building learning hypotheses. However, there are also notable contradictions to the predictions for learning. This chapter, firstly, discusses the support for and contradictions of each hypothesis in turn and speculates on the learning mechanisms in play. These results should be interpreted carefully as the experiment is a simplification of a real simulation study. Thus this section also looks at the differences between the experiment and real simulation studies and how factors such as the timing of attitude measurement and novice versus expert knowledge affect applicability of results.

7.1 Hypothesis 1: Correct Direction of Attitude Change

The largest difference between conditions found in the experiment contradicts the hypothesis that decision makers learn more from involvement in model building. In particular it appears that MBL participants experience less attitude change, in the correct direction, than MR and MB on the *TradeUtil* measure. There was also

some graphical evidence that MBL participants experienced less attitude change on $MaxUtil$. In fact, the findings suggest that more rather than less experimentation aided participants' learning about managing resource utilisation in the A&E department case study. An explanation for this may lie in the strength of the initial attitude and the complexity of the concept to be learnt; section 7.1.1 discusses this mechanism.

The results for the final attitude measure ($ElimVar$) are more difficult to interpret, due to contradictions in results and the possibility of a regression to the mean effect. Analyses of the data suggest two differences between the model building conditions and the model reuse condition. Firstly, the data support the hypothesis that MB undergoes more attitude change than MR; although the result was no longer significant after multiple comparison control is applied to the results, graphical analyses also supported this view. Secondly, although inference tests do not support significant differences between MBL and MR, graphical analysis of the distributions illustrate that the lower quantiles of indicate more attitude change in MBL. As pre-test attitudes for $ElimVar$ were neutral - indicating that the factor was outside of the participants mental models, an explanation may be related to the discovery and novelty of the factor during model building. In fact, a number of discovery moments were frequently observed during model building.

Some caution should be taken with this interpretation of the $ElimVar$ results, however, as analyses also show that regression to the mean may be particularly problematic with the $ElimVar$ measure. In fact, the adjusted results suggest that attitude change was similar in MB and MR. Therefore, instead of concluding which condition learns 'more', section 7.1.2 discusses the behaviour of some participants in the experiment and how this may relate to the novelty and discovery mechanism discussed above.

7.1.1 Mechanism: Complexity

The first two attitude measures in the experiment relate to the management of A&E resource utilisation. Participants held strong opinions on this topic (pre-test measures $MaxUtil$ mean = 30; $TradeUtil$ mean = 22), although they fail to recognise a relationship between the two.

This result agrees with the survey work of Suri (1998): out of a sample of 450 manufacturing managers and CEOs, 75% of participants believed that 100% utilisation of assembly line machines led to fast and flexible production of their goods. Suri's result indicates the relationship is a complex one as domain experts can also fail to grasp it.

The utilisation attitude measures provide little support for the prediction that involvement in model building increases attitude change in the correct direction. There is some evidence that MB participants are more likely to drop their intention to push all resource utilisation to 100%; however, this was only a small effect when compared to other differences found in the experiment. Additionally MBL participants undergo less learning in the correct direction than either MB or MR on the $TradeUtil$ measure.

One mechanism may be the complexity of the concept/relationship to be understood; two pieces of evidence point in this direction. Firstly, MB participants spend an extra hour on the case study problem. This extra time may allow participants, even those with lower ability, to have the opportunity to think twice about the effect of pushing for 100% utilisation. Thus we see a slight increase in learning on $MaxUtil$ for the MB condition.

Secondly, given that correct attitude change on $TradeUtil$ appears limited under the lower experimentation time of MBL, it would seem that empirical understanding of a utilisation/performance relationship benefits from experimentation. Two reasons may explain this result. Firstly, understanding some complex concepts may

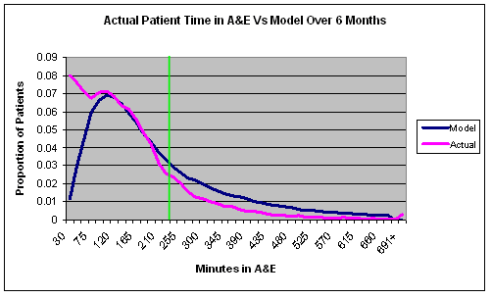
benefit from a classic test, experience, reflect and reformulate learning cycle (Kolb, 1984); resource utilisation may fit into this category. Secondly, the participant's involvement in model building may only offer a hint of this relationship and change attitude only slightly. For example, in a debrief after the experiment, participant MB9 stated that they had noticed during model building that maximum utilisation was not working and that it appeared 'surprisingly' to be a better option to obtain more resource. MBL participants had very little opportunity to test this idea.

7.1.2 Mechanism: Discovery and Novelty

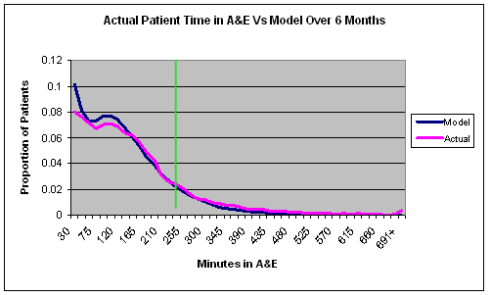
All participants held a fairly neutral $ElimVar$ attitude pre-test (mean = 13.9). This was expected given the experience with the pilot participants. A plausible explanation is that participants have little experience in thinking of the world in terms of variability. Some limited evidence of this can be seen in the order of choices participants make when involved in building the model. For example, the majority of participants initially opt to model physical aspects of the system in more detail (e.g. doctors and treatment categories) as opposed to the inter-arrival profile of patients. As there is a possibility of regression to the mean affecting the $ElimVar$ results, for the moment let us assume that each condition effected $ElimVar$ to a similar extent; however, the mechanism for that learning may be different across the conditions. In particular, it is noted that MB and MBL participants may discover the importance of radiology to system performance during model building, while MR (and also MB) learn about the radiology from experimentation.

As an example, consider that, during model building participants review a number of supplied criteria when assessing if the current version of the model is fit for purpose. Although no direct measurement was taken, it is noted that participants paid particular attention to a frequency distribution chart comparing actual performance of the system (time spent in A&E) to simulated performance. As the level

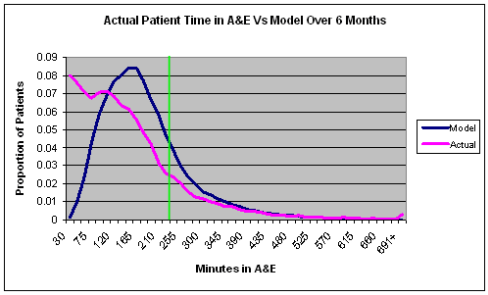
of detail included in the model changes so does the fit of the simulated output to actual. As an example, consider Figure 7.1; the charts represent before and after increasing the level of detail in modelling the radiology department. Depending on the route participants took to building the model they may see one of these two sets of charts. Notice that the two 'before charts' contain a large proportion of patients spending too great a time in A&E (resulting in the poor fit in the left hand side of the charts). The 'after charts' appear to have closer fit to the real data - although route one is closer than two.



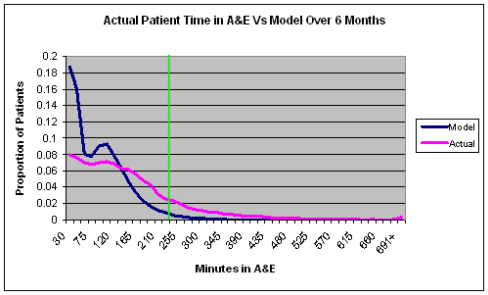
(a) Before (Route 1)



(b) After (Route 1)



(c) Before (Route 2)



(d) After (Route 2)

Figure 7.1: Output Examples: Before and After Increasing Radiology Level of Detail

Indeed it appeared that several model builders noted this difference in fit; typically stating 'ah, that is much better' and then reviewing their last change to the model and remaining simplifications. Given their initial neutral attitudes to the radiology improvement scenario, participants may be mildly surprised by the im-

portance of modelling radiology in detail to build a valid model. This is a possible mechanism for learning during model building. Novel and surprising events are remembered better (Butterfield and Metcalfe, 2006; Tulving and Kroll, 1995) and research has shown that novel and surprising events can help with more permanent and strong correction of errors - a term Butterfield and Metcalfe (2006) call hypercorrection. The importance of the factor in memory increases attention paid to feedback about the factor (Butterfield and Metcalfe, 2006; Tulving and Kroll, 1995; Chaiken et al., 1989). In this experiment that translates to model builders paying greater attention to the results of the radiology scenario than model reusers. However, it is acknowledged that participants could still follow a heuristic route to attitude change (Chaiken et al., 1989). Specifically 'apply more weight to the results of a radiology scenario, given that I discovered its importance in model building' or put more succinctly a *novelty means importance* heuristic.

This heuristic may explain why the majority of the difference in attitude change between the MBL and MR condition lies in the lower quantiles. It may be that these MBL participants have inflated attitudes simply due to the model building demonstrating an unexpected effect of modelling radiology in detail. The MR participants either experience very little attitude change as they never see the impact of modelling radiology in low and high detail or they learn about the effect in experimentation and experience a lot of attitude change.

It is also very likely that the measures included in the experiment did not capture all of the learning that model builders gained through novel ideas discovered during model development. One such example is the effect of patient prioritisation and variable inter-arrival times on the time spent in A&E. In early versions of the model patient prioritisation is not included and, due to the length of treatment, major emergency patients spend longer in A&E than minor patients. Once patient prioritisation is added (and also when inter-arrival time variation is increased) this

reverses. Minor emergency patients wait time for resources increase (both on average and in variance) due to prioritisation of major cases. This change in behaviour helped trigger inquiry by model builders. Participant MB19 provides an example of a common response:

> 'This [the proportions of minor and major emergency target breaches] is not very realistic. Of course, it takes a longer time to treat major emergencies than the minor ones. So what is the catch? Where is the mistake in the model?'

Participant MB19 then spent an extended time in V&V; watching the model run, pausing it and looking at the contents of queues proved to be useful. This is perhaps unsurprising given the marketed benefits of Visual Interactive Simulation (Belton and Elder, 1994). However, this was not always enough; some participants also chose to simulate the system without prioritisation included to see the effect on performance. Of course the effect of patient prioritisation could be questioned by model reusers, but it was only model builders who became, initially, very confused by the output. This is possibly due to the difference between model outputs once greater detail has been included.

7.2 Hypothesis 2: Incorrect Direction of Attitude Change

It was also predicted that involvement in model building would limit attitude change in the incorrect direction; although some support for this hypothesis was found, the results are inconsistent across the two model building conditions. This suggests two possible mechanisms that interact with learning: the level of experimentation, and face saving processes.

7.2.1 Mechanism: Confirmation Bias and Hypothesis Fixation

If utilisation is a difficult concept to understand and it is aided by experimentation then it is striking that MBL was less persuasive than MR and MB in both the correct and *incorrect* direction of change in $TradeUtil$. The constrained and controlled experimentation of MBL led to minimal learning (either right or wrong) about the relationship between resource utilisation and system performance. This indicates that the increased experimentation of MR and MB both aided and hindered learning.

Possible mechanisms for this difference in learning may be related to two issues that Bell and O'Keefe (1995) believe are important in discrete-event simulation experimentation: confirmation bias and hypothesis fixation (a subset taken from Fraser et al., 1992). These concepts are analogous to single-loop learning systems: individuals seek to remain within their predefined definition of effective performance. In other words individuals seek to confirm they are correct about managing utilisation rather than test that knowledge and possibly receive negative results (Argyris and Schön (1996) would describe this as a 'win not lose' governing variable leading to a learning process inhibited by face saving strategies). For example, participant MB14 ($TradeUtil$ = -20) quickly disregarded results showing reduction in performance from a self choice resource reallocation scenario. Appearing to be quite embarrassed that they were wrong MB14 quickly moved onto something more obvious (adding extra resources) that gave them an improvement in performance. MB14 then continued to choose this type of scenario only. The participants in the MBL condition had no opportunity to repeatedly reinforce any erroneous beliefs they may hold.

Alternatively, participants may fail to change their minds about effective performance even in the face of results from a number of disconfirming scenarios (hypothesis fixation). One area where this occurred most notably was the dedication or pooling of cubicles, nurses and doctors between patients. One participant, MB1,

simulated three scenarios where the total number of resources were divided up and dedicated to particular emergency streams. This always led to the same conclusion - lower performance as some patients were waiting in one queue while other resources were idle (due to process variation). Afterwards, during the post-test questionnaire, MB1 was asked what would happen if resources were dedicated - just as he or she had simulated. The answer below suggests MB1 had not changed his or her view despite the results of three scenarios chosen by them.

> 'Performance target breaches would decrease [improve performance] because minor patients would have their own cubicles and major patients would not be affected.'

The participant did not simulate all possible combinations of resources on the above problem; so one argument is that MB1 may believe it is simply a combinatorial problem. Note, however, that in all scenarios simulated by MB1 performance target breaches increase substantially (from 15% to around 33%). MB1 (who rated the model as highly credible) showed no signs of self doubt in his or her answer, given this unexpected result.

This result is notable given past views on the benefits of different types of experimentation. It has been argued and even empirically shown that proper statistical analysis and reporting has benefits over VIS (see Bell and O'Keefe, 1995) when locating an 'optimal' solution. Proponents of VIS have also stated that they are beneficial for discovery, clarification, change and creation of decision makers' views and ideas about system management (Belton and Elder, 1994). In this experiment participants had both statistical (although no formal, for example 2^k, experimental designs were used) and visual support for understanding results. Yet we see instances where attitudes moved in quite the wrong direction to benefit the system.

It is acknowledged that different models may give different results for learning about the relationship between resource utilisation and performance. On reflection

the A&E model has a difficult objective, a 98% service level, to obtain - as in real life departments - and as such really pushes the decision maker to think about the trade off between utilisation and performance. A different model, such as a factory, with different performance targets and constraints may make it easier for decision makers to learn. Alternatively the lack of a difficult target may allow such relationships to go unnoticed.

7.3 Hypotheses 3 and 4: Credibility Judgements

The results do not suggest any large differences between the median of the credibility assessment measure across the conditions. This was the expected result, given the pilot investigation. However, there does appear to be a difference in the variation of the credibility assessment measure. The most consistent results were given when model development was followed by limited (and controlled) experimentation. The model reuse condition produced the highest variation in assessment.

One area that a difference was expected was in the self confidence that participants had in their own assessments. Firstly, it was hypothesised that the more thorough the verification and validation (V&V) of the model by the participants the higher their self confidence reports would be. Secondly, it was hypothesised that a mechanism for increasing the sufficiency threshold of V&V was increasing the personal importance of the model to the client via involvement in model building. Put more simply it was expected that it would be more difficult for the model to pass model builders face validity checks than those of the model reusers due to their involvement in model building.

The results do not support these expectations. In fact there is some weak evidence that the model reuse condition produces higher self confidence than the model builders if experimentation is limited and controlled. Reflecting on the experiment it could be argued that the conditions provide different levels of process structure

to the participant. Figure 7.2 speculates that this process structure is linked to participants self confidence in and the consistency of their credibility assessment. Participants in the MBL condition were given the most process structure. Firstly, a task of reviewing modelling assumptions, simplifications and results as well as suggesting increases in the level of detail to reach sufficient accuracy. Secondly, a pre-defined set of experiments. The least structured condition was the MR condition. Participants were given an introduction to the model as well as a set of predefined experimentation. MR participants are allowed to verify and validate and use the model as they wished. The MB condition is a midpoint between these two conditions. As the level of structure increases so too does the consistency of the credibility assessment. There is a trade-off, however, with self confidence.

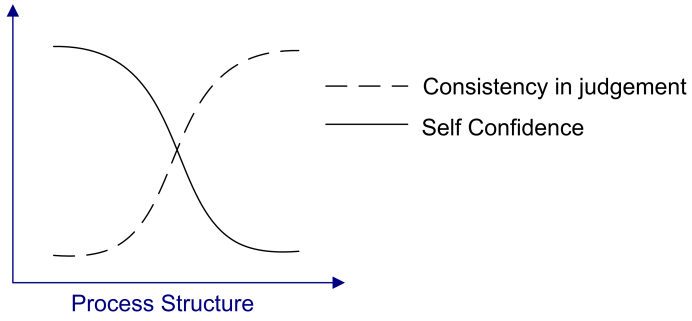


Figure 7.2: A Possible Credibility Mechanism in the Experiment

This is a speculative relationship and especially given the issues with influential points encountered in the analysis should be subject to further empirical investigation. This speculative relationship perhaps modifies the original hypothesis of the credibility assessment mechanism rather than refutes it. A possible addition to the hypothesis is the effect of simulation results that are different from a client's expectation. This may initially reduce a client's self confidence in their understanding of the model. More experimentation can help clients test differences in expectations and results and improve self confidence. However, it would seem that this increased experimentation should be structured in some way to improve consistency in the

outcome.

7.4 Applicability to real world simulation studies

7.4.1 Timings of Attitude Measurement

One factor to consider when assessing the applicability of results to real world simulation studies relates to the time frame of the experiment. In particular the experiment measures the participant's attitude immediately after the completion of the simulation study. This could be quite different from a study in industry: real decision makers may wait for, at least, a small period of time before committing to any implementation action.

Dikjksterhuis et al. (2006) report on a series of experiments testing the 'deliberation without attention' hypothesis. For example, in one experiment, when deciding on which car to buy one group of participants were given time to consciously process ('deliberate') information about the car. A second group were given the same time, but were distracted (they solved anagrams). Participants in this experiment (and field work) made better and more consistent decisions about complex purchases when they were distracted. One possible mechanism leading to this result is believed to be the relatively low capacity of conscious compared to subconscious processing. Thus, if the participants in this research were asked to 'sleep on it' and then report their attitudes the next day a different result may have occurred due to subconscious processing.

The studies discussed by Dijksterhuis et al (2006) do not invalidate the current research, but it does mean that the conclusions drawn from the results must be considered carefully. Clearly in a real simulation study, incorporating either development or reuse, clients would most probably have time to process (both consciously and subconsciously) the results for longer.

One area the experiment can apply to is the snap or early judgements made by decision makers. This is important because in real simulation studies decision making is likely to be made in groups. Thus, as detailed by theories of persuasion (E.g. Chaiken et al., 1989) and behaviour (E.g. Ajzen, 1991), a decision maker will take into account the early judgements of other members of the group. The early judgements of influential members of a group are likely to be important in how other members of the group choose to view the results after they have 'slept on it'. This makes the mechanisms for learning discussed very important, if one is interested in implementation actions following a simulation study. Especially since we cannot expect decision makers to systematically review how and if they have changed their minds (Rouwette, 2003).

7.4.2 Novice Decision Makers

A second difference between the experiment and real life is the level of expertise or perceived self expertise that a client has in the behaviour of the system of study. Empirical studies support the view that the greater the knowledge an individual has about a topic (or the greater the self confidence in one's own knowledge) then the lower the likelihood that advice is accepted (Yaniv, 2004; Kantowitz et al., 1992). This likelihood drops even further as the advice moves further away from the opinion of the individual (Yaniv, 2004). The hypothesis put forward to explain credibility assessment in simulation studies accounts for this by including the effect of perception of expertise on the sufficiency threshold for V&V, i.e. more persuasion is needed for individuals with higher belief in their knowledge of the system.

Clearly the participants in the experiment of this study are system novices and, given the empirical evidence, their credibility assessments are easier to manipulate. This means that the higher variability in MR credibility assessment is more striking - the process was unreliable at producing a consistent view even when participants

were fairly open to taking advice from the model. Similarly, the (albeit weak) evidence of a lower self confidence in assessment of the model given the more rigid structure of MBL might be magnified in a study with 'domain experts'. This would be especially so if the simulation model results were drastically different from those, implicitly, expected by the client.

A similar perspective can be taken when interpreting the attitude change results. The manufacturing managers and CEOs that partook in the Suri (1998) questionnaire might find it more difficult than students to reflect on their beliefs about resource utilisation and be persuaded that 100% utilisation results in large queues. In fact the students are system novices and it should be easier to persuade them to do the right thing. Thus results that indicate failures to change the attitudes of novices are useful as this suggests that experts would certainly not be persuaded.

Lastly, it is likely that the students will have reduced 'buy-in' (or motivation to understand) to the case study problem compared to real managers and decision makers with their own ideas and beliefs about how to improve performance. This extra motivation may improve the chances of attitude change. Although motivation in the experiment may not be as high as found in a real study some confidence in motivation levels can be drawn from both observations in the experiment and other empirical studies of statistical reasoning.

The first point that indicates that participants were highly motivated comes from the MBL condition. Anecdotally it was pointed out that many of the MBL participants wanted to simulate additional scenarios beyond the three that were prescribed without any prompting. Several of the participants became quite frustrated when they were not allowed to perform the additional experimentation believing that they had ideas that would improve performance. One participant, MBL21, took this a step further and e-mailed a number of scenarios (regarding changing the shift times and staff allocation) through asking for feedback on the performance.

In addition to the anecdotal evidence of this study, Brase et al. (2006) conducted four experiments of participant recruitment methods and their impact on statistical reasoning performance. Findings suggest that performance is influenced by the ranking of the (U.S.A based) university (i.e. students from higher ranked universities perform better) and financial incentives (i.e. students are more likely to engage when paid or when a chance to win more money is present). As participants of the current study come a top ten ranked U.K University and were given several financial incentives some confidence can be placed in level of engagement achieved.

7.5 Summary and Implications

In order to meet a decision maker's budget, simulation practitioners often face the difficulty of running a simulation study in a short space of time. If the technical difficulties can be overcome, model reuse offers the potential for practitioners and decision makers to use this time to concentrate on learning through experimentation. Alternatively practitioners can choose to develop a new model, involve the decision makers in the process, and perform, time permitting, some carefully chosen experimentation. This is an important decision for practitioners and the benefits of model building and experimentation beyond cost should be considered when making it. Although there are some limitations to this research, it is believed that when the results are carefully interpreted they offer four areas worthy of further study and for practitioner's consideration; these are summerised in Table 7.1.

Firstly, there is some evidence that more rather than less experimentation can aid learning when the participants' attitudes about a complex topic are strong. However, notably this can backfire; learning can be inhibited by the desire to win rather than lose in scenario selection and hypothesis fixation.

Secondly, it appeared that when building and validating a model, participants may experience some novelty in their thinking. Factors outside of their mental

model may have significant influence on system performance and focus attention or inflate the importance of related results. This persuasive power may be somewhat diluted - at least initially - when decision makers are no longer involved in building a model. The importance of these factors can also be demonstrated by experimentation: results suggested that an increase in the number of related scenarios may not be as convincing, but can sometimes prevent the factor from being disregarded as unimportant compared to when experimentation is limited.

A third concern was that when participants were limited in the amount of experimentation they could carry out they typically report less self confidence in their assessments of the simulation model than participants who reused the model. However, given the issues with outliers in the credibility data, it would seem that further investigation is required before implications can be explored. In the meantime, it is worth reiterating the argument that substantial verification and validation procedures are needed in any simulation project if decision maker confidence is to be achieved (Robinson, 2004).

Lastly, it is acknowledged that the experimental results relate to the early judgements of novice decision makers. This means that results should be interpreted carefully. Two suggestions are made here. Firstly, as the attitudes of novice decision makers are reported to be more easily influenced than experts then the results that indicate failures to learn/gain confidence may be the most significant. Secondly, the initial reactions of decision makers to simulation results are most important in group decision making. Discussion of early reactions to the model and results are likely to influence the attitudes of all members of the group in the long term.

To avoid issues caused by incorrect snap judgments it would seem plausible that some help with problem structuring is important for decision makers. The difficult task for a simulation practitioner to overcome is determining what judgements have been made by the decision maker(s). Richmond's approach of putting a 'stake

in the ground' (Richmond, 1997) may be a good option here. This would simply involve the practitioner recording decision maker expectations of performance prior to performing the analysis. For example, during early usage planning meetings in model reuse studies or as and when new scenarios come to light during model development. Once analysis is complete, a review of the results should include discussion of differences with *a priori* expectations.

This would also seem to offer some benefit even if decision makers have learnt the 'right thing'. Evidence suggests that often decision makers do not recognise that they have changed their minds about implementation options unless confronted with initial views (Rouwette et al., 2010). This would seem necessary to encourage learning beyond attitude change in both model reuse and model building studies.

So far this thesis has concentrated on analysis and discussion of the single-loop learning results. The following three chapters deal with double-loop learning and analyse the transfer of learning results. Chapter 8 describes double-loop learning by condition before proceeding to a detailed comparison in chapter 9. Chapter 10 then discusses the double-loop results.

Table 7.1: Important Findings

1. Some concepts are best learnt through experimentation, but are inhibited by
    - Complexity (and ability of decision maker);
    - Strong attitudes towards the best course of action;
    - Decision makers favouring expected positive results over unexpected negative results;
2. Novelty introduced via model building can aid learning by
    - Increased scrutiny of experimental factors related to novelty;
    - Application of a *novelty inflates importance* heuristic.
3. More work is needed to understand credibility
    - Weak evidence to suggest MBL held least confidence in their assessment of model.
4. Applicability of Results
    - Early judgement of simulation results and implementation actions;
    - Early judgements are more important in a group decision making situation;
    - Decision makers were novices:
        - Expert decision makers would find it more difficult to learn;
        - Failures to convince novices may be most important.

Chapter 8

Double-Loop Learning Results

8.1 Introduction

Chapters 5 and 6 presented the results of the single-loop learning variables. Single-loop learning is achieved when a new strategy to solve a business problem is adopted. Although the participants may have found several improvements to the fictional A&E department this does not automatically suggest that they will transfer this learning to future decision making situations. For successful transfer to occur the participant must have undergone a degree of double-loop learning: a change in their definition of effective performance. A problem with surface and structural similarity to the case study should cue transfer in these participants. This transfer should become more likely the closer the surface similarity of the transfer scenario to the case study problem (an A&E department). Furthermore, it is expected that the greater the participants involvement in model development the easier it is for participants to recognise the cues and successfully transfer.

Double-loop learning is analysed in three ways: transfer success, the confidence participants report in their reasoning and relationships between the single-loop and double-loop variables. This chapter firstly justifies the analysis approach used. This

is followed by the results for the three conditions in turn followed by a summary of the main findings. Three areas are described for each condition: firstly, high level aggregate results for transfer success and confidence. Interesting results from this section are then explored in an analysis of transfer and confidence results at the scenario level. Lastly, the relationship between the transfer, attitude and process variables are explored and discussed in a correlational analysis. The purpose of this chapter is to provide a descriptive account of the conditions only. Chapter 9 provides a comparison of conditions and a test of predictions.

8.1.1 Participants and Procedure

The same participants are used as reported in the single-loop learning results. On completion of the attitude questionnaire participants are firstly asked to predict and explain the change in performance of the A&E department if resource pooling is removed. Secondly, participants answer eight reasoning scenarios testing for transfer of concepts learnt from the simulation study. Thirdly, each participants provides a subjective rating of the confidence he or she has in their reasoning answer. The full procedure is detailed in the experimental design chapter; for convenience an overview of the transfer scenarios is summarised in Table 8.1.

Table 8.1: Summary of Transfer Scenarios

	Scenario	**Context**	**Reasoning required for transfer success**
Close	$S1$	GP's Surgery	Process Variation linked to Performance;
	$S2$	A&E department	Resource Utilisation and Performance;
	$S3$	Operating Theatre	Resource Utilisation and Performance;
	$S4$	NHS walkin Centre	Process Variation linked to Performance;
Far	$S5$	Pie Factory	Resource Utilisation and Performance;
	$S6$	Police Call Centre	Process Variation linked to Performance;
	$S7$	Pie Factory	Process Variation linked to Performance;
	$S8$	Call Centre	Resource Utilisation and Performance;

8.1.2 Reasoning Variables

There are two types of reasoning variables; a prediction relating to the A&E case study ($PredictPerc$) and transfer of learning to analogous decision making situation ($CorrectAnswers$ and $TransferSuccess$). Table 8.2 details the meaning and interpretation of these variables.

Both the transfer of learning measures can be broken down into answers for close and far scenarios. All scenarios share a structural similarity to the case study, but the surface similarity differs. The first four questions are set in a healthcare domain (close to the case study) and the remaining four are divided between a call centre and manufacturing domain (far from the case study).

Table 8.2: Reasoning Measures

Variable	Description
$PredictPerc$	The percentage of participants that correctly predict the behaviour of the A&E department after resources are dedicated to emergency streams instead of pooled.
$CorrectAnswers$	The number of correct multiple choice answers provided for the scenarios given out of eight. This can be subdivided into correct answers for close and far scenarios (each out of four);
$TransferSuccess$	The number of correct multiple choice answers, out of eight, that are accompanied with correct reasoning relating to the case study problem. This can be subdivided into close and far transfer success (each out of four);

8.1.3 Supporting Variables: Confidence

In addition to the reasoning variables that measure learning outcomes, participants report the confidence they hold in their answer (out of eight) for each scenario. If

a high confidence answer is subjectively defined as seven or greater (Chapter 9 performs a sensitivity analysis of this assumption) then it is possible to examine errors and transfer success at high and low confidence. These new variables provides quick information on any interesting behaviour within (and between) the condition(s).

8.1.4 Supporting Variables: Single-Loop Learning

In terms of double-loop learning the analysis is interested in any relationships between single-loop learning variables (i.e. the attitude change variables) and the transfer variables. That is, the analysis seeks to understand if a change in attitude i correlates with any transfer variable scores. The significance of bivariate correlations is assessed using the non-parametric spearman's rho r_s.

It is predicted that if double-loop learning has occurred then correlations will exist between corresponding attitude and transfer variables. For example, if the participants have grasped the relationship between utilisation and service level then the change in attitude $TradeUtil$ should be correlated to the participants score on questions testing this concept.

8.1.5 Coding of Transfer of Learning

Two codings of the reasoning answers have taken place in November 2009 and April 2010 respectively. These had high reliability, $\alpha = .87$, and had a high intraclass correlation coefficient (ICC), $.838 \leq r \leq .895$ (or more simply 91% of the codings were the same). In addition to minimise any judge bias (as the experimental hypothesis is known) all participant details were hidden from view and the order of answers was randomised in each coding. The differences between the codings typically reflected cases where the participants had provided minimal detail on their reasoning; differences were resolved by reviewing comments made in each coding and a combined index was produced.

8.2 Model Building

8.2.1 Transfer Success Summary

Table 8.3 summarises the transfer of learning results at the aggregate level. Results are split down by the number of correct multiple choice answers (out of eight), transfer success (out of eight), close transfer success and far transfer success (both out of four). The results show that MB participants typically select correct multiple choice answers 50% of the time. This drops slightly when we consider which scenarios contain clear signs of transfer success. As expected, the majority of transfer success occurs in scenarios with close surface similarity to the case study and simulation model. Section 8.2.2 explores the source of the drop between correct answers and transfer success in detail.

Table 8.3: MB Transfer Success Summary

Measure	*Mean*	*S.E*
CorrectAnswers	4.0	0.4
TransferSuccess	3.4	0.4
CloseTransferSuccess	2.3	0.2
FarTransferSuccess	1.1	0.2

CorrectAnswers and TransferSuccess out of 8

Close and Far Transfer out of 4

Table 8.4 summarises mean transfer success and multiple choice error by high and low confidence answers (high confidence is defined as seven and upwards). The majority of transfer success is accompanied by a high confidence report. There are also a similar number of high confidence errors.

Table 8.4: Mean Transfer Success by Confidence Level for MB participants

	Success	**Failure**
Low Confidence	1.0	2.1
High Confidence	2.4	2.6

High Confidence Cut-off $\geq$ 7 out of 9

8.2.2 Transfer Success by Scenario

Figure 8.1 summarises the results of the reasoning questions by the percentage of participants with correct answers (dark gray bar), transfer success (light gray bar), and the median confidence in their answer (line). The scenarios are shown in descending order of transfer success. That, is the first bar represents the scenario where participants achieved the most transfer success and the last bar represents the scenario where participants achieved the least transfer success. Using transfer theory it is assumed that the scenarios where transfer success is high are perceived quite similar to case study by participants. Similarly those scenarios where transfer success is low have low perceived surface similarity.

Figure 8.1 indicates that, in line with predictions, MB participants generally perform better on the close transfer scenarios (scenarios one to four) than the far transfer scenarios (scenarios five to eight). Notably, however, participant performance on scenario eight is a substantial improvement over the other far transfer scenarios.

Scenario eight is set in a call centre and tests the transfer of participants learning about resource utilisation and performance. It is similar to the case study as there are four staff shifts across the day. Participants see the performance (percentage of calls answered in a target time) of each shift individually along with the utilisation of staff. The options are to hire more staff for the two underperforming shifts, reallocate resources from other shifts or split the call centre into two specialised

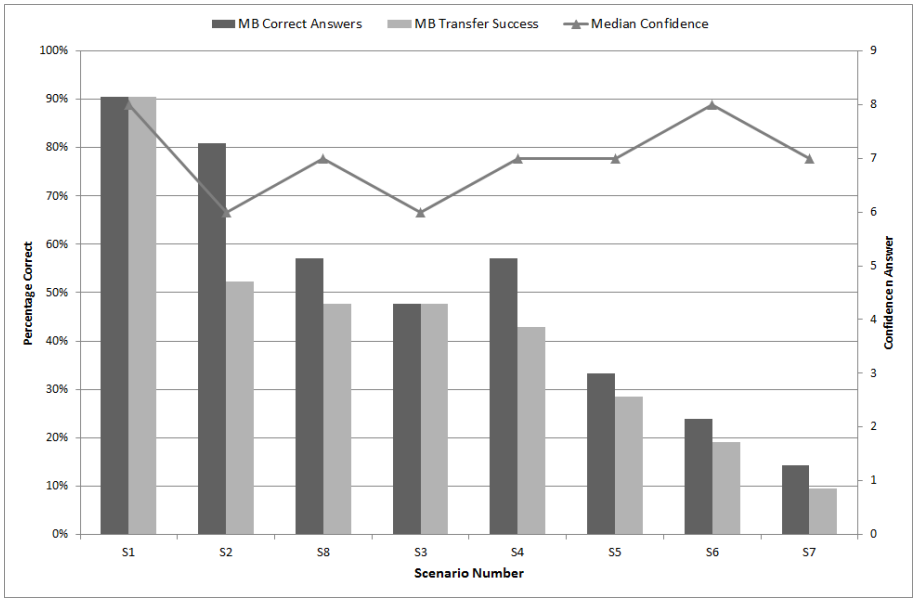


Figure 8.1: MB Participants: Percenatage of Transfer Success and Median Confidence by Scenario. (Ordered by Transfer Success)

lines (total number of resources remains the same). Transfer success is achieved by recognising that any drop in the high staff utilisation of the underperforming shifts will affect the performance target. In addition any increase in staff resource utilisation on the shifts currently meeting targets will reduce their performance against the call answering target. Figure 8.1 illustrates that nearly half the MB participants successfully transferred learning from the simulation model case study. An explanation for this increase in performance relative to the other far transfer scenarios is that scenario eight may be the closest in surface similarity to the case study; i.e. the case study had specific examples about shift reallocation. One participant even referred to the behaviour of the simulation model outputs when answering.

The average difference between correct answers and transfer success was 8% per scenario. However, as can be seen in Figure 8.1, scenario two has a large difference (29%) between correct answers and transfer success (having a substantial influence

over the mean value, median = 5%). As a reminder, scenario two is concerned with what should happen to staffing levels on an A&E shift where demand is expected to increase. Participants are told that the A&E department is subject to the same performance target (98% of patients must be through A&E in less than four hours) as the case study A&E. Transfer success is achieved by recognising that any increase in utilisation of staff will affect the performance target - especially when utilisation is high. Additional staff should be considered to maintain current performance.

Reasons for the large difference in scenario two transfer success and correct answers are in every case due to the participant considering if resources can 'cope' with demand; i.e. the participant considers any utilisation below 100% as capacity that should be used up before it is worth considering extra resource. This is fine in one sense as capacity may not be enough, but the answer does not transfer any learning from the simulation model used in the case study. In particular these answers show no consideration of the performance target (essentially the speed at which entities can travel through a process) even though it stated as an objective in the scenario.

Turning to the median confidence in answers, it is notable that Figure 8.1 indicates that scenario six consistently received one of the highest ratings (eight out of nine) yet has the second lowest number of correct answers. This is more interesting when compared with scenario four which tested the same transfer, but within a healthcare setting. As expected, MB participants perform better in the close transfer scenario; however, the median confidence in their answers is one point lower on the scale compared to scenario six. Figure 8.2 illustrates the distribution of differences between confidence scores. This is negatively skewed, but indicates that there is a slight increase in confidence from scenario four to six.

As a reminder, scenario six tested participants reasoning about the relationship between performance and variability of arrivals to several regional police call centres.

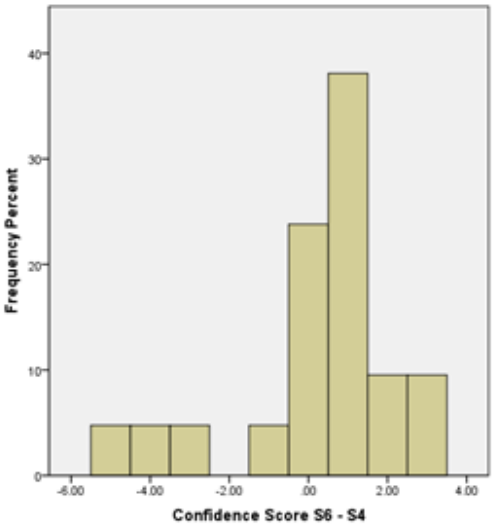


Figure 8.2: MB Participants: Distribution of Differences between S4 and S6 Confidence

Similar to scenario four, and the cubicle setup in the simulation model, participants needed to recognise that pooling resources is beneficial for performance and splitting resources is likely to increase waiting time. As expected, participants performed better in the close transfer scenario (S4) as opposed to the far transfer scenario (S6).

8.2.3 Relationship with Single-Loop Learning

There were several notable correlations between the transfer of learning variables and the attitude change and process variables from the single-loop learning analysis. Estimates of the magnitude of these correlations and their associated significance can be found in Appendix C, Table C.2. In summary these correlations indicate the following relationships.

1. Participants with large beneficial changes in $TradeUtil$ and $ElimVar$ have high transfer success reported with high confidence and vice versa;

2. Participants with large beneficial changes in $ElimVar$ have high transfer suc-

cess in the resource utilisation scenarios and vice versa;

3. The more scenarios simulated in the experimentation *the lower* the total (and far) transfer success and vice versa;

4. The more creative the participant in experimentation (i.e. selecting and combining new variables with given variables for experimentation) *the lower* close transfer success and vice versa;

5. Participants who recognised that variation in inter-arrival time was important in the first stage of model building typically have *low* close transfer success and vice versa.

The first two points listed indicate that some MB participants are probably undergoing double-loop learning. In line with the predictions attitude change predicts transfer success and the confidence of transfer success.

The third and fourth points relate to the experimentation stage of the simulation process. Total scenarios simulated is also negatively correlated with the average number of times a participant drills into results, to view more detailed charts, per scenario ($r_s = -.577$, $p < .05$). Thus one explanation is that these participants do not take time to fully reflect on the outcomes of scenarios before proceeding.

Point four is perhaps explained by the lack of focus on the core transfer concepts that are measured in the experiment. This does not necessarily indicate that participants have not undergone a degree of double-loop learning; however, this may be in an area unrelated to resource utilisation and process variation.

Point five relates to the model building process undergone by MB participants. An area participants found particularly challenging to recognise in model building was that inter-arrival times and staff shifts are initially modelled in insufficient detail. The MB participants that recognise this early in the process perform poorly in the transfer scenarios. This seems counterintuitive. In fact it might be expected that

these participants are the brightest students and thus should perform the best within the group in the transfer scenarios. Some weak evidence agrees with this point: the correlation between the stage of model building where this detail is included and the students exam results for QAM (see appendix B.2) is .440; however, it is non-significant.

One explanation may be that the difficulty involved in the discovery of the need to increase model detail in this way is related to learning. Some participants may find this choice and subsequent choices 'obvious'. Potentially the ease of this process leads to a lack of self-reflection and concentration on the simulation model results. Those participants finding this process difficult may question their beliefs and definitions of effective performance to a greater extent in turn increasing their attention to the results and chances of learning. Some weak quantitative evidence also points towards this speculation: the correlation between the stage of model building where this detail is included and the level of scrutiny given to results is -.404 (non-significant). This again would seem to be related to the novelty mechanism speculated upon in Section 7.1.2; i.e. when novel aspects of the problem are discovered during model building it increases the level of scrutiny participants pay to the problem and factors.

8.3 Model Building with Limited Experimentation

8.3.1 Transfer Success Summary

Table 8.5 summarises the transfer of learning results at the aggregate level. Results are split down by the number of correct multiple choice answers (out of eight), transfer success (out of eight), close transfer success and far transfer success (both out of four). MBL participants typically select correct multiple choice answers 40% of the time. This number drops, in a similar manner to MB, slightly when we

consider which scenarios contain clear signs of transfer success. Again, as expected, the majority of MBL transfer success comes from the close transfer scenarios. These results are explored in detail in the following section.

Table 8.5: MBL Transfer Success Summary

Measure	*Mean*	*S.E*
CorrectAnswers	3.1	0.3
TransferSuccess	2.3	0.2
CloseTransferSuccess	1.4	0.1
FarTransferSuccess	0.9	0.2

CorrectAnswers and TransferSuccess out of 8

Close and Far Transfer out of 4

A more interesting result is found in the level of confidence attributed to transfer success and error rates that are summarised in Table 8.6. The largest number in Table 8.6 corresponds to the average number of high confidence errors. This number is particularly noteworthy as it is slightly higher than the total number of correct answer and appears to be substantially higher than the total transfer success detailed in Table 8.5. This indicates that MBL participants may be overconfident in their decision making ability.

Table 8.6: Mean Transfer Success by Confidence Level for MBL participants

	Success	**Failure**
Low Confidence	0.5	2.0
High Confidence	1.8	3.7

High Confidence Cut-off $\geq$ 7 out of 9

8.3.2 Transfer Success by Scenario

Figure 8.3 summarises the results of the reasoning questions by the percentage of MBL participants with correct answers (blue bar), transfer success (red bar), and the median confidence in their answer.

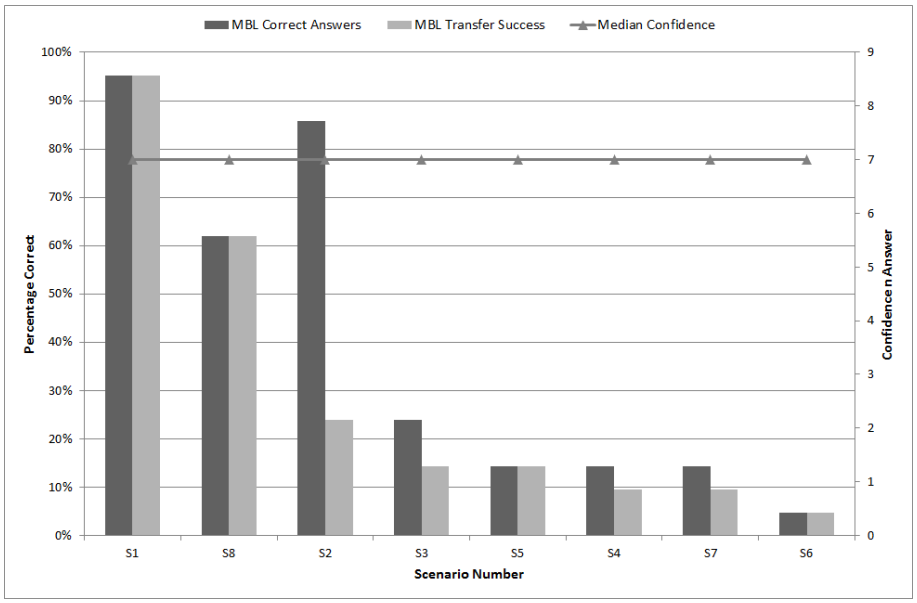


Figure 8.3: MBL Participants: Percenatage of Transfer Success and Median Confidence by Scenario

Figure 8.3 illustrates that the median confidence MBL participants had in their own answers was constant (seven out of nine) across the eight scenarios. Note also that MBL participants perform poorly on scenarios three to seven (and two if only transfer success if considered). This suggests that MBL participants were slightly overconfident in their own answers. Again like MB the largest difference between correct answers and transfer success is in scenario two answers. This difference of 62% has a substantial effect on the estimate of the average difference (10%) when typical performance was much closer and is best considered using the median (2%).

Participants perform best in scenario eight of the far transfer scenario (scenarios five to eight). This is substantially larger than the other far transfer scenarios: 62%

relative to the second highest of 14%. Similar to the result of MB this suggests that scenario eight contained the most surface similarity to the case study. Again one participant (MBL16) specifically referred to the case study simulation in their answer.

8.3.3 Relationship with Single-Loop Learning

There were only three significant correlations between the transfer of learning variables and the attitude change and process variables from the single-loop learning analysis for the MBL grouping. Estimates of the magnitude of these correlations and their associated significance can be found in Appendix C, Table C.3. In summary these correlations indicate the following relationships.

1. Participants who recognised that variation in inter-arrival time was important in the first stage of model building typically have high transfer success within the group and vice versa;

2. Furthermore, this higher transfer success is found in the far transfer scenarios;

3. Participants who viewed more output charts per scenario performed better in the questions where utilisation was the target of transfer.

In contrast to the MB condition, MBL does not contain any significant correlations with attitude change. However, there are again correlations between transfer success and the stage in model building that participants recognise inter-arrival times and staff shifts are initially modelled in insufficient detail (points one and two). This is the reverse relationship found in the MB condition: if participants recognise and choose to correct this deficiency in early model building then they performed better than those participants that recognised it later.

The difference in direction to MB for this correlation may be related to the fact that there is no evidence of a correlation between attitude change and transfer

success. It has already been speculated that the participants making this model building choice early are the brightest within the group. If there is only limited double-loop learning within the MBL condition then it is likely that the brightest students can transfer the most learning from the experiment. Another explanation is that double-loop learning may be present, but participants found it difficult to access the relevant knowledge based on the transfer cues present in the scenarios.

The final point indicates that increased scrutiny of scenario results was particularly beneficial to transferring concepts of resource utilisation across from the simulation study.

8.4 Model Reuse

8.4.1 Transfer Success Summary

Table 8.7 summarises the transfer of learning results at the aggregate level. Results are split down by the number of correct multiple choice answers (out of eight), transfer success (out of eight), close transfer success and far transfer success (both out of four). MR participants typically select correct multiple choice answers 43% of the time. This number drops, in a similar manner to MB and MBL, by one when we consider which scenarios contain clear signs of transfer success. Again, the majority of MR transfer success comes from the close transfer scenarios. These results are explored in detail in the following section.

Table 8.8 summarises mean transfer success and multiple choice error by high and low confidence answers. The majority of transfer success is accompanied by a high confidence report. The average number of high confidence errors appears to be less than the total number of correct answers reported in Table 8.7, but of a similar size to transfer success.

Table 8.7: MR Transfer Success Summary

Measure	*Mean*	*S.E*
CorrectAnswers	3.5	0.3
TransferSuccess	2.6	0.3
CloseTransferSuccess	1.5	0.2
FarTransferSuccess	1.1	0.2

CorrectAnswers and TransferSuccess out of 8

Close and Far Transfer out of 4

Table 8.8: Mean Transfer Success by Confidence Level for MR participants

	Success	Error
Low Confidence	0.8	2.7
High Confidence	1.8	2.8

High Confidence Cut-off $\geq$ 7 out of 9

8.4.2 Transfer Success by Scenario

Figure 8.4 summarises the results of the reasoning questions by the percentage of MBL participants with correct answers (blue bar), transfer success (red bar), and the median confidence in their answer.

Similar to the MB result the median credibility score is one point higher in scenario six (S6) than in scenario four (S4) (as a quick reminder, scenarios four and six test the same concept, but scenario four has higher surface similarity to the case study than scenario six). A difference from MB is that scenario six has higher transfer success than scenario four. This, at times, led to contradictions in reasoning about queues. In scenario four, for example, MR7 correctly reasoned that separate receptions and variable inter-arrival times leads to some patients waiting in one queue while receptionists serving the second queue are idle. MR7 then contradicted this reasoning by opting to split one or more police call centres into smaller regions as

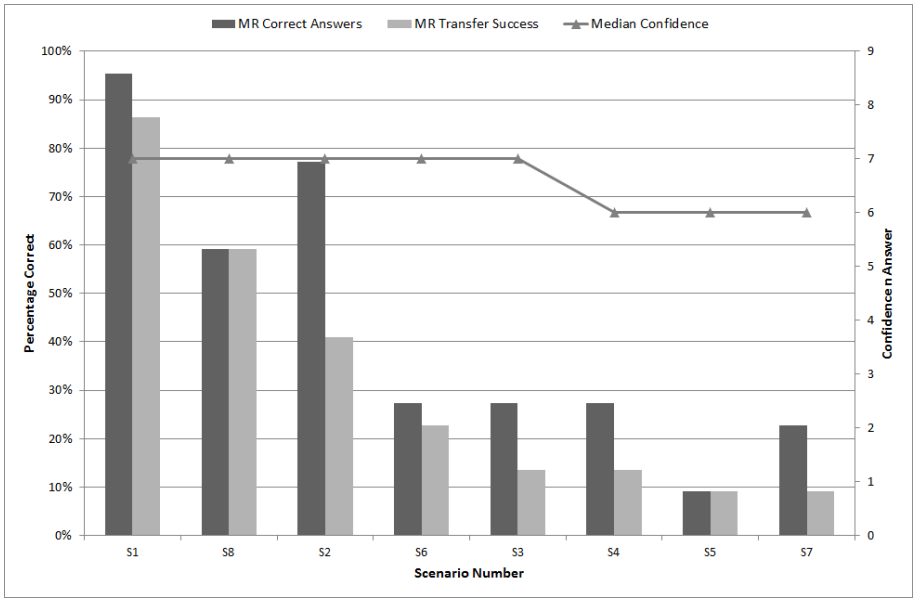


Figure 8.4: MR Participants: Percenatage of Transfer Success and Median Confidence by Scenario

'the number of calls can be handled more easily and without putting a strain on the same department in a short space of time'. Note, however, that this contradiction could also be reversed: MR8, for example, provided correct reasoning for S6 and an incorrect choice and reasoning for S4.

8.4.3 Relationship with Single-Loop Learning

Estimates of the magnitude of correlations and their associated significance can be found in Appendix C, Table C.4. In summary these correlations indicate the following relationships within the MR group.

1. Participants with large beneficial changes in $TradeUtil$ and $ElimVar$ have high close transfer success and vice versa;

2. Furthermore it appears that participants with large beneficial changes in $ElimVar$ have high transfer of learning related to process variation and vice versa;

3. When participants had a high average number of output statistics views per

scenario *they performed poorly* in the transfer scenarios (particularly on far transfer scenarios) and vice versa;

4. When participants had a high average number of output statistics views per scenario they had a *low number* of transfer successes with high confidence and vice versa;

5. The more creative the participant in experimentation (i.e. selecting new variables for experimentation) the higher their far transfer success and vice versa;

The first two of these points indicate that some participants have undergone a degree of double-loop learning about the relationship between resource utilisation and performance. The participants who experience the most change in their intention to improve A&E performance, but trade-off A&E resource utilisation in the case study, are able to transfer this learning to the new business scenarios. This indicates that these participants have refined their definition of effective performance for a system.

The remaining points refer to the experimentation stage of the simulation. Points three and four appear to be a counterintuitive results: although it would seem intuitive that learning would be higher the greater the scrutiny and attention participants pay to scenario results, learning and scrutiny are negatively correlated. A simple explanation for this result is that more time spent on scrutiny equals less time for running scenarios ($r_s = -.612$, $p < .05$). Although the total number of scenarios simulated is not significantly correlated with transfer success, it is positively correlated with $TradeUtil$ ($r_s = .459$, $p < .05$).

A possible explanation for point five is that the choice of new experimental factors is correlated with general intelligence. Participants identifying high numbers of novel experimental factors may be able to abstract what they are seeing in simulation results and transfer to new scenarios.

8.5 Summary

Table 8.9 details the main findings of the double-loop analysis by condition. Findings are grouped into three areas: transfer success, confidence and evidence of double-loop learning via relationships with single-loop learning. Although a thorough descriptive account of the double-loop results and supporting variables has been given in this chapter, further analysis is required to address the specific hypotheses discussed in Section 4.3.5. This is the topic of Chapter 9: a comparison of double-loop variables across the conditions.

Table 8.9: Summary of Descriptive Results

Transfer Success

- As predicted all conditions have high close transfer success relative to far transfer success;
- The largest discrepancy between correct multiple choice answer and transfer success was found in scenario two;
- Scenario eight appears to have the closest surface similarity out of the far transfer scenarios to the case study;

Confidence

- MBL make a large number of high confidence errors relative to transfer success;

Evidence of double-loop learning

- Evidence points to a relationship between the attitude change in $TradeUtil$ and transfer success in MB and MR conditions;
- No evidence of correlations between attitude change and transfer success is found in the MBL condition;
- Creativity in experimentation had opposite relationships in MB and MR. In MB higher creativity is linked to lower transfer success. In MR higher creativity is linked to higher transfer success;

Chapter 9

Double-Loop Learning Comparison

9.1 Introduction

The previous chapter presented a descriptive account of the double-loop results within each condition. This involved an analysis of the typical transfer success to a set of analogous decision making scenarios, an analysis of the confidence participants placed in the validity of their answers and a correlation analysis between single-loop and double-loop variables. The focus of this chapter is on comparing the conditions using graphical and formal tests of inference.

This chapter is split into four sections. The first section discusses the analysis methodology employed and discusses the use of effect sizes and sensitivity analysis. The second section presents a graphical comparison between conditions. This is done using simple bar charts for means, confidence intervals for means, and boxplots. The third section presents the inferential results concerning the predictions and exploratory tests. The chapter ends with a summary of the results linked backed to the predictions summarised in section 9.1.1.

9.1.1 Predictions and Exploratory Tests

The double-loop learning comparison tests five predictions and performs three additional inference tests based on the exploratory work of the previous chapter and the graphical analysis included in this chapter; these are summarised in Table 9.1. When the descriptive statistics show no difference between conditions, no formal inference procedures are employed. These are discussed on a case by case basis.

Table 9.1: Summary of Predictions and Exploratory Tests

Hyp	**Measure(s)**	**Prediction**
d.1	Prediction Success	MB & MBL > MR
d.2	Total Transfer Success	MB & MBL > MR
d.3	Close Transfer Success	MB & MBL > MR
d.4	Far Transfer Success	MB & MBL > MR
d.5	Correlation: attitude change and transfer	Correlation present in all conditions
d.6	High Confidence Errors	MBL produced the highest number
d.7	Performance on Scenario 7	MR > MBL
d.8	Performance on Scenario 6	MR > MBL

9.2 Analysis Considerations

9.2.1 Inference using Categorical Variables

The final two exploratory tests listed in Table 9.1 are tests of categorical variables (i.e. variables that can take the value 0 or 1). For example, a participant either achieves transfer in scenario seven (coded as 1) or not (coded as 0). The standard inference approach for categorical variables between two conditions is the chi-square test of association. This is based on a simple contingency table - a breakdown of the number of participants achieving transfer by condition. Effect sizes for the chi-square test of association are typically given as an odds ratio (Field, 2009).

The odds ratio has a very simple interpretation. For example, a test for transfer on scenario x between MR and MBL has an odds ratio of 3.5. This is interpreted as MR participants are 3.5 times as likely to achieve transfer success as MBL. Appendix D.1.1 details the procedure for estimating an odds ratio from a contingency table.

9.2.2 High Confidence Cut-Off

The scale for participants to rate their confidence in their answers runs from 1 (no confidence) to 9 (extremely confident). It was decided that participants are deemed to have high confidence if their answers are greater than seven. As the choice of this cut-off is subjective (for example, six or eight could have been chosen as the cut-off) a sensitivity analysis of high confidence errors to the cut-off is conducted for each comparison. Where the difference between the numbers of high confidence errors in conditions appears to be sensitive, a range of inferential results are provided.

9.3 Graphical Analysis

9.3.1 MR versus MB

Prediction Results

Figure 9.1 compares the proportion of participant's correctly predicting the performance of the system if cubicles and staff resources were dedicated to specific emergency streams rather than pooled. Figure 9.1 illustrates both the proportion of participants correctly predicting the outcome, shown in blue, and the proportion of participants also supplying the correct reasons for the outcome.

Figure 9.1 illustrates that there is little difference in the outcome between MB and MR in making the correct prediction - MB has slightly more success (this is 10% or two participants). When correct reasoning is considered the outcome reverses slightly. Both conditions have a drop in success rate with more MR participants

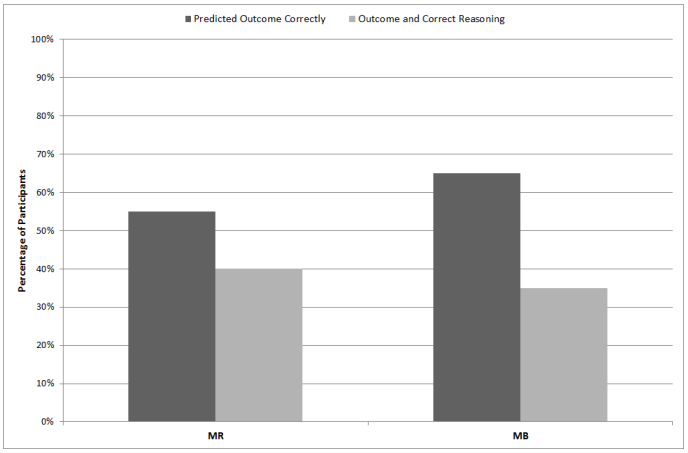


Figure 9.1: Prediction Success MR versus MB

supplying the correct reasoning. Note that this is not in line with the prediction (d.1).

Aggregate Transfer Results

Figure 9.2 compares the mean transfer success of the MR and MB participants at the aggregate level as well as the split into the close and far transfer success. Error bars are bootstrapped (B=2000) 90% confidence intervals for the means.

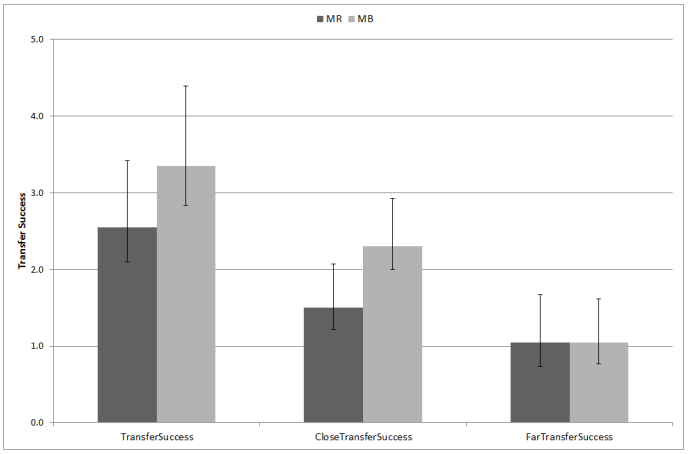


Figure 9.2: MR versus MB: Mean Transfer Success. It was predicted (d.2, d.3 and d.4) that $MB > MR$

The estimate of mean $TransferSuccess$ is slightly higher in the MB condition. However, there is a reasonably large overlap of the confidence intervals. The difference is much clearer if only $CloseTransferSuccess$ is considered. Here the mean difference is the same as for $TransferSuccess$ (Difference (D) = 0.8), but the confidence interval overlap is substantially smaller. This indicates that, in line with predictions, MB participants have higher transfer success in the close scenarios (d.3). Note that, counter to predictions (d.4), there is no difference in $FarTransferSuccess$.

Results by Scenario

The aggregate results suggested a difference in the performance of MR and MB participants in the close transfer scenarios. Figure 9.3 is used to investigate this result further: the percentage transfer successes for each scenario by condition and the median confidence in answers are presented.

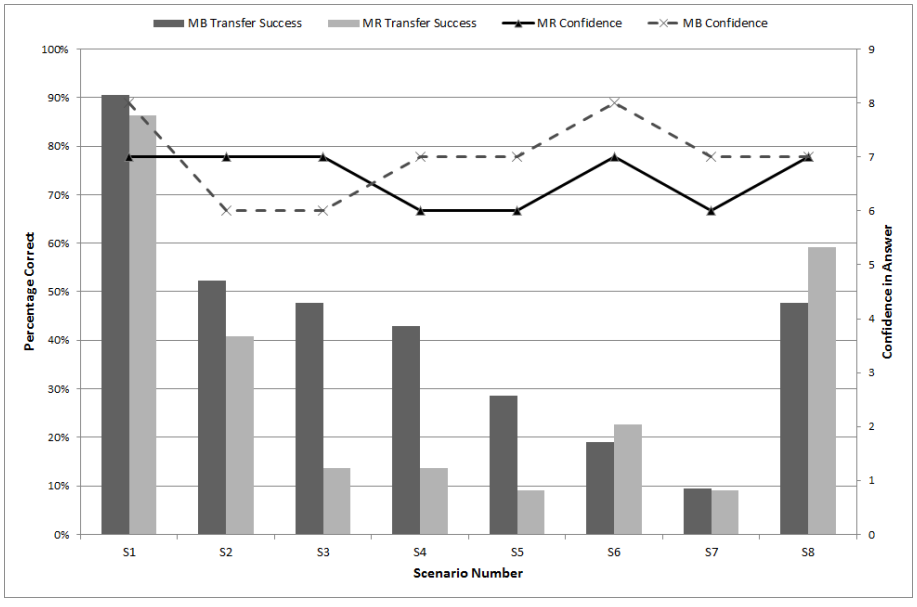


Figure 9.3: MR versus MB Transfer Results by Scenario

Figure 9.3 illustrates that the difference in close transfer success largely comes

from the scenario three (D = 34%) and scenario four (D = 29%), (A two-tailed significance test supports this statement: $r = .50$, $p < .01$). As a reminder

- Scenario three tested participants reasoning about the relationship between performance (operating room overruns) and resource utilisation in an operating theatre setting. Higher performance is achieved by considering additional resources.

- Scenario four tested participants reasoning about relationship between performance and the variability of arrivals to an NHS walk-in clinic and GP surgery receptions within a shared building. Higher performance (lower average queuing time and increased average staff utilisation) is achieved by pooling the reception staff.

Mistakes on scenario three typically involved participants believing the operating rooms where not used enough (current average utilisation is stated as 82%). For example, MR5 answered:

> '82% usage means that there is enough free time to perform the operations during the day, but the time is not utilised enough, thus spare slots should be used up.'

Participants also considered increasing the duration of operating slots. Indeed this would cut waiting times due to operation overruns; however, these participants always failed to consider the potential impact on the waiting time to book an operating slot.

Mistakes on scenario four typically took the form of achieving speed through 'specialisation' of resources. For example, MR1 answered:

> 'Separating queues will reduce waiting times as patients know what [service] they are queuing for. Segregation of tasks increase efficiency and speed.'

Overall both conditions are least successful with the manufacturing scenarios (five and seven) in far transfer. MB participants appear achieve higher transfer success on scenario five (D = 20%).

Confidence

Figure 9.4 illustrates the distribution of high confidence errors in answering the transfer scenarios for the MR and MB conditions using boxplots. The main point to notice is that the median number of HCE errors is similar across the conditions (mdn diff = 0.1). There appears to be some difference in the distributions as MB is positively skewed (illustrated by the longer lower whisker) and MR shows some negative skew (illustrated by the longer upper whisker). However, in general the distributions have substantial overlap.

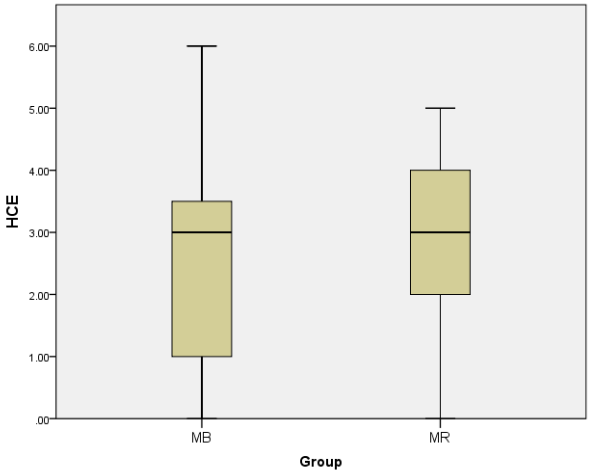


Figure 9.4: MR Vs. MB Boxplot of High Confidence Errors

A sensitivity analysis of the high confidence cut-off is illustrated by Figure 9.5. This simply illustrates the mean number of high confidence errors in each condition if the cut-off is varied between five and eight. A lower bound of five was chosen as the scale is out nine; any lower than five is less than half way down the scale does not seem to be a plausible choice for high confidence. An upper bound of eight was

chosen as it seemed intuitive that a maximum choice on the scale (i.e. nine) was not a sensible choice for the cut-off.

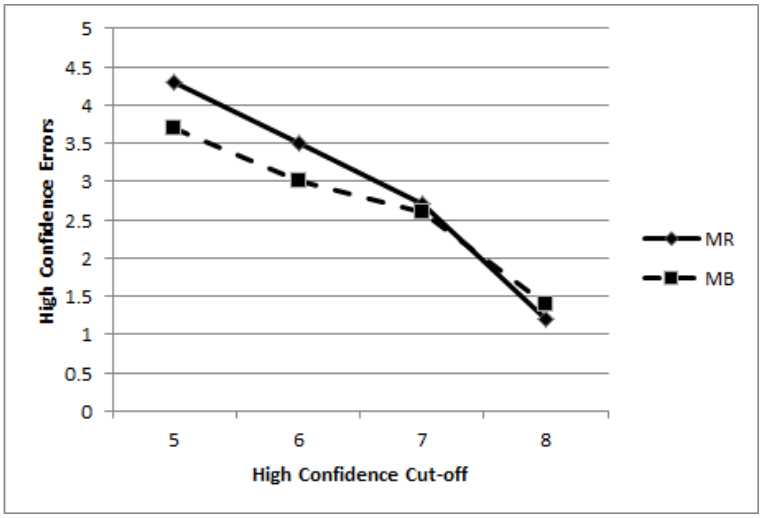


Figure 9.5: MR Vs. MB Sensitivity of High Confidence Cut-off

Figure 9.10 suggests that MR is the most sensitive to a lowering in the cut-off from seven. This leads to a larger mean difference between the conditions at a cut-off of six ($D = 0.5$) and slightly larger again difference at five. Both conditions are highly sensitive to an increase in the cut-off; participants are much more unlikely to make a high confidence error at eight or above. However, the difference between the conditions does not change. Thus it would seem that the choice of this cut-off is important in how the comparison data are interpreted. If a cut-off of six was chosen then there is a larger difference between conditions; although this is not exceptionally large.

9.3.2 MR versus MBL

Prediction Results

Figure 9.6 compares the proportion of participant's correctly predicting the performance of the system if cubicle and staff resources were dedicated to specific

emergency streams rather than pooled.

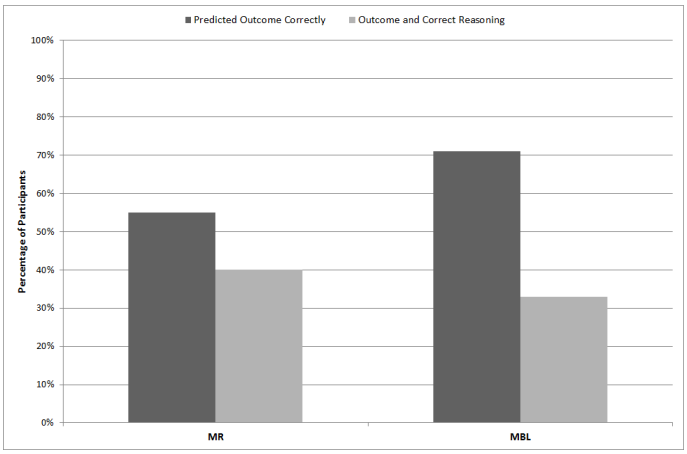


Figure 9.6: Prediction Success MR versus MBL

The MBL participants make the most correct predictions (70% versus 55%); however, this proportion notably falls when correct reasoning is considered. The highest proportion of correct reasoning was supplied by the MR condition.

Aggregate Transfer Results

Figure 9.7 compares the mean transfer success of the MR and MBL participants at the aggregate level as well as the split into the close and far transfer success. Error bars are bootstrapped (B=2000) 90% confidence intervals for the means.

Figure 9.7 illustrates little difference between the two conditions at the aggregate level. Although there is a minor difference between the estimates of the means the confidence intervals overlap to a large degree. Thus it is highly unlikely that any difference will be detected using inference procedures.

Results by Scenario

Figure 9.8 is used to investigate the aggregate results further: the percentage transfer successes for each scenario by condition and the median confidence in answers are presented.

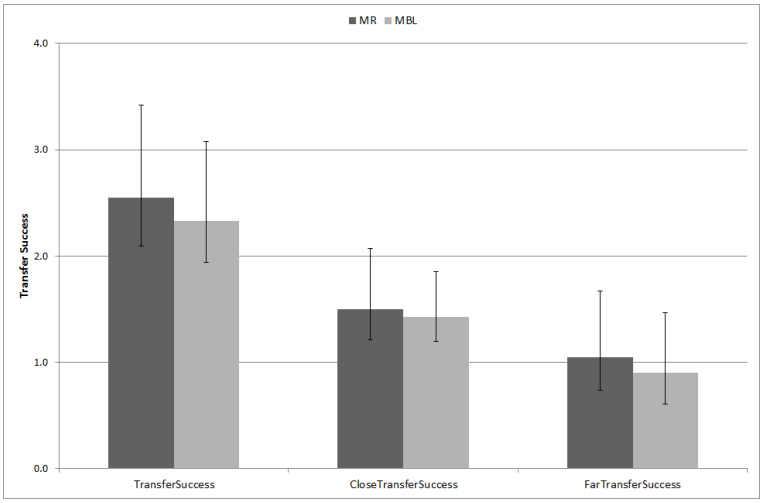


Figure 9.7: MR versus MBL: Mean Transfer Success. It was predicted (d.2, d.3 and d.4) that $MB > MR$

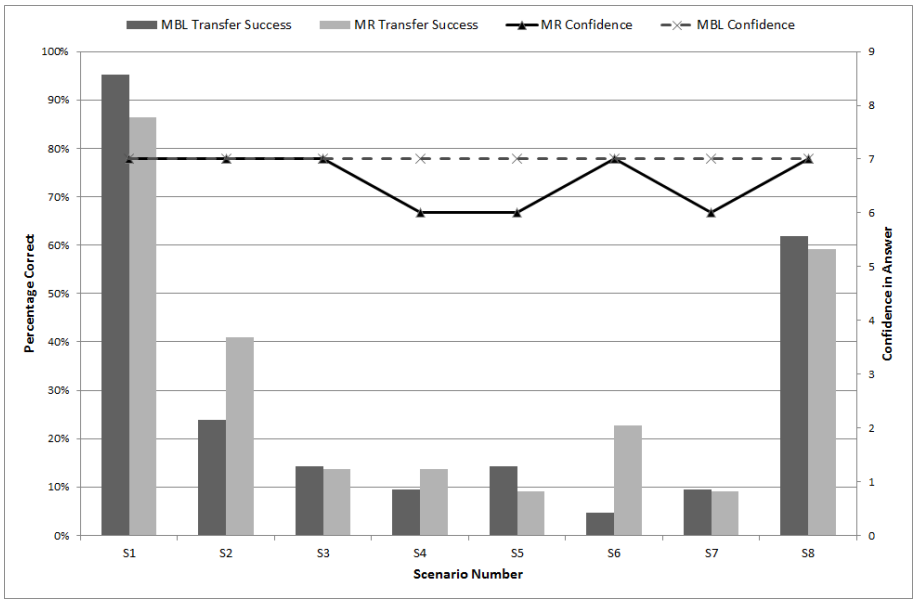


Figure 9.8: MR Vs. MBL Transfer Results by Scenario

The largest differences in transfer success are found in scenarios two (D = 17%) and six (D = 18%). The result for scemario six is unusual as, noted in Section 8.4, MR performs better in the far transfer scenario than scenario four - the scenario testing for the same transfer at the close level of surface similarity. In fact there were

no significant associations (chi-square test of association) between the performance in scenarios four and six in either condition.

Confidence

Figure 9.9 illustrates the distribution of high confidence errors in answering the transfer scenarios for the MR and MBL conditions using boxplots. The main point to notice is that the median number of HCE errors is slightly higher in the MBL condition (mdn diff = 1.0). However, another indication of difference is that the lower quartile for MBL is the same as the median for MR: indicating that the MBL distribution sits to the right of MR.

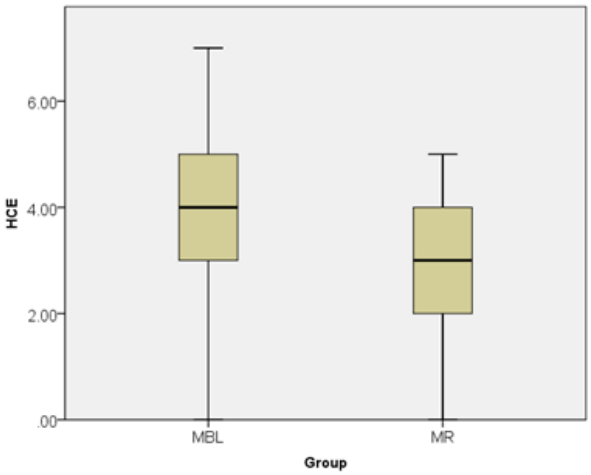


Figure 9.9: MR Vs. MBL Boxplot of High Confidence Errors

A sensitivity analysis of the high confidence cut-off is illustrated by Figure 9.10. This simply illustrates the mean number of high confidence errors in each condition if the cut-off is varied between five and eight. A lower bound of five was chosen as the scale is out nine; any lower than five is less than half way down the scale does not seem to be a plausible choice for high confidence. An upper bound of eight was chosen as it seemed intuitive that a maximum choice on the scale (i.e. nine) was not a sensible choice for the cut-off.

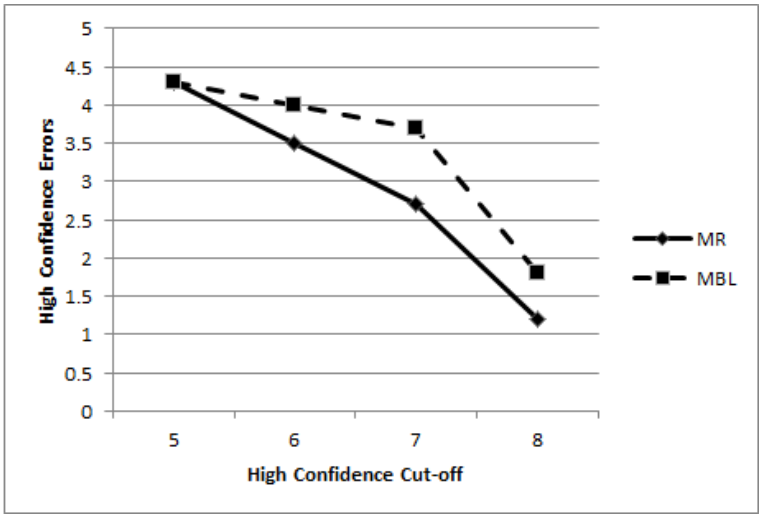


Figure 9.10: MR Vs. MBL Sensitivity of High Confidence Cut-off

Figure 9.10 suggests that MR is the most sensitive to a lowering in the cut-off from seven. This leads to a smaller mean difference between the conditions at a cut-off of six ($D = 0.5$) and no difference at five. Both conditions are highly sensitive to an increase in the cut-off; participants are much more unlikely to make a high confidence error at eight or above. Thus again, it would seem that the choice of this cut-off is important in how the comparison data are interpreted. The high sensitivity of both conditions to an increase in the cut-off indicates that eight or above on the scale reflects extremely high confidence - an area where only a few of the participants rate themselves. However, we could also interpret high confidence errors at six or above. This leads to a smaller difference between the conditions.

9.3.3 MB versus MBL

Prediction Results

Figure 9.11 compares the proportion of participant's correctly predicting the performance of the system if cubicle and staff resources were dedicated to specific emergency streams rather than pooled. The performance of the two conditions is very

similar for both correct prediction and correct reasoning. There is only a difference of 7% (1.91 participants) in the correct prediction and 2% (< 1 participant) in the correct reasoning outcomes.

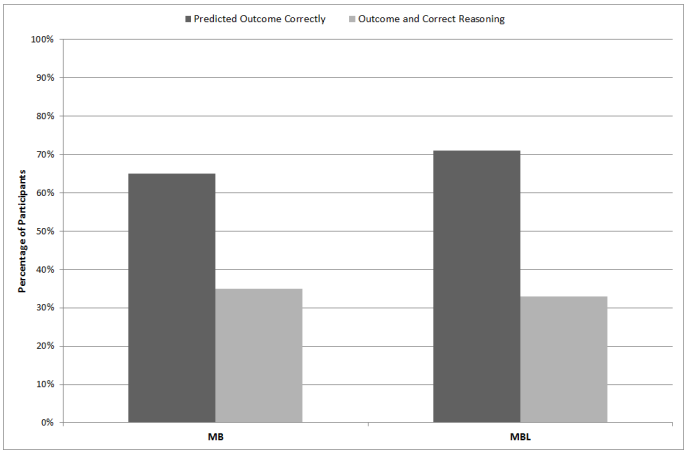


Figure 9.11: Prediction Success MB versus MBL

Aggregate Transfer Results

Figure 9.12 compares the mean transfer success of the MB and MBL participants at the aggregate level as well as the split into the close and far transfer success. Error bars are bootstrapped (B=2000) 90% confidence intervals for the means.

Figure 9.12 suggests that the difference in typical *TransferSuccess* and *Close-TransferSuccess* is approximately one scenario higher in the MB condition. The confidence intervals offer some support of this statement: there is only a small overlap between MB and MBL on $TransferSuccess$ and no overlap on $CloseTransferSuccess$.

Transfer by Scenario

The aggregate results suggested a difference in the performance of MR and MB participants in the close transfer scenarios. Figure 9.13 is used to investigate this result further: the percentage transfer successes for each scenario by condition and the median confidence in answers are presented.

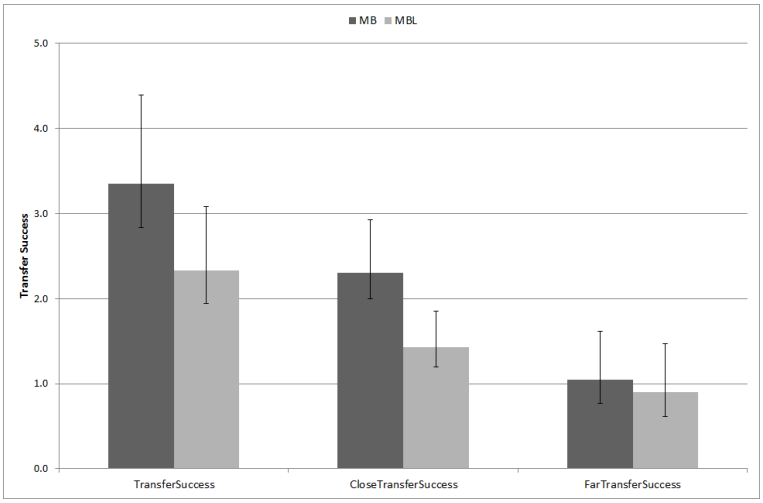


Figure 9.12: MB Vs. MBL - Mean Transfer Success

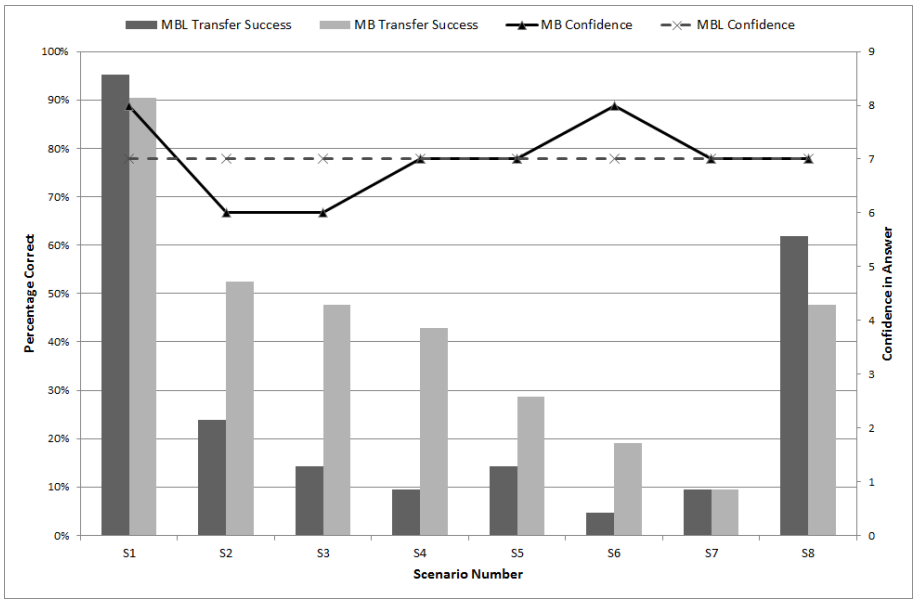


Figure 9.13: MB Vs. MBL Transfer Results by Scenario

Figure 9.13 indicates that MB out performs MBL on close scenarios two to four and far transfer scenario five (D = 14%) and six (D = 14%). A higher percentage of MBL participants have transfer success on scenario eight (D = 14%) even though it tests for the transfer of the same concept as scenarios two, three and five - where MB has the highest transfer success.

Confidence

Figure 9.14 illustrates the distribution of high confidence errors in answering the transfer scenarios for the MB and MBL conditions using boxplots. Again like the comparison between MR and MBL the main point to notice is that the median number of high confidence errors is slightly higher in MBL condition (mdn diff = 1.0, mean diff = 1.1). The difference in skew between the conditions also indicates a potential difference. The long upper and short lower whiskers of the MB boxplot indicate positive skew - the majority of the data lie at the lower end of the scale. Although not as severe as MB, the MBL boxplot displays negative skew - the majority of the points lie at the upper end of the scale.

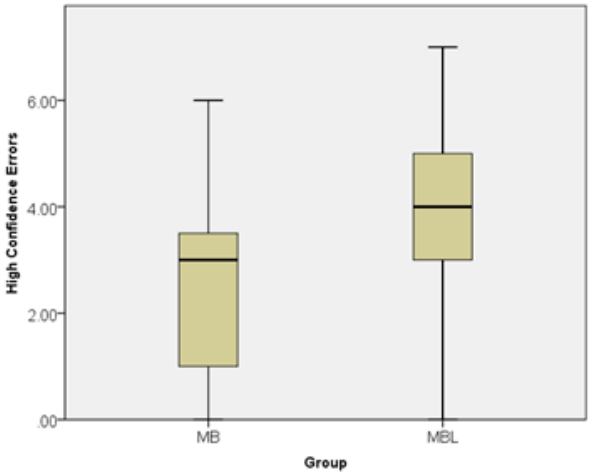


Figure 9.14: MB Vs. MBL Boxplot of High Confidence Errors

Figure 9.15 illustrates a sensitivity analysis of the high confidence cut-off in the MB and MBL comparison. The points represent the mean number of high confidence errors in each condition if the cut-off is varied between five and eight. The reasons behind the choice of lower and upper bounds for the analysis are detailed in Section9.3.1

Figure 9.10 suggests that both conditions are highly sensitive to an increase

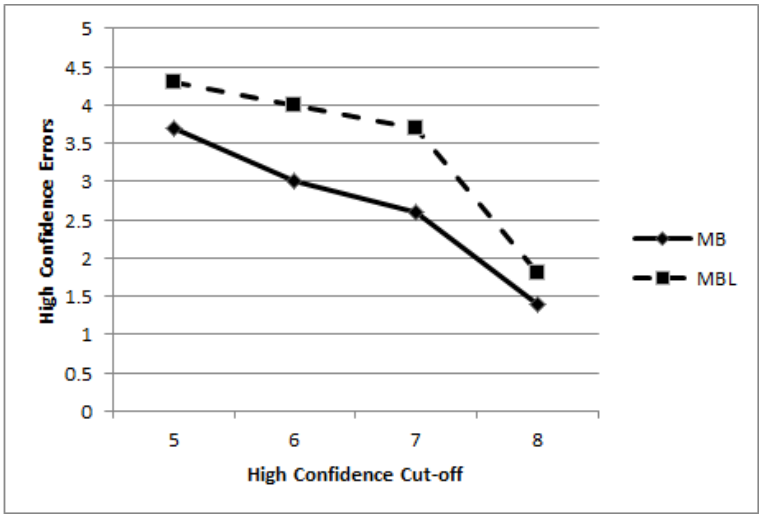


Figure 9.15: MB Vs. MBL Sensitivity of High Confidence Cut-off

in the high confidence cut-off to eight. As seen in the other comparison sections, participants are much more unlikely to make an error and report it with a confidence of eight or above. However, both conditions are fairly insensitive to a drop in the cut-off to six. Thus it seems that seven is a fairly sensible choice to analyse. Using eight as the cut-off is too high and interpretation remains the same if the cut-off were dropped to six.

9.4 Inferential Results

This section builds on the graphical analysis by providing formal inferential results for the predictions made about learning. The information presented here outlines support, lack of support or contradictions of predictions given by the bootstrap and non-parametric tests. Where disagreement exists between the inference tests, typically in the form of the non-parametric test failing to find a difference where the bootstrap does, discussion of the graphical results and effect sizes are used in an attempt to provide an explanation.

9.4.1 MR versus MB

Table 9.2 summarises the inferential results for the test of predictions in the MR and MB comparison. The colour of the cells indicates agreement between the non-parametric and bootstrap tests: green indicates agreement and red indicates disagreement.

Table 9.2: MR versus MB Summary of Inferential Support for Predictions

Hyp	Measure	Result	Effect Size (r)
d.1	PredSuccess	No Difference	.10
d.2	TransferSuccess	$MR < MB^*$	.30
d.3	CloseTransferSuccess	$MR < MB^*$	.47
d.4	FarTransferSuccess	No difference	.00

*Asterisked results indicate support for predictions

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

The tests disagree in only one case: the total transfer success. The bootstrap concludes that there is evidence to suggest that the difference in means is highly unlikely to be due to chance alone, while the non-parametric test cannot conclude that there is evidence.

Although there is a disagreement the data do suggest a medium effect size is present for total transfer success (r = .30). So the disagreement is likely to be due to a lack of sufficient power in the Mann-Whitney test. The graphical analysis of Section 9.3.2 indicated that there was a fairly large overlap of the 90% confidence intervals for mean transfer success between the conditions. However, this is explained by very similar performance in the far transfer scenarios. The real difference lies in the performance on close transfer success - where the tests agree - where a medium to large effect is present (r = .47).

The graphical analysis of the data also indicates that the typical confidence of

the MR and MB participants was similar. Inference results summarised in Table 9.3 support this interpretation: no difference was found in the number of high confidence errors between the conditions at either a cut-off of six or seven.

Table 9.3: MR versus MB Support for Differences in Confidence

Hyp	Measure (cut-off)	Result	Effect Size (r)
d.6	High Confidence Errors (7)	No difference	.06
d.6	High Confidence Errors (6)	No difference	.19

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

9.4.2 MR versus MBL

Table 9.4 summarises the inferential results for the test of predictions in the MR and MB comparison. In all cases the tests agree - no significant differences are present across the three measures. Thus no support for the prediction that MBL will transfer more learning from the simulation case study is found.

Table 9.4: MR versus MBL Summary of Inferential Support for Predictions

Hyp	Measure	Result	Effect Size (r)
d.1	PredSuccess	No Difference	.17
d.2.	TransferSuccess	No difference	.07
d.3.	CloseTransferSuccess	No difference	.03
d.4	FarTransferSuccess	No difference	.10

*Asterisked results indicate support for predictions

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

Although the results show no difference at the aggregate level the graphical analysis of results by scenario indicated a possible difference in scenario two and scenario six. Table 9.5 summarises the results of chi-square tests of association.

These support the view that MR has higher performance on scenario six ($p < .1$). In fact the odds ratio suggests that the MR participants are 6.7 times more likely than MBL participants to achieve transfer on scenario six. However, the analysis did also suggest possible issues with the chi-square test results due to small group sizes within the contingency table (i.e. counts < 5). Thus it is difficult to say conclusively that the results are not due to random variation. For a discussion of the test results and to view the contingency table, see Appendix D.3.

Table 9.5: MR versus MBL - Summary of Differences in Scenario Results

Hyp	**Measure**	**Result**	**Odds Ratio**
d.7	Performance on Scenario 2	No difference	2.1
d.8	Performance on Scenario 6	MR > MBL	6.7

The graphical analysis of the data also indicated that MBL participants were slightly overconfident in their answers to the scenarios. Inference results for high confidence errors summarised in Table 9.6 support this interpretation at a scale cut-off of seven: the higher number of mean high confidence errors in MBL is highly unlikely to be due to chance alone (r = .28). As illustrated by the graphical results MR is sensitive to a drop in the cut-off. The inference results in Table 9.6 support this view. At a cut-off of six, the effect size becomes small ($r = .14$) and the result is no longer significant.

Table 9.6: MR versus MBL Support for Differences in Confidence

Hyp	**Measure (cut-off)**	**Result**	**Effect Size (r)**
d.6	High Confidence Errors (7)	MR < MBL	.28
d.6	High Confidence Errors (6)	No difference	.14

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

9.4.3 MB versus MBL

Table 9.7 summarises the inferential results for the test of differences between the MB and MBL comparison. As there are no specific predictions for outcome the results are presented in a format indicating the direction of the difference. For example, the result MB > MBL indicates the MB achieved higher transfer success than MBL on the variable.

Table 9.7: MB versus MBL Summary of Inferential Support for Predictions

Hyp	**Measure**	**Result**	**Effect Size** (r)
d.1	PredSuccess	No Difference	.07
d.2.	TransferSuccess	MB > MBL	.36
d.3.	CloseTransferSuccess	MB > MBL	.50
d.4.	FarTransferSuccess	No difference	.08

*Asterisked results indicate support for predictions

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

In all cases the tests agree; two results indicate significant differences. As indicated by the graphical analysis of the data there is a difference in the average transfer success of the MB and MBL participants. This is largely explained by the higher performance of MB participants in the close transfer scenarios (r = .50).

The graphical analysis also suggested that, similar to the MR versus MBL comparison, the MB participants made less high confidence errors. Inference results for high confidence errors summarised in Table 9.6 also support this interpretation (given a cut-off of six or seven): the higher number of mean high confidence errors in MBL is highly unlikely to be due to chance alone (r = .28).

This result on its own may seem unremarkable as Table 9.7 shows that MB make less errors anyway. However, the previous section illustrates that this result is also found in the MR versus MBL comparison, given a cut-off of seven. Here the transfer

Table 9.8: MB versus MBL Support for Differences in Confidence

Measure (cut-off)	Result	Effect Size (r)
High Confidence Errors (7)	MB < MBL	.28
High Confidence Errors (6)	MB < MBL	.30

Shading indicates agreement between non-parametric and bootstrap inference.

Comparison Key	Agreement	Disagreement

success level is the same. Thus there may be some effect leading to a higher number of high confidence errors in MBL (or a lower number in MR).

9.5 Summary of Group Differences

Table 9.9: Summary of within and between group differences

Support for Predictions for Transfer

- No support for the superiority of both MB and MBL over MR in predicting transfer performance;
- Total and Close Transfer Success is higher in MB than MR;
- However, MBL and MR have a very similar number of total and close transfer successes;
- Some weak evidence that MR is more successful than MBL on scenario six;

Support for Predictions about Double-Loop Learning

- MR participants transfer success is related to the $TradeUtil$ and $ElimVar$ attitude change variables in the case study;
- MB participants transfer success is related to the $TradeUtil$ and $ElimVar$ attitude change variables in the case study;
- No relationships between transfer success and attitude change are found in MBL;
- Experimentation (search method and creativity) seems to play a role in MB and MR transfer success - although it affected the conditions differently;

Table 9.9 – continued from previous page

Results for Confidence
• MBL participants make a slightly higher number of high confidence errors than the other two conditions;

Chapter 10

Discussion of Double-Loop Results

The results from the previous two chapters show that transfer performance is only improved when time is sufficient to allow for both involvement in model building and experimentation. That is, transfer performance is highest in MB relative to the MBL and MR conditions. In fact MBL and MR appear to have similar transfer success.

A simple conclusion from this result is that time is essential for the participant to build transferable knowledge from the simulation case study. MB has more time allocated than MBL and MR; hence transfer success is higher in MB. However, this ignores the lack of evidence to support the prediction that model builders, MB and MBL, should have a positive correlation between attitude change and transfer success. In fact, MB and MR data support this correlation, but MBL does not.

To draw some conclusions and implications from these results this chapter reflects on the differences between both double-loop and single-loop learning and discusses the differences from the real world simulation studies. It is necessary to include single-loop learning results, i.e. attitude change, as it is inherently linked to double-

loop learning (i.e. two loops of learning incorporating attitude change and transfer).

10.1 Differences in learning between conditions

In order to explain the differences in double-loop learning between conditions it is necessary to consider the transfer success, attitude change and approach to experimentation taken by the participants. Table 10.1 summarises the results for double-loop, single-loop and experimentation variables across the three conditions by rank. For example, in the first column the MB condition is given the rank of one indicating that MB achieved the highest level of transfer success; MR and MBL both have the rank two indicating that the conditions have similar transfer success, but that it was less than MB. The sixth column in Table 10.1 states if the data support any correlation between attitude change and transfer. The final two columns in Table 10.1 summarise some details about typical approaches to experimentation in the conditions. The column labelled 'most learning' indicates the approach that aided participants learning the most. For example, in MB scrutiny of results was correlated with transfer success. The column labelled 'focus' indicates the type of scenarios simulated and the focus of scrutiny. For example, MB participants both focussed on familiar factors (variables) such as resource utilisation and unfamiliar factors discovered during building or the radiology scenario (which was less familiar than resource utilisation factors, but more familiar than those discovered during model building).

The model building condition experienced the highest transfer success and provides evidence of a correlation with the resource utilisation and process variation attitude variables. The three hour condition clearly provides the participants with the time to explore the solution space thoroughly; both the familiar and unfamiliar aspects of the problem were explored. However, the results for the approach to experimentation demonstrated that it was those participants that considered the

results in detail (and hence simulated fewer scenarios) that learnt the most. Participants who simulated a large number of scenarios (and hence scrutinized results less) learnt less possibly due to the confirmation bias mechanism outlined in the single-loop learning discussion (section 7.2.1) or a lower ability at analytical problem solving. The results also show that scrutiny of results was higher for those participants who found it harder to identify that inter-arrival times must be modelled in more detail (or at least chose to include it in the model at a later model building stage). It was speculated, backed up by some limited quantitative data, that this increased scrutiny was again an indication of the discovery and novelty mechanism increasing reflection as discussed in section 7.1.2. Section 10.2 discusses learning through the novelty mechanism in more detail.

In the MR condition there are three outcomes to consider. Firstly, attitude change in the correct direction about resource utilisation was to a greater extent than in the MBL condition, but correct attitude change about process variation (ElimVar) was similar, i.e. attitude change was higher in the more familiar direction than the less familiar. Secondly, a correlation exists between transfer success and attitude change about resource utilisation. Lastly, although attitude change about the trade-off between resource utilisation and performance was similar in MB, the MR participants typically have less transfer success.

Table 10.1 also highlights that MR participants typically investigate experimental factors that are most familiar to them, i.e. those that manipulate the utilisation of resources. Participants who investigate the most scenarios typically experience higher attitude change that in turn was related to higher transfer success. This suggests that a quick cycle of experimentation followed by validation was more beneficial for double-loop learning than extra scrutiny of results (leading to less scenario coverage). However, time pressure and the need to incorporate verification and validation (V&V) into experimentation made it difficult for all MR participants to

explore this area of the solution space exhaustively. Exploration of the unfamiliar factors of the problem also suffered due to time pressure i.e. there was less attitude change (although as a caveat see the discussion in section 7.1 about regression to the mean) and transfer about the process variation in radiology diagnostics as well as less identification of new variables for experimentation, relative to the MB condition. One way of summarising this result would be that MR participants learn about the model (and problem) via experimentation; hence, the more scenarios they see the better. This approach stands in contrast to the MB participants who benefitted from more scrutiny of results.

Turning to MBL it is notable that transfer success is similar to MR. Instead MBL differentiates itself from the other two conditions by making more high confidence errors in the reasoning questions and a lack of a relationship between attitude change and transfer.

In comparison the single-loop results were also similar between MBL and MR, except for two results. Firstly, MBL appears to experience slightly less attitude change about managing resource utilisation ($TradeUtil$). Secondly, although overall attitude change in $ElimVar$ is similar it was speculated, in section 7.1.2, that a difference may lie in the mechanisms for change in attitude (single-loop learning) and understanding (double-loop learning). Specifically the model building conditions (MBL and MB) are discovering unfamiliar factors in the case study problem that have an influence on system performance during model building (e.g., process variation stemming from radiology usage or prioritisation of emergencies). The mechanism described in the single-loop results also offers an explanation of the difference in double-loop results found between MBL and the MB and MR conditions. In MB the discovery of unfamiliar and novel factors appears to lead to more scrutiny in experimentation and hence a correlation between attitude change and transfer. In MBL experimentation is limited hence all that is seen is a heuristic inflation of

attitude due to the novelty of the factors discovered in model building. The consequence is a lack of a correlation between attitude change and transfer and an overinflated confidence.

After some reflection it was found that the best way to consider and understand the differences and mechanisms for learning outlined in this section was to construct a simple qualitative model of the dynamics of learning across the conditions. The following section provides the resultant model and a detailed description.

10.2 A model of learning in the experiment

A hypothesised mechanism for learning across the conditions is illustrated by the simple model in Figure 10.1: a stock and flow diagram (sometimes referred to as levels and rates) of the importance (attitude) a decision maker places in the impact a factor (i.e. the radiology department) has on performance and the deeper transferable knowledge (understanding) they have about it. Stocks or levels are illustrated as boxes. Flows or rates are illustrated as hour glasses with arrows and represent the rate of change of the stock. For example, the greater the (constructive) experimentation rate with a model the more one's stock of understanding builds up. The arrows represent influence and feedback; delays are indicated by double lines across the arrows. For example, understanding influences the refining rate of an attitude; however, there may be a delay in building sufficient understanding for this to take effect on the refinement rate.

The first stage of the model in Figure 10.1 is a novelty means importance heuristic. Factors, such as radiology or time dependent arrivals in the case study model, and their influence on system performance may have been discovered during model building; hence they have an element of novelty to a decision maker. Novelty inflates the interest and importance of the factor above what it would have been. This can be a good thing as an inflated attitude towards a factor of the system

Table 10.1: Summary of difference across conditions

	Transfer		Correct Attitude Change				Experimentation	
	Success	OvConf	MaxUtil	TradeUtil	ElimVar	Corr	Most Learning	Focus
MB	1	2	1	1	1	Yes	Scrutiny	All
MBL	2	1	2	2	2	No	Scrutiny	Fixed
MR	2	2	2	1	2	Yes	More scenarios	Most familiar

Notes: 1 = highest in group. Ranks of the same value indicate that groups are tied.

OvConf = overconfidence. A rank of 1 indicates that the condition had the most high confidence errors.

Corr = correlation between attitude change and transfer.

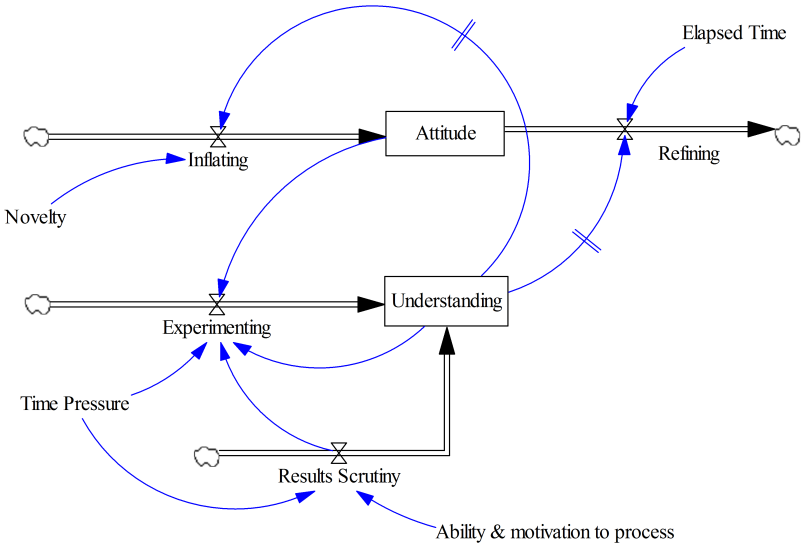


Figure 10.1: Novelty Heuristic Related to Learning

should translate into the intention to experiment with that factor using simulation. The MBL condition, however, is limited in time (i.e. limited scenarios) so the positive attitude does not have an opportunity to influence the stock of understanding through experimentation. On the other hand the MB participants, discovering the same novelty, have less time pressure and are able to experiment as they see fit. Scrutiny of the results over a number of scenarios led to a greater understanding as well as a refinement of attitude (i.e. a correlation between attitude change and transfer performance).

The described attitude refining nature of experimentation may also be a possible explanation of the increased number of high confidence errors MBL participants make in their transfer answers. One way to think of the effect of the novelty heuristic is as model builder's pride. Participants may not be thinking in terms of the problem, but have gained confidence simply because of the novelty introduced via model building. Experimentation acts as the mediator to this 'pride' and, given some time (illustrated by the delays in Figure 10.1), refines the confidence in MB participants. The opposite may be true for the model reuse participants. If model reuse implies automatic distrust of the model (the so called not invented here syndrome (Robinson et al., 2004; Sutcliffe, 2003)) then perhaps experimentation refines confidence upwards to a more realistic level.

Of course, Figure 10.1 should only be viewed as an illustrative model of the novelty heuristic and it's link to time pressure and delay in double-loop learning. Additional factors should also be considered. Firstly, the ability and motivation of the participant will affect the effectiveness and rate of results scrutiny (Chaiken et al., 1989). Secondly, the rate that the attitude is refined may be influenced by elapsed time; for example, attitudes accompanied by little understanding may decrease quite rapidly over time (see section 10.4 for a discussion of implications). The model also leaves out many important factors that affected the participants in

the experiment. For example, when the type of experimentation was purely self-confirmatory - scenarios that demonstrated the participants were obviously correct - then this reduced participants understanding.

10.3 Experiment versus Real World Decision Making

The experiment is a simplification of a real simulation study and results should not be generalised and used as expectations for every model reuse or model building study. This section discusses some of the key differences between the experiment and real simulation studies and the implication when interpreting double-loop results.

10.3.1 Level of Experimentation

The MBL condition provides only enough time to simulate three improvement scenarios. In real simulation studies there may be time to simulate a greater number. Thus, decision makers may possess (and remember) a greater number of examples to assist transfer of insights. This would tend to agree with the studies from the transfer of learning literature. The greater the number and variety of examples the more likely that transfer will occur (Gick and Holyoak, 1987). While further work is required to understand this area in more detail the experimental design and results do indicate some underlying mechanisms at work during model building and experimentation. Firstly, it is hypothesised that model building can help decision makers identify factors outside of their mental model that may be important in influencing system performance. Secondly, it is hypothesised that experimentation is required to refine understanding and correct self confidence in reasoning.

10.3.2 Type of Experimentation

The approach to experimentation used in the research has been framed from an informal position. Participants are given a number of pre-defined scenarios (in the

case study these are presented as options suggested by knowledgeable workers from the system). Participants are also free, in the MB and MR conditions, to choose their own additional scenarios. An alternative approach sometimes adopted in real world DES studies is to use an efficient experimental design; for example, a 2^k factorial or Latin hypercube design may be used to reduce the number of scenarios necessary to estimate the effect of a factor (see Law (2007) for a detailed review). The choice of factors to include in the experimental design may, of course, be biased towards those with which the decision maker is most familiar. However, an experimental design may reduce the learning problems exhibited by those participants who preferred to only simulate scenarios that showed they were correct. The downside might be that the experimental design approach (and similarly a heuristic search approach) may reduce scrutiny of results.

If novel factors are included in the experimental design, either at the choice of the modeller or as a result of decision makers identifying them in model building, then this might also mitigate time pressure issues with scrutinising results.

10.3.3 Model Complexity

The choice of system and its subsequent simplification were made for three reasons. Firstly, so that there was a learning curve for participants i.e. this was something new. Secondly, so that students could understand the system and model in the time available. Thirdly, the system was chosen so that the model could be constructed within a reasonable time frame.

Real world simulation models are likely to have higher complexity than the model used in the experiment. For example, a model of a semi-conductor fabrication process may have an extremely large number of inter-connected activities and resources all subject to process variation. An alternative explanation of the MBL results is therefore that the complexity of the model was too low for the benefits of

involvement in model building to take effect. For example, in a large process decision makers may be more likely to think in terms of local optimisation and fail to recognise (possibly delayed) effects further downstream. Simply building the model may help improve this understanding.

Reusable models may also contain a large number of inter-connected activities and resources subject to process variation; for example, a model of a full hospital (Günal and Pidd, 2010). Again the model used in the experiment may lack sufficient complexity for decision making trust issues to really take hold. If these trust issues can be overcome, the experiment does indicate an underlying mechanism for learning. It is hypothesised that decision makers are most likely to pursue experimentation with factors with which they are most familiar. This would seem important regardless of the complexity of the model.

10.3.4 Background Knowledge of Decision Makers

A key factor effecting transfer is the background knowledge of the decision maker (Gick and Holyoak, 1987). The participants in the experiment lacked familiarity with the system. Thus a proportion of their cognitive processing power was put to work on building familiarity with the system while building understanding about DES in general and A&E system performance. Participants may also expend effort attempting to work out the experimental hypothesis (Field and Hole, 2003); in other words students may expend effort second guessing what the researcher is up to. Clearly this is not the case in real world studies.

To an extent the increased mental effort is compensated for by the simplified case study problem; however, decision makers within a real system may find it easier to concentrate on understanding the problem and hence improve transfer likelihood.

Empirical evidence also suggests the opposite can happen: increased background knowledge of decision makers can lead to transfer difficulties. Bakken et al. (1994)

tested for transfer of insights between analogous System Dynamics models using students and experts. They found that students performed better as the experts would 'take the actions they would have taken in real life' rather than experiment with different approaches to see the effect; Bell and O'Keefe (1995) might call this hypothesis fixation.

10.3.5 Choice of Transfer Concepts

The concepts tested for transfer were chosen because they were felt to be important underlying behavioural factors found in real world systems typically studied using DES. That is, the performance of many queuing systems is uncertain due to activities and resources being subject to process variation over time. There was also some evidence from the manufacturing sector that managers/decision makers found the topics of process variation and resource utilisation difficult to understand (Suri, 1998). These concepts, resource utilisation in particular, may benefit from a more experimentation orientated approach. Other concepts found in real world problems may benefit from a model building approach. For example, a participant could be given a problem containing a small part of a larger system and asked if they think a local optimisation would improve overall performance. Transfer success would be achieved if the participant points out that improvement depends on factors/activities further upstream and their interconnectedness with the local optimisation.

10.4 Implications

This thesis aims to contribute to the debate about model reuse and model building. Specifically if time is limited in a simulation study then how do our modelling choices affect decision maker learning? This thesis explores two possible approaches: building a model and limiting experimentation or increased experimentation through model reuse. The main findings are summarised in Table 10.2, under these head-

ings, along with some pragmatic advice to assist modellers working on real world simulation projects. This section considers these implications.

1. Increased experimentation through reuse
• *Hypothesis*: Decision makers will focus on factors with which they are most familiar;
• *Hypothesis*: Variety and number of scenarios explored assists validation and learning;
• *Advice*: Advocate planning for model validation and use;
• *Advice*: Document decision maker predictions before experimentation;
2. Build model and limit experimentation
• *Hypothesis*: Involvement in model building inflates importance of novel factors;
• *Hypothesis*: Experimentation and scrutiny of factors is necessary to convert attitude into understanding;
• *Advice*: Extend validation to include validation of decision maker understanding;

Table 10.2: Summary of Generated Hypotheses and Advice

10.4.1 Increased experimentation through reuse

There are two hypotheses that require attention regardless of model complexity or approach to experimentation that is taken. The first is that decision makers, expert or novice, may be inclined to focus on factors that are already felt to be important (e.g. resource utilisation). The second is the approach to validating model results, by scrutiny of individual scenarios or by quick interactive simulation, may influence learning. In the experiment quick interactive use of the simulation model was the most effective at increasing the degree of double-loop learning.

The first hypothesis may not be an issue in many studies as the factors felt to be important may indeed be the most important in determining the performance of the system of study. A conceptual modelling stage for model reuse may help with the overhead of validation. Here modellers would encourage planning of how the model is to be used in validation and experimentation. Regardless of the validation approach adopted, it would seem that some benefit could be gained from understanding the decision maker's expectations about performance by asking them for predictions at this stage (Rouwette et al., 2010; Richmond, 1997). Simulation results different from decision maker expectations could receive the most attention within a simulation project report or validation/experimentation workshop to aid learning.

10.4.2 Building a model and limiting experimentation

The main model building hypothesis generated by this research is that involvement in model building may help inflate the importance of factors, particularly novel factors discovered during model building, affecting system performance within a decision maker's mental model. In this experiment limiting experimentation led to attitude change only and typically the greatest number of high confidence errors in the transfer scenarios. That is, there appears to be single-loop learning, but little to no double-loop learning about the concepts measured.

In the case of single-loop learning the simulation modeller may be concerned about the longevity of the attitude; i.e. will the attitude translate into implementation actions or will they simply fizzle out. For example, Robinson (1994) discusses a study where DES suggested an alteration to a manufacturing process. The recommendation was initially incorporated into implementation plans, but disappeared shortly afterwards. After repeating the explanation the recommendation was reincorporated, but again disappeared after some time. One explanation of the issue in Robinson's case study may be that the decision maker has an initial positive

attitude, but with little deep understanding. Thus after a short time the decision maker may struggle to remember why the recommendation was a good idea! Of course, this may only be problematic when implementation of results will happen after a prolonged duration.

Single-loop learning may be even more problematic when the manager involved is not the key decision maker. The transfer of insights into implementation actions may be more problematic if the manager involved in the problem struggles to explain the solutions to the key decision makers.

Balci (1994) defines 10 stages of verification and validation in the lifecycle of a simulation study. The last of these is presentation validation, that is, justifying that the simulation results are interpreted, documented and communicated with sufficient accuracy. The results of the MBL condition emphases the importance of this process, but also the difficulty in achieving a successful outcome. It would seem that some guidance to help maximise the longevity of the knowledge and improve the chances of implementation is needed.

Chapter 11

Conclusions

11.1 Introduction

This thesis has explored differences in how and what decision makers learn when involved in simulation studies where models are built and simulation studies where models are reused. In general it has been a test of the *high involvement hypothesis*: the proposition that high involvement of decision makers in model building leads to high levels of learning. This test is undertaken using an experimental approach comparing attitude change and transfer of learning in simulation model building and model reuse processes.

This final chapter firstly presents a summary of the research objectives and main findings. These findings are used to outline the contribution to theory of the current study. This is focussed on the value of model building and model use in learning and the debate about the value of model reuse. As the current study is only one possible test of the high involvement hypothesis, this is followed by a discussion about the limitations of the results. Next, further work to address the limitations, test the hypotheses generated by the current study and refine the approach taken are outlined. Final comments discuss the views of model building and (re)use by

simulation modellers.

11.2 Summary of Research Objectives

Chapter 2 provides a rationale for this thesis by reviewing modelling literature dating back to the late 1960's. The literature suggest a high involvement hypothesis: that decision makers learning may be aided substantially when they are involved in model building. Although there is a lack of empirical evidence to support this view, the message is repeated quite often by experienced modellers and academics.

This message, it would appear, is at odds with the benefits of reusing a simulation model, for example a generic model of an A&E department, to aid decision making. The reuse, if possible, is desirable as it leads to reduced cost as there is no longer the need to build the model. This, theoretically, benefits the decision maker as the time saved can be used for extra experimentation.

At a general level the objective of this thesis is to test the high involvement hypothesis against the benefit of increased experimentation offered by reusing a model when total project time is limited. If the high involvement hypothesis is true and substantial in effect size, it should be possible to recreate the effect in a laboratory setting. This is done via a comparison of model building (with limited and extended experimentation) against a model reuse process. In addition to the general test of the high involvement hypothesis this research aims to:

1. Empirically identify mechanisms that aid decision maker learning within model building;

2. Empirically identify mechanisms that aid decision maker learning within model reuse/experimentation;

3. Empirically identify mechanisms that inhibit learning from DES models;

Learning is defined using the theory of action framework developed by Argyris and Schön (1996). The main assumption of this framework is that individuals have a predefined theory of effective performance. Argyris and Schön use this assumption to introduce the concept of learning at two levels: single and double-loop.

At the first level of learning the individual's theory-of-action remains the same; however, their actions and attitudes towards achieving those objectives may change. This is called single-loop learning and can be interpreted as short term learning to solve an immediate business problem. In the experiment this is inferred through attitude change variables. The second level of learning refers to a change in the individuals governing variables. That is, it defines a deeper level of understanding and reflection. This is defined as double-loop learning; a change in the long term decision making and behaviour of the individual. In the experiment this is inferred through transfer of learning to analogous problem scenarios and a correlation between attitude change and transfer.

11.3 Summary of main findings

This section presents the main findings of the research. These are categorised by the objectives detailed in Section 11.2.

Objective 1: Empirically identify mechanisms that aid decision maker learning within model building

- Model building aided participants in identifying/discovering novel factors since they saw the impact on performance as it was added to the model. This discovery inflated the importance of the factor, but participants may not fully understand why the factor was important.
- Experimentation and scrutiny (i.e. careful examination of results) was necessary to convert the inflated attitude into a deeper understanding that was

transferable.

Objective 2: Empirically identify mechanisms that aid decision maker learning within model reuse/experimentation

- Model reuse participants focused experimentation on the factors with which they are most familiar;
- A quick cycle of experimentation followed by validation helped the MR participants learn about the model and thus helped attitude change and transfer.

Objective 3: Empirically identify mechanisms that inhibit learning from simulation models

- The overhead of face validation in model reuse made it difficult for some participants to investigate everything they wished to. This was especially problematic if the participant spent a lot of time carefully examining results as opposed to interacting with the model.
- Participants appeared to learn most about resource utilisation during experimentation. Attitudes changed relatively little in the model building condition where experimentation was limited.
- If a novel factor was discovered in model building it was necessary to incorporate the factor into multiple simulation scenarios to aid deeper understanding. The model building participants who could not do this (i.e. MBL) could not transfer the concept.
- The approach participants used for experimentation affected learning. Some participants preferred to be correct all the time (i.e. their scenarios improved performance over the base model). These participants typically reinforced attitudes counter to good system performance rather than build new knowledge.

- No substantive evidence was found to support the so called 'not invented here syndrome' affecting model reuse. However, this may be due to lack of sufficient model complexity to trigger the mechanism. Some evidence was found to support differences in the process of model use between the conditions to aid learning.

11.4 Contributions to Simulation Theory

This section discusses the contributions of the current study to the simulation literature. Section 1.4 listed four areas that where contributions were expected to be made. This section discusses each under the corresponding headings:

1. An empirical comparison of the impact of model building and reuse on learning;
2. Insight into how generic models aid learning;
3. Generation of focussed hypotheses on learning mechanisms;
4. A framework for future research;

11.4.1 An empirical comparison of the impact of model building and reuse on learning

Learning from the interaction of model building and use

The first contribution of this study is the attempt to draw out some of the differences between the learning mechanisms found in simulation model building and (re)use. Often it is assumed that model building produces a deeper transferable knowledge of the system under study than model use or reuse (Alessi, 2000). Although simulation analysts tend to think this is true, there is a lack of empirical evidence within the simulation literature to support this view. In fact, studies of managerial learning from simulation have typically focused on experimentation. System Dynamics

research has tended to perform and analyse experiments of learning (e.g. Rouwette et al., 2004; Bakken et al., 1994; Paich and Sterman, 1993; Sterman, 1989*b*). While discrete-event simulation has presented discussions on the use of simulation as a gaming tool (Van der Zee and Slomp, 2009), the benefits of visual interactive simulation (Belton and Elder, 1994) and experiments of learning (Bell and O'Keefe, 1995). There are also data rich examples of learning during group model building interventions in System Dynamics (Rouwette, 2003) and finite element simulation (Thomke, 1998). However, it is difficult to draw out where (i.e. model building or use) learning occurs during the interventions. This study generated the hypothesis that involving decision makers in model building aids identification of novel factors for experimentation. Of course, this proposed mechanism should be subject to further empirical test (see Section 11.7 for a discussion), but it provides a plausible explanation for the 'hunch', expressed by many simulation authors, that model building aids learning.

Involvement in model building does not guarantee learning

The second contribution of the study is that at times model building plus unconstrained experimentation resulted in participants forming incorrect attitudes (i.e. attitudes that were detrimental to system performance). This result was quite unexpected for two reasons. Firstly, the experiment used student participants. It is well documented that novices are easier to manipulate than experts (Yaniv, 2004; Kantowitz et al., 1992). Thus it was expected that involvement in model building and experimentation would shift their attitudes quite significantly (as opposed to political factors, group pressures and self-confidence mediating attitude change). Secondly, the results from experimentation were communicated in a very simple manner. Improvements in performance are highlighted in green and reductions in performance are highlighted in red (see Section 4.5.2 for an overview of the results

screen). Thus even those participants with the lowest ability should have been able to discern detrimental from beneficial action.

Other studies have also shown that students can experience difficulties in learning from simulation. Paich and Sterman (1993), for example, conclude that when dynamic complexity is high students need to become modellers not merely players in a simulation game. Of course, the participants in this experiment were not 'model builders' in this sense. However, participants still had substantial involvement in the conceptualisation of the model that should have improved transparency. The learning difficulties expressed by the model builders in this study illustrate that we should be careful of automatically falling back on such conclusions.

11.4.2 Insight into how generic models aid learning

Increased experimentation time can aid learning

The first contribution of this thesis, to the debate about model reuse, is a demonstration of the potential benefit to learning if experimentation time is increased. A lot of modern research effort has gone into developing generic models that can be reused by decision makers in different, but 'similar' organisations (e.g. Günal and Pidd, 2010; Fletcher et al., 2007). If the technical and conceptual difficulties of model reuse can be overcome, as they were in the studies listed, then this study supports the hypothesis that increased experimentation time can aid decision makers' learning about a system factor or relationship between factors (such as the relationship between resource utilisation and performance). This seemed to work best when the participants used a quick cycle of experimentation followed by face validation (i.e. not delving too deeply into model output).

This contribution is somewhat counter-intuitive, as many simulation modellers may expect involvement in model building to yield more learning than model reuse. Sections 7.1.1 and 10.2 speculated two possible explanations for this difference.

Firstly, learning about concepts such as resource utilisation may benefit from a classic Kolb learning cycle (Kolb, 1984) (hypothesing,experimenting, experiencing and reflecting) and DES experimentation fits more clearly into this framework. Secondly, model building may provide decision makers with hints of relationships between variables, but interaction with experimentation was needed to help participants thoroughly test these ideas. The extra time spent on experimentation allowed (most) participants to work out experimentation strategies to improve their learning (i.e. double-loop) and build transferable knowledge. For example, experimentation provided an opportunity for participants to run 'validation scenarios'. The model building participants could use early scenarios to test uncertainties they had about the model and reserve the later scenarios for exploring potential relationships once they gained sufficient confidence; the model reuse participants might find that, after an initial review of results, it was better to run many scenarios and quickly review the results to improve their understanding of, firstly, the DES model and, secondly, potential relationships between variables that they found during this process.

Bounded rationality and the focus of experimentation

A second contribution to the debate about model reuse concerns the approach to experimentation taken by decision makers. It is typically accepted in Operational Research / Management Science literature that decision makers are boundedly rational (see Pidd, 2009, for a discussion). Thus it is expected that the initial focus of decision makers will be on the factors that decision makers are aware of and feel are of most importance. In this study the participants who learnt the most from reusing the A&E model did so by a quick cycle of experimentation followed by (face) validation. As expected the initial focus of this effort was typically on the factor(s) with which the participants were most familiar. However, this effort of validation meant that model reusers were less likely to get the time to learn about relatively

novel factors.

The overhead of model verification and validation when reusing a model is discussed frequently in the simulation literature. In fact, model reuse is often believed to be as or more time consuming than building the model from scratch (Robinson et al., 2004). This research adds to this debate with the hypothesis that, in reuse studies, the face validation effort required by boundedly rational decision makers may hinder the search of the solution space. This must be weighed against the benefit of focussed experimentation: namely that decision makers may learn quite a lot about the focus of their experimentation. There are, of course, some limitations to this contribution; these are discussed in Section 11.5.

11.4.3 Generation of focussed hypotheses on learning mechanisms

Section 10.4 provides a detailed discussion of the hypotheses generated by this research. In summary these are:

- Decision makers will focus on factors with which they are most familiar;
- Variety and number of scenarios explored assists validation and learning;
- Involvement in model building inflates importance of novel factor;
- Experimentation and scrutiny of factors is necessary to convert attitude into understanding;

11.4.4 A framework for future research

Model building procedure

The study used an innovative approach to building simulation models with participants in a learning laboratory. Other studies of learning have mainly focussed on experimentation (see Rouwette et al., 2004, for a review of the SD literature). Where model building has been involved in the experiment it has been of a qualitative manner (e.g. Shields, 2001, 2002).

The approach used in this study could quite easily be adapted for building quantitative simulation models in alternative experiments, if desired. In addition to producing the models and materials, researchers would need to tailor the introduction to simulation to be consistent with their context. Limitations might be the availability of simulation software and simulation approach. For example, the software used in this study was Simul8 - a discrete-event package. Simul8 was particularly useful for the early simulation models that were - very quickly - built and run in front of participants. Researchers would need to consider if their package provides the same development speed. In particular, researchers in a SD context may wish to only perform direct building for basic simulation education.

Measuring Single-Loop and Double-Loop Learning

Other studies measuring learning from simulation have either focussed on attitude change (Rouwette, 2003) or transfer (Bakken et al., 1994). This study provides a combined approach in order to explore the differences in single-loop and double-loop learning in a laboratory setting. The approach can be adapted to measure learning in different simulation contexts (e.g. System Dynamics or Agent Based Systems) or different concepts within discrete-event simulation. The downside is that the process does require a substantial amount of effort to setup. This includes developing attitude questionnaires, transfer scenarios and models as well as a lengthy pilot phase.

11.5 Limitations

Although the thesis provides a contribution to the debate about model reuse and model building there are a number of limitations to the work. This section discusses these limitations and their possible effect on the results. These relate to the results from one experiment, the sample size employed, the use of students, the use of indi-

viduals rather than groups and approach to measuring learning. These limitations are considered here.

11.5.1 External Validity versus Results from one Experiment

An objective of the research is to test the claims found in the simulation literature that involving clients in model building is beneficial for learning. The research does indeed test this claim, but it is only a single test of this central hypothesis. The external validity of the results can only be tested through further experiments (Mook, 1983; Berkowitz and Donnerstein, 1982; Gadlin and Ingle, 1975). In particular, it would seem necessary to rigorously test the hypotheses generated by this research. Section 11.7 discusses further work to meet this requirement. Additionally, as the research is the result from one experiment, there are also limitations to the procedure used in the experiment. Specifically the choice of experimental materials (e.g. model and transfer questions) may affect learning.

Model Context, Objective and Display

All experimental results relate to the use of a single model. The accident and emergency (A&E) model is a service system with a tough service level target (98% of patients must spend less than four hours in A&E). Three limitations to the single test approach are the effects of model context (e.g. healthcare versus manufacturing), objective function and visual display. A different choice of model from a different domain may have influenced learning in other ways. For example, in a service system the majority of variability is driven by arrivals to the system (e.g. patients). If a model was selected from manufacturing, for example, where much of the variation is derived from cycle times, breakdowns, repair times and so on, then perhaps this would influence attitudes and aid/hinder transfer differently. It may be easier for participants to learn about process variation and its effect on system performance

by watching variable machine breakdown and repair times than the equivalent in the A&E model.

The choice of context may also be important due to possible differences in objectives. In the case study, as in real life, the NHS face a challenging time based target coupled with high variation in arrivals and the need for value for money from resources. This high service level requirement meant that the resource utilisation 'problem' (i.e. its relationship to performance) was a constant thorn in the side of participants; they always ran into it during experimentation. A model from a different context may not have an objective that puts them at odds with the resource utilisation problem in such an aggressive manner. Thus the participant may be able to focus on other aspects of the problem - aspects in which experimentation is not as important as model building.

Lastly, it is difficult to determine if the choice of model display influenced learning. The A&E model is built with a simple process flow worldview. This was done for two reasons. Firstly, it was assumed participants would be familiar with a general process flow diagram. Secondly, it simplified the model building procedure. There was some evidence that the approach helped learning. For example, if participants made alterations to upstream activities in the process (e.g. created a new triage room to reduce queues) then they saw the problem move along to downstream activities. However, it is difficult to know if having a model showing the layout of the A&E would have aided other types of learning. For example, a layout model may have demonstrated how radiology could be blocked to A&E patients due to the variable arrival of, and usage by, non-A&E patients.

Ordering of Transfer Scenarios

Participants were always presented with the transfer scenarios in the same order. This went from close transfer, in a healthcare context, to far transfer, in call centre

and manufacturing contexts (although it appears that participants found scenario eight very similar to the case study problem). It is uncertain if the ordering of these questions had an effect on results. For example, scenarios four and six both considered the resource pooling problem in healthcare and call centre contexts respectively (analogous to the pooling of cubicles in the experiment). It is difficult to say if participants would be more likely to gain successful transfer on scenario four if they answered scenario six beforehand. One theory might be that when participants answer the reordered scenario four they would have two examples of the problem in different contexts to draw from; a healthcare example from the case study and a call centre example from scenario six. Transfer of learning theory would suggest this aids transfer success (see Barnett and Ceci, 2002; Bassok, 2003, for review of relevant research). However, this may require that the participant recieves feedback on their answers before proceeding to the next scenario.

Timing of Credibility Measurement

The pilot experiments suggested that both model builders and model reusers provided the same ratings of how credible the model appeared to them. Persuasion theory (Chaiken et al., 1989) suggested that differences were likely to be found in sufficiency thresholds of verification and validation (V&V) that participants held. That is, involvement in model building should increase the importance of the problem to participants and thus their V&V sufficiency threshold. Higher self-confidence was then expected in the model builders, as the level of verification and validation that a model needed to have undergone to satisfy model building participants was higher than that of the model reusers. This also seemed to agree with the behaviour seen in the pilot experiments. However, this prediction was not supported by the results.

One explanation is that the self-confidence of the participants was measured

immediately after the experiment. It is difficult to know what the results would be if the credibility measurement was undertaken after a delay (e.g. a day). This may generate the expected behaviour. If it is true that model reusers followed a less rigorous V&V then when they think back about the model they may not remember why they were so confident. However, the model builders who are expected to have followed a more rigorous and exhaustive questioning of the model may be able to remember why they were so confident.

The participants in the MR condition also seemed to learn most effectively when they performed a quick cycle of experimentation and V&V. These participants may indeed be as confident in their assessment of the model as the model builders. Again changing the timing of the measurements may be helpful in getting round this limitation. One possible approach would be to include a credibility measurement after initial introduction to the model and one after experimentation.

11.5.2 Sample Size

The sample size was a constraint on the research due to the funds available and the time required to run the individual experiments. Inference procedures are aided through the complementary use of non-parametric procedures (i.e. the exact version of the Mann-Whitney test), bootstrapping and effect sizes. In particular, the exact version of the Mann-Whitney test helps test for possible bias effects simply due to the randomisation (see Appendix B.1.3 for a detailed explanation), the bootstrap procedures provide sufficient levels of statistical power to enable rigorous control of multiple comparisons and the trade-off between Type I and Type II errors), while the effect size helps understand differences in outcomes of the test. Ultimately, however, more confidence in the results could be gained by employing a larger sample size. Power analysis suggests that a sample size of around 40 participants in each condition (i.e. a total of 80 in a comparison and a total sample size of 120) is required

to detect a medium size effect (Hair et al., 2006). Using the current payment scheme this would total £1400 in costs and, assuming that one participant is seen at a time, 330 hours of lab time.

11.5.3 Use of Students

The discussion chapters for single-loop learning (Section 7.4) and double-loop learning (Section 10.3) both discuss the limitations of employing novice decision makers in such an experiment; these points will not repeated here. There are two additional points to consider.

The first point is that there is a threat to the internal validity of the experiment due to the chance of collusion between students. That is, a student may tell a friend who is also participating in the experiment about the details and what they should do. While this can never be fully controlled (in any psychology study that does not run completely in parallel) the current study did take several steps to attempt to minimise it. The first step was to randomly assign students to conditions. Thus there is a chance that the colluding students are in separate conditions and experience a different simulation process. One weakness is with this alone is that the collusion may still 'average out' or mask the differences between conditions. Thus the second step was to offer a cash prize for the best performance (highest transfer with a tie breaker of best explained reasoning) as an incentive not to collude. The last step was to remind each student after the experiment that there was the potential for them to win some extra money and that collusion reduced their chances of winning.

The second point is similar to the limitation of Section 11.5.1. It refers how social psychologists build up confidence in the external validity of their theories. In particular, although the use of students is a non-controversial approach in experimental social psychology, many experiments are carried out (by different researchers) in attempts to explain, for example, inconsistencies in the results of other studies.

Hence the use of students in psychology is less problematic as a body of research and theory is built up that one can have more confidence in. The simulation model building aspect of this research has little to no experimental studies for comparison in the simulation literature. Thus a further limitation is that it is difficult to know if this experiment caused unusual behaviour in the students.

11.5.4 Individuals versus Groups

In many OR interventions the decision making will be performed by groups. It has been repeatedly demonstrated that double-loop learning is particularly inhibited by group learning systems (Argyris, 1992; Argyris and Schön, 1996). Specifically group decision making is much more defensive due to tacit and automatic face saving mechanisms. There is some anecdotal evidence from the experiments that some participant's learning was inhibited by face saving mechanisms during experimentation. This led to avoidance of scenarios with negative results and a bias towards confirming obvious results. It is difficult to know the effect of combining groups and DES on the learning approach taken. One theory, largely originating from the SD literature on group model building (e.g. Senge, 1990; Rouwette, 2003; Rouwette et al., 2010), is that the model building, model use and discussion around it allows the groups to build consensus on issues and expose hidden assumptions of the group, subgroups and individuals.

11.5.5 Measurement of learning

The variables used in the experiment to measure single and double-loop learning are best thought of as proxy measures. That is, they are indirect measures providing an indication of the degree of single-loop and double-loop learning that occurs.

One limitation might be in the measurement of transfer success to infer a degree of double-loop learning (i.e. changes in reasoning). Measurement is operationalized

using simple analogous scenarios specially constructed to include cues to aid transfer that are undertaken immediately after the case study. Clearly in the real world 'new scenarios' would not be as simple as those encountered in the experiment and transfer cues may not be obvious to decision makers. This view would tend to agree with the transfer of learning literature which has found little transfer mathematics and physics text books to real life analogous problems (Bassok, 2003).

Although transfer success may be more difficult to achieve with real problems, an alternative indicator of double-loop learning may be the approach that decision makers use to structure a problem. In particular it would seem that the search for specific transfer cues, even if they are not found, indicates learning. For example, within a DES context, this may be asking questions about process variation or considering the inter-connectedness of the process. This would give an indication that the decision maker understands the importance of these concepts within queuing type systems.

Turning to single-loop learning a limitation may be the exclusion of the two other variables from the theory of planned behaviour: subjective norm (similar to peer pressure) and perceived behavioural control (i.e. does the participant believe they have any influence over implementing the changes to the system). Although learning is measured at an individual level, a measure of subjective norm may be relevant as participants were required to interact with a researcher during model building and use. Hence the presence of a researcher may have affected some participant's behaviour; for example, participants may have felt more pressure to find scenarios that improved performance and reduced their investigation of interesting scenarios that reduced performance.

Similarly perceived behavioural control may have influenced participant's attitudes. For example, participants may have felt that agreeing a radiology usage policy with the rest of the hospital was particularly difficult or not possible at all. Thus

even though the scenario improved performance the participant's attitude did not change as the participant did not feel it was worth pursuing due to implementation issues.

The questionnaire could, of course, easily be adapted to include subjective norm or perceived behavioural control. The downside is that the experiment would now measure additional variables and the questionnaire length would be extended. One way around this problem would be to limit the experiment to focus on one or two behaviours as opposed to the three measured here.

11.6 Further Work to Address Limitations

There are several variations to the experimental procedure and materials that can be made to help address the limitations discussed in the previous section. These can be categorised into the following three areas:

- Manipulating the learning and transfer domains;
- Extending the experiment to include groups;
- Inclusion of real managers and decision makers;

This section discusses each of these categories and the different variations to the experiment that can be made within each.

11.6.1 Manipulating the learning and transfer domains

A simple way to test if the learning domain has an effect on learning is to repeat the experiment, but this time use an alternative model and case study. For example, many manufacturing models exist in the DES literature. A suitable case study could be selected and used as the basis for the experiment. In this simple manipulation of the learning domain it is necessary that the objective of the new simulation

model be highly similar to that of the A&E model. That is, it should be a target associated with the speed at which entities can travel through the queuing process. This approach redefines far transfer: this now shifts to call centres and healthcare.

A more complicated experimental design is appropriate if the difficultly in the achievement of the objective of the model is manipulated; difficulty in achievement might be varied, for example, by performance targets in the A&E model that are easy and hard to meet.

Possible approaches to the design of the experiment are the addition of a new condition(s) or introduction of repeated measures. A repeated measures design would firstly subject the participant to one objective, measure the outcomes and then repeat for the alternative objective. To avoid bias the participants are randomly divided into two groups. One group receives the hard to achieve objective treatment first while the other group receives the easy to achieve target first. The benefit of this approach is that the sample size required - relative to a between groups design - is reduced. This approach could also be adapted for the visual display and complexity issues.

11.6.2 Extending the Experiment to Include Groups

There are three issues that must be considered if the experiment is altered to include groups of three or four participants: dependency within the data, sample size and the scope of the attitude questionnaire.

The first issue relates to a grouping of participants introducing dependency into the collected data. That is, the attitude(s) of an individual in one group is more likely to be similar to the others in their group than an individual from another group. This means that the standard statistical tests, that assume independent observations, cannot be used to assess difference between individuals. One approach to deal with this issue is to study outcomes only at the group level, as groups can be

considered as independent. The drawback is that vast amounts of data related to the individuals in the groups are ignored. One solution is to employ a complicated statistical model in the analysis of the data. Multi-level linear models appear to be a viable approach for this type of issue (Field, 2009). However, there are no hard and fast rules on the sample size that is required for this approach. The consensus seems to be the larger the sample size the better (Field, 2009). Thus the second issue with extending the experiment to include groups relates to the sample size required. Given this, researchers would need to consider the feasibility of the approach.

The last issue is that the attitude questionnaire should be extended to incorporate additional or all of the aspects of the theory of planned behaviour. This would produce a questionnaire similar to the one used by Rouwette (2003). Hence the revised questionnaire would be longer and contain participants perceptions of how important the thinking of others in the group is to them and their perception of whether they have control over the outcomes. As the experiment already contains a large number of variables, researchers should consider the benefits of removing some variables if more are introduced. Additional variables potentially mean a greater number of multivariate outliers that require removal; given the more complicated analysis, and hence larger sample size that would be required for groups, this is something that researchers should try to avoid.

Although not directly addressing a limitation of the research, researchers may also wish to qualitatively analyse the content of the experiment, i.e. explore the process of group validation and (re)use of a discrete-event simulation model. This would require the careful development of a coding scheme, or adaptation of an existing coding scheme from the literature, to determine the proportion of time groups spent discussing, for example, the validity of an assumption in the model and the actual system. This might also show up interesting differences in the behaviour of groups given exposure to model building or reuse. Of course, this modification

would require substantial setup and piloting of the coding scheme.

11.6.3 Inclusion of real managers and decision makers

One option to overcome the limitations of using students is to include managers in the experiment. The problem, of course, is sample size and access to the participants for an appropriate amount of time. One option is use executive MBA students (as in Bakken et al. (1994)) and incorporate the experiment into the MBA course. However, as this may still yield smaller sample sizes than the current experiment it might be prudent to conduct more focussed experiments in an attempt to minimise the variance and number of dependent variables; section 11.7 provides some alternatives. It is also still questionable if this approach yields real 'decision makers' as though the executive MBA students have decision making experience it may have little to do with a case study in an experiment.

The other option is to explore the use of DES with managers and workers in a particular application domain. For example, healthcare clinicians from an A&E department could be involved in the development and reuse of a simulation model of an A&E department. Again, it is likely that sample size will be small, if access is even possible, so all results would need to be interpreted carefully and simplifications to the experiment or a more focussed experiment should be considered.

If access for experimentation is not possible, then survey work could be undertaken to aid the construction of future experiments. The questionnaires could be similar to the reasoning scenarios incorporated into post-test of the current experiment. It would seem prudent to keep these scenarios focussed and relevant to the participants; for example, all within a healthcare domain.

Of course, this would still require access and substantial piloting. The latter may be particularly problematic when the target population is quite small, as the relevant participants may be largely used up in the pilot.

11.7 Further Work Testing Generated Hypotheses

11.7.1 Test of Novelty Heuristic

The first area of investigation suitable for testing is that involvement in model building can help decision makers identify novel factors (i.e. factors not felt to be important) influencing system performance. This work should also be undertaken with experiments that look at disconfirming the proposed mechanism that novelty initially inflates the importance of a factor. Table 11.1 and Table 11.2 outline potential experimental designs for tests of the novelty identification and heuristic hypotheses respectively.

Table 11.1: Novelty Identification Test

Setup and Pilot Work
• Identify factors of a problem that are novel to participants
Independent Variables
• Process (Build / Reuse)
Dependent Variables
• Factor identified as important
Predictions
• Higher proportion of model building group identifies novelty.

A large amount of setup would be required for the experiments. This would involve writing a case study and testing pilot participants' initial mental models for aspects and factors that are novel. Methods for doing this could involve structured questionnaires, interviews, testing (via transfer type scenarios) and observation of

simulation usage. It seems advisable that automated records of simulation usage are generated in some manner so that these can be analysed easily.

Table 11.2: Novelty Heuristic Test

Setup and Pilot Work

- The setup work undertaken for the novelty identification test could be reused.
- Preparation of simulation experiments and transfer scenario.

Independent Variables

- Novelty introduction (Yes/No)
- Experimentation (Yes/No)

Dependent Variables

- Attitude towards factor
- Transfer of learning to analogous problem context

Predictions

- Novelty (YES) and experimentation (YES) - highest transfer
- Novelty (YES) and experimentation (NO) - highest attitude (i.e. novelty inflates importance)
- Novelty (NO) and experimentation (NO) - low attitude; low transfer
- Novelty (NO) and experimentation (YES) - attitude and transfer higher than control group.

The two experiments could be combined into one single experiment. However, as they involve the tests of focussed predictions it seems sensible to separate them. Reducing the number of independent and dependent variables is likely to ease the pressure on sample size requirements and data analysis issues - there should be less variation in results and a lower number of multivariate outliers.

11.7.2 Test of Variety of Experimentation

The second area that requires testing is the hypothesis that the variety in and number of scenarios explored in experimentation can aid learning from a reused model. This experiment would be slightly simpler than the one conducted in this study as there would be no need to include a model building element. Furthermore, the choice of scenarios for experimentation could be completely controlled. Table 11.3 outlines a potential approach to test the hypothesis.

The main choice facing researchers is how to define variety. As an example consider resource utilisation in a call centre. The closest parallel to the experiment is to run different scenarios for the same model demonstrating the performance of the call centre and call operator utilisation; e.g., altered shift patterns, reallocation between shifts and increases and decreases in resource. Low variety is then defined as only a subset of this list; e.g. the effect of increasing or decreasing resource. As an alternative, researchers may wish to define variety as examples from a different context. For example, a participant could be shown the impact of reallocating resource on the performance of a call centre and A&E department. Low variety would then be defined as the results from only one context.

The results of both approaches would seem to be useful for understanding the impact of variety on learning from simulation. One drawback is that the controlled approach to experimentation limits what can be drawn out about the process of face validation of reused models. The experiment could be modified so that variety is controlled by providing different levels of options for participants to use in experimentation. Specifically, some groups have access to a large variety of options and others have access to only a few. The trade-off is similar to the current study: extra exploration of process, but introduction of variability in outcomes.

Out of the approaches to define variety listed the development of multiple models is likely to require the most time to setup and pilot. However, it may also provide

insight into how decision makers involved in multiple simulation projects learn. An interesting modification may not be change context, but to remain in the same context and model a different system (it would seem plausible that this is more likely to occur in the real world) and then assess transfer.

Table 11.3: Variety of Experimentation Test

Setup and Pilot Work

- Creation of simulation model(s)
- Identification of a suitable attitude and transfer concept
- Creation and trial of measurement instruments

Independent Variables

- Number of scenarios (High/Low)
- Variety in scenarios (High/Low)

Dependent Variables

- Attitude towards factor
- Transfer of learning to analogous problem context

Predictions

1. Number (HIGH) and variety (HIGH) - highest transfer
2. Number (LOW) and variety (HIGH) - medium attitude, transfer higher than in 3.
3. Number (HIGH) and variety (LOW) - high attitude, but low relative transfer to 2.
4. Number (LOW) and variety (LOW) - lowest transfer

11.8 Further Work to Refine the Experiment

11.8.1 Refining the Model Building Procedure

The model building procedure required a great deal of effort from the researcher. This could be reduced as much of the work was scripted, i.e. videos could be used to

introduce the students to discrete-event simulation, the results screens, the model and smaller versions of the model during construction. In fact, this approach was considered during the development of the experimental procedures. Due to project time constraints the only feasible option out of those listed was to produce videos for an introduction to simulation. It was decided that this would not be done as it did not provide opportunities for participants to ask questions. However, it was noted that during the explanation questions from participants were minimal. Hence it appears quite feasible to introduce an automated element to the heavily scripted parts of the experiment.

The option is also available to remove all interaction with the researcher from the experiment. This would require the development of 'virtual simulation modeller'. In essence a central console allowing the participants to control the simulation model, view results and view the conceptual model (i.e. assumptions and simplifications). The console would also need to allow the participant to ask two types of question. The first type of question would be an enquiry about if an assumption or simplification can be removed from the model. If the simplification is not removable (for example, due to lack of data) then the virtual modeller would report the reasons why it cannot be removed. The virtual modeller could also give generic advice on sensitivity analysis if a participant is unhappy about an assumption. The second type of question the console would need to handle is explanation of model logic. Thus some sort of simplified representation *of the model* must be incorporated into the console.

While this is a fairly straightforward programming problem it will require some effort to setup. Thus some interaction with the researcher may be preferred in the procedure.

11.8.2 Adding a Control Condition

Replications of the experiment may wish to include a control condition where participants are not exposed to simulation. The motivation would be to look beyond the relative differences between conditions and determine the benefit of simulation. In addition to increased sample size the experimental design would need to consider the effects of subconscious and conscious processing of the case study. To understand this consider the extension to the experimental design displayed in Figure 11.1.

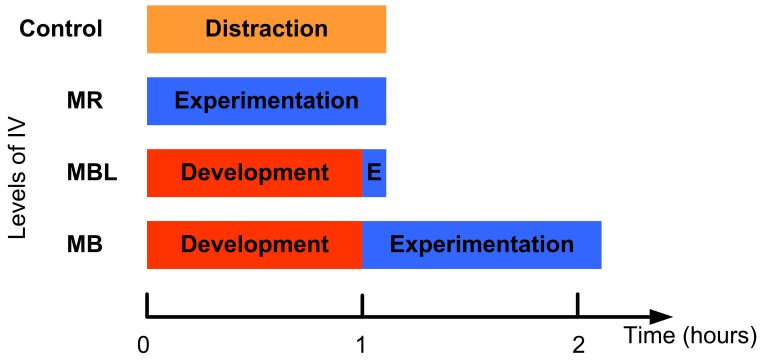


Figure 11.1: Control Condition

A typical approach in psychology is to run the control condition for an equal time to the other conditions. Thus participants complete the pre-test and post-test questionnaires at the same time as the MR and MBL conditions. To fill the time in-between questionnaires it is usual for a distraction to be put into place e.g. participants watch a video of an unrelated topic.

In a similar experimental design, concerned with decision making, Dikjksterhuis et al. (2006) showed the distracted participants made better and more consistent decisions when they were distracted. This was due to the relative power of unconscious processing to conscious processing. It is uncertain how this would be affected by a more complex process suitable for simulation. Future work might compare the relative trade-offs between distraction and conscious processing of the case study (i.e. a workshop style intervention) and simulation modelling.

11.8.3 Adding a Transfer Hint

Some of the transfer of learning literature uses the idea of a transfer hint to help assess the difference between learning and recognising the similarity between the 'training' and transfer scenarios (e.g. Novick and Holyoak, 1991; Reed, 1987). If the hint grouping has higher transfer performance than the no-hint grouping then it can be concluded that the participants have learnt something, but have difficulty in recognising the similarity between the 'training' and transfer scenarios.

This approach could also be used to test the learning that MBL and MR participants experience. For example, this could test if experimentation aids abstraction of the concepts learnt in model building.

11.8.4 Comparing Methods for Experimentation

A further option to refine the experiment is to consider the type of experimentation that is undertaken. Section 10.3.2 noted that a potential difference between the experiment and real simulation studies is that model use was framed in an informal manner. The alternative is to consider the impact of formal methods for experimentation, i.e. experimental design (potentially used to construct a meta-model) and heuristic search of scenarios. Section 10.3.2 speculated that the formal methods could reduce the bias of participants to stick to scenarios with which they were most familiar and that proved them to be obviously correct. The drawback was speculated to be a reduction in the scrutiny of results.

The simplest way to explore this area would be to focus only on model use as opposed to the interaction between model building and the type of experimentation (reducing the sample size required and the number of variables measured). One approach would be to run the experiment with three conditions. These would consist of informal experimentation, experimental design and heuristic search. The second and third of these conditions, i.e. the formal methods, may not even include any

interaction with the model, but rather a briefing of the model followed by a summary of the results that can be viewed and explored by the participant (e.g. viewing the best results, checking the robustness of solutions and drilling into the data). An alternative (or addition depending on sample size available) is to permit informal experimentation using the model after the experimental design and heuristic search groups have reviewed results. Measurement could focus on solution understanding (possibly via a written test) and what actually happens in the experiment (e.g. scrutiny of results, choice of informal experimentation or time taken to reach a decision). Similar to this thesis the objectives would be gain some insight into the mechanisms that aid or inhibit learning as opposed to determining the 'best approach' to experimentation.

The inclusion of a heuristic search condition means that some consideration must be given to the fair comparison between conditions. That is, heuristic search tends to focus on quantitative variables only; such as numbers of resources available, capacity of conveyer belts, routing percentages (for reference, OptQuest that ships with Simul8 (2009) Professional only includes examples of these types of variables in the instructions). Changes in the policies and process flow of a system, common in informal experimentation, could be incorporated in a similar manner to that used in experimental design: binary variables where a zero value represents that the change is not present and one represents that the change is present. This type of automated search may be beyond the standard 'optimizers' that ship with commercial simulation software (such as OptQuest) hence researchers could either choose to write an updated search engine or focus only on quantitative variables (the allocation of resources across shifts might be a suitable choice).

11.9 Final Comments

Much of the simulation literature reviewed for this research took the view that it is critical to involve decision makers (e.g. a manager) in model building in some manner to aid learning. Anecdotally, it is noted that a similar message was repeated in several discussions with simulation and general OR practitioners/academics at conferences and workshops. Indeed, it seems that the current study offers some support for this view in the form of the novelty heuristic.

One issue with taking this view, however, is that it appears that to some extent the benefits of involving decision makers in experimentation are not considered. In the current study experimentation aided learning and at times *complemented* learning gained in model building. One interesting difference was that (most) model reusers used experimentation to learn about the model and hence about the problem. This process is different from model building, but equally valid. This outcome should also make us optimistic about the reuse of large generic simulation models (e.g. a full hospital) for decision making. Clearly, in these cases development time and hence cost is a big issue; however, reusing a model to gain experimentation time may be beneficial for learning even if time is lost to validation. Future research may wish to consider how experimentation (and validation) can be structured and supported to aid learning from reusable models.

References

Abdulla, R., Garrison, B., Salwern, M., Driscoll, P. and Casey, D. (2004), Online news credibility, *in* M. Salwern, B. Garrison and P. Driscoll, eds, 'Online News and the Public', 1 edn, Routledge, chapter 5.

Ajzen, I. (1988), *Attitudes, personality and behavior*, Open University Press, Buckingham.

Ajzen, I. (1991), 'The theory of planned behavior', *Organizational Behavior and Human Decision Processes* **50**(2), 179–211.

Ajzen, I. (2001), 'Nature and operation of attitudes', *Annual Review of Psychology* **52**, 27–58.

Alessi, S. (2000), Building versus using simulations, *in* M. Spector and T. Anderson, eds, 'Intergrated and Holistic Perspectives on Learning Instruction and Technology', Kluwer Academic Publishers, pp. 175–196.

Andrews, L. and Gutkin, T. (1991), 'The effects of human versus compter authorshop on consumers' perceptions of psyhcological reports', *Computers in Human Behaviour* pp. 311–317.

Argyris, C. (1992), *On Organisational Learning*, Blackwell Publishers, Cambridge, Massachussetts.

Argyris, C. and Schön, D. A. (1996), *Organisational Learning II: Theory, Method and Practice*, Addison Wesley.

Arons, H. S. and Boer, C. A. (2001), 'Storage and retrieval of discrete-event simulation models', *Simulation Practice and Theory* **8**(8), 555–576.

Bakken, B., Gould, J. and Kim, D. (1994), Experimentation in learning organisations: A management flight simulator approach, *in* J. Morecroft and J. D. Sterman, eds, 'Modelling for Learning Organisations', Productivity Press, Portland.

Balci, O. (1994), 'Validation, verficiation, and testing techniques throughout the life cycle of a simulation study.', *Annals of Operations Research* **53**, 121–173.

Balci, O. and Nance. N, E. (2008), Accomplishing reuse with a simulation conceptual model, *in* S. Mason, R. Hill, L. Mönch, O. Rose, T. Jefferson and J. W. Fowler, eds, 'Proceedings of the 2008 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc, Piscataway, New Jersey, pp. 959–965.

Balci, O. and Nance, R. (1985), 'Formulated problem verification as an explicit requirement of model credibility', *Simulation* pp. 76–86.

Balci, O. and Ormsby, W. F. (2007), 'Conceptual modelling for designing large-scale simulations', *Journal of Simulation* **1**(3), 175–186.

Barnett, S, M. and Ceci, S, J. (2002), 'When and where do we apply what we learn? a taxonomy for far transfer.', *Psychological Bulletin* **128**(4), 612–637.

Bassok, M. (2003), Analogical transfer in problem solving, *in* J. Davidson and R. Sternberg, eds, 'The psychology of problem solving', Cambridge University Press, Cambridge, pp. 343–369.

Bell, D., Mustafee, N., de Cesare, S., Lycett, M. and Taylor, S. (2008), 'Ontology engineering for simulation component reuse', *International Journal of Enterprise Information Systems* **4**(4), 47–61.

Bell, P. and O'Keefe, R. (1995), 'An experimental investigation into the efficacy of visual interactive simulation', *Management Science* **41**, 1018–1038.

Belton, V. and Elder, M. (1994), 'Decision support systems: Learning from visual interactive modelling', *Decision Support Systems* **12**, 355–364.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing.', *Journal of the Royal Statistical Society. Series B (Methodological).* **57**(1), 289–300.

Benjamini, Y. and Hochberg, Y. (2000), 'On the adaptive control of the false discovery rate in multiple testing with independent statistics', *Journal of Educational and Behavioral Statistics.* **25**(1), 60–83.

Berkowitz, L. and Donnerstein, E. (1982), 'External validity is more than skin deep: Some answers to criticisms of laboratory experiments', *American Psychologist* **37**(3), 13.

Bonate, P. (2000), *Analysis of Pretest-Posttest Designs*, Chapman & Hall//CRC, Washington, D.C.

Brase, G., Fiddick, L. and Harries, C. (2006), 'Participant recruitment methods and statistical reasoning performance', *The Quarterly Journal of Experimental Psychology* pp. 965–976.

Brown, N. and Powers, S. (2000), Simulation in a box: a generic reusable maintenance model, *in* J. A. Joines, R. R. Barton, K. Kang and P. A. Fishwick, eds, 'Proceedings of the 2000 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc, Piscataway, New Jersey, pp. 1050–1056.

Butterfield, B. and Metcalfe, J. (2006), 'The correction of errors commited in high confidence', *Metacognition and Learning* **1**, 69–84.

Cassidy, W. (2007), 'Online news credibility: An examination of the perceptions of newspaper journalists', *Journal of Computer-Mediated Communication* **12**, 478–498.

Chaiken, S., Liberman, A. and Eagly, A. (1989), Heuristic and systematic processing within and beyond the persuasion context, *in* U. JS and B. JA, eds, 'Unintended thought', Guilford Press, New York, pp. 212–252.

Chaiken, S. and Maheswaran, D. (1994), 'Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgement', *Journal of Perseonality and Social Psychology* pp. 541–546.

Chaung-Stein, C. and Tong, D. (1997), 'The impact and implication of regression to the mean on the design and analysis of medical investigations', *Statistical Methods in Medical Research* **6**, 115–128.

Churchman, C. and Schainblatt, A. (1965), 'The researcher and the manager: A dialectric of implementation', *Management Science* pp. 69–87.

Chwif, L. and Barretto, M. (2003), Simulation models as an aid for the teaching and learning process in operations management, *in* S. Chick, P. Sanchez, D. Ferrin and D. Morrice, eds, 'Proceedings of the 2003 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey.

Cochran, J., Mackulack, G. and Savory, P. (1995), 'Simulation project charactoristics in industrial settings', *Interfaces* pp. 104–113.

Cohen, J. (1990), 'Things i have learned (so far)', *American Psychologist* **45**, 1304–1312.

Conover, W. and Iman, R. (1982), 'Analysis of covariance using the rank transformation', *Biometrics* **38**(3), 715–724.

Dhir, K. (2001), 'Enhancing management's understanding of operational research models', *Journal of the Operational Research Society* pp. 873–887.

Dikjksterhuis, A., Bos, M. and Van Baaren, R. (2006), 'On making the right choice: The deliberation-without-attention effect', *Science* **311**, 1005–1007.

Field, A. (2009), *Discovering Statistics Using SPSS*, 3 edn, Sage Publications Ltd.

Field, A. and Hole, G. J. (2003), *How to Design and Report Experiments*, Sage Publications Ltd, London.

Fletcher, A., Halsall, D., Huxham, S. and Worthington, D. (2007), 'The dh accident and emergency department model: a national generic model used locally', *Journal of the Operational Research Society* **58**(12), 1554–1562.

Fletcher, A. and Worthington, D. (2009), 'What is a 'generic' hospital model? - a comparison of 'generic' and 'specific' hospital models of emergency patient flows', *Healthcare Management Sciencel* **12**, 374–391.

Ford, D. and Sterman, J. (1998), 'Expert knowledge elicitation to improve formal and menal models', *System Dynamics Review* pp. 309–340.

Forrester, J. (1994), Policies, decisions, and information sources for modeling, *in* J. Morecroft, J.D.W; Sterman, ed., 'Modeling for Learning Organisations', Productivity Press, Portland, Oregon.

Fowler, M. (1999), *Refactoring: Improving the design of existing code*, 1st edn, Addison Wesley, London.

Fraser, J., Smith, P. and Smith Jr, J. (1992), 'A catalog of errors', *International Journal of Man-Machine Studies* **37**, 265–307.

Gadlin, H. and Ingle, G. (1975), 'Through the one-way mirror: the limits of experimental self-reflection', *American Psychologist* **30**(10), 6.

Gaziano, C. and McGrath, K. (1986), 'Measuring the concept of credibility', *Journalism Quarterly* **63**, 451–462.

Gick, M. and Holyoak, K. (1983), 'Schema induction and analogical transfer', *Cognitive Psychology* pp. 1–38.

Gick, M. and Holyoak, K. (1987), The cognitive basis of knowledge transfer, *in* S. Cormier and J. Hagman, eds, 'Transfer of Learning: Contemporary Research and Applications', Academic Press, London, pp. 9–46.

Gick, M. and Holyoak, S. (1980), 'Analogical problem solving', *Cognitive Psychology* pp. 306–355.

Größler, A. (1998), Structural transparency as an element of business simulators, *in* 'Proceedings of the 1998 System Dynamics Conference', p. 39.

Größler, A. (2004), 'Don't let history repeat itself - methodological issues concening the use of simulators in teaching and experimentation', *System Dynamics Review* pp. 263–274.

Größler, A., Maier, F. and Milling, P. (2000), 'Enhancing learning capabilities by providing transparency in business simulators', *Simulation and Gaming* pp. 197–229.

Günal, M. and Pidd, M. (2006), Understanding accident and emergency department performance using simulation, *in* L. R. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol and R. M. Fujimoto, eds, 'Proceedings of the 2006 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc, Piscataway, New Jersey, pp. 446–452.

Günal, M. and Pidd, M. (2009), 'Understanding target-driven action in emergency department peformance using simulation', *Emergency Medical Journal* **26**, 723–727.

Günal, M. and Pidd, M. (2010), 'Dghpsim: Generic simulation of hospital performance', *Transactions on Modeling and Simulation (TOMACS)* p. [forthcoming].

Hair, J., Black, W., Babin, B., Anderson, R. and Tatham, R. (2006), *Multivariate Data Analysis*, 6th edn, Pearson Education International, New Jersey.

Johnson, T. and Kaye, B. (1998), 'Cruising is believing?: Comparing internet and traditional sources on media credibility measures', *Journalism and Mass Communication Quarterly* **75**(2), 325–340.

Johnson, T. and Kaye, B. (2002), 'Webelievability: A path model examining how convenience and reliance predict online credibility', *Journalism and Mass Communication Quarterly* **79**(3), 619–642.

Kantowitz, B., Hanaowski, R. and Kantowitz, S. (1992), 'Driver acceptance of unreliable trafic information in familar and unfamilar settings', *Human Factors* **39**, 164–176.

Kaylani, K., Mollaghasemi, M., Cope, D., Fayez, S., Rabadi, G. and M, S. (2008), 'A generic environment for modelling future launch operations - gem-flo: a success story in generic modelling', *Journal of the Operational Research Society* **59**(10), 1312 – 1320.

Kelly, C. and Price, T. (2005), 'Correcting for regression to the mean in behaviour and ecology', *The American Naturalist* **166**(6), 700–707.

Kelman, H. and Hovland, C. (1953), "reinstatement' of the communicator in delayed measurement of opinion change', *Journal of Absnormal and Social Psychology* pp. 327–335.

Kiousis, S. (2001), 'Public trust or mistrust? perceptions of media credibility in the information age', *Mass Communication and Soceity* **4**(4), 381–403.

Kolb, D. (1984), *Experiential Learning. Experience as The Source of Learning and Development*, Prentice Hall, Englewood Cliffs.

Krause, A. and Pinheiro, J. (2007), 'Modeling and simulation to adjust p values in presence of a regression to the mean effect', *The American Statistician* pp. 302–307.

Lahteenmaki, S., Toivonen, J. and M, M. (2001), 'Critical aspects of organizational learning research and proposals for its measurement', *British Journal of Management* **12**, 113–129.

Lane, D. (1994), Modeling as learning: A consultancy methodology for enhancing learning in management teams, *in* J. Morecroft, J.D.W; Sterman, ed., 'Modeling for Learning Organisations', Productivity Press, Portland, Oregon.

Lane, D. (1995), 'On a resuragance of managment simulations and games', *Journal of the Operational Research Society* pp. 604–625.

Law, A. (1993), 'A forum on crucial issues in the simulation modeling', *Industrial Engineering* pp. 32–36.

Law, A. M. (2007), *Simulation Modelling and Analysis*, 4th edn, McGraw-Hill International, Boston.

Lunneborg, C. (2000), 'Data analysis by resampling: Concepts and applications'.

Malak, R. and Paredis, C. (2004), Foundations of validating reusable behavioural models in engineering design problems, *in* R. G. Ingalls, M. D. Rossetti, J. S. Smith and B. A. Peters, eds, 'Proceedings of the 2004 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 420–428.

Meyer, P. (1988), 'Defining and measuring credibility of newspapers: Developing an index', *Journalism Quarterly* **65**, 567–572.

Miller, G. A. and Chapman, J. (2001), 'Misunderstanding analysis of covariance', *Journal of Abnormal Psychology* **110**(1), 40–48.

Monks, T., Robinson, S. and Kotiadis, K. (2009), Model reuse versus model development: Effects on credibility and learning, *in* M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds, 'Proceedings of the 2009 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey.

Mook, G. D. (1983), 'In defense of external invalidity', *American Psychologist* **38**, 379–387.

Musselman, K. (1990), Position statement on interactive modelling, *in* O. Balci, R. Sadowski and R. Nance, eds, 'Proceedings of the 1990 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey.

Neuhauser, J. (1976), 'Business games have failed', *The Academy of Management Review* pp. 124–129.

Nisbett, R. and Wilson, T. (1977), 'Telling more than we can know: verbal reports on mental processes', *Psychological Review* pp. 231–259.

Novick, L. and Holyoak, K. (1991), 'Mathematical problem solving by analogy', *Journal of Experimental Psychology: Learning, Memory and Cognition* **17**, 398–415.

O'Neill, K. (1992), Real-time tragedies: a simulated commons learning laboratory, *in* 'Proceedings of the 1992 System Dynamics Conference', pp. 25–34.

Ozdemirel, N. (1991), Measuring the user acceptance of generic manufacturing simulation models by review of modeling assumptions, *in* B. L. Nelson, W. D. Kelton and G. M. Clark, eds, 'Proceedings of the 1991 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 419–417.

Paich, M. and Sterman, J. (1993), 'Boom, bust and failures to learn in experimental markets', *Management Science* **39**(12), 1439–1458.

Paul, R. J. and Taylor, S. J. E. (2002), Improving the model development process: what use is model reuse: is there a crook at the end of the rainbow?, *in* E. Yücesan, C. H. Chen, J. L. Snowdon and J. M. Charnes, eds, 'Proceedings of the 2002 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 648–652.

Petty, R., Briñol, B. and Tormala, Z. (2002), 'Thought confidence as a determinant of persuasion: The self-validation hypothesis', *Journal of Personality and Social Psychology* **82**(5), 722–741.

Petty, R. and Cacioppo, J. (1986), 'The elaboration likelihood model of persuasion', *Advances in Experimental Social Psycholology* **19**, 123–205.

Pidd, M. (1988), 'From problem structuring to implementation', *Journal of the Operational Research Society* pp. 115–121.

Pidd, M. (1992), 'Guidelines for the design of data driven generic simulators for specific domains', *Simulation* **59**(4), 237–243.

Pidd, M. (2004), *Computer Simulation in Management Science*, 5th edn, John Wiley and Sons, London.

Pidd, M. (2009), *Tools for Thinking*, 3rd edn, John Wiley and Sons, London.

Pidd, M. and Carvalho, A. (2006), 'Simulation software: not the same yesterday, today or forever', *Journal of Simulation* **1**(1), 7–20.

Pierce, N. G. and Drevna, M. J. (1992), Development of generic simulation models to evaluate wafer fabrication cluster tools, *in* J. J. Swain, D. Goldsman, C. C. R and J. R. Wilson, eds, 'Proceedings of the 1992 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 874–878.

Priester, J. and Petty, R. (1995), 'Source attibutions and persuasuion: Perceived honest as a determinant of message scrutiny', *Personality and Soial Psychology* pp. 175–190.

Quade, D. (1967), 'Rank analysis of covariance', *Journal of the American Statistical Association* **62**(320), 1187–1200.

Radnor, M., Rubensein, A. and Tansik, D. (1970), 'Implementation in operations research and r&d in government and business organisations', *Operations Research* pp. 967–991.

Reed, S. (1987), 'A structure-mapping model for word problems', *Journal of Experimental Psychology: Learning, Memory and Cognition* **13**, 124–139.

Reinhard, M. and Sporer, S. (2008), 'Verbal and nonverbal behaviour as a basis for credibility attibution: The impact of task involvement and cognitive capacity', *Journal of Experimental Social Psychology* **44**(3), 477–488.

Richmond, B. (1997), 'The strategic forum: aligning objectives, strategy and process', *System Dynamics Review* **13**, 131–148.

Robinson, S. (1994), *Successful Simulation*, John Wiley and Sons.

Robinson, S. (1998), 'Measuring service quality in the process of a simulaiton study: The customer's perspective', *International Transactions in Operational Research* pp. 357–374.

Robinson, S. (2002), 'General concepts of quality for discrete-event simulation', *European Journal of Operational Research* **138**(1), 103–117.

Robinson, S. (2004), *Simulation: The Practice of Model Development and Use*, John Wiley and Sons.

Robinson, S. (2008), 'Conceptual modelling for simulation part 1: definitions and requirements', *Journal of the Operational Research Society* **59**(3), 278–293.

Robinson, S., Nance, R. E., Paul, R. J., Pidd, M. and Taylor, S. J. E. (2004), 'Simulation model reuse: definitions, benefits and obstacles', *Simulation Modelling Practice and Theory* **12**(7-8), 479–494.

Rocconi, L. and Ethington, C. (2009), 'Assessing longitudinal change: Adjustment for regression to the mean effects', *Research in Higher Education* **50**(4), 368–376.

Rogers, A. (1980), 'Regression towards the mean and the regression-effect bias', *New directions for testing and measurement.* pp. 59–82.

Rogosa, D. (1988), Myths about longitudinal research, *in* K. Schaie, R. Campbell, W. Meredity and S. Rawlings, eds, 'Methodological issues in aging research', Springer Publishing Co, pp. 171–209.

Rouwette, E. (2003), Group Model Building as Mutual Persuasion, PhD thesis, University of Nijmegan.

Rouwette, E., Fokkema, E., Kuppevelt, H. and Peters, V. (1998), 'Measuring marco polis management game's influence on market orientations', *Simulation and Gaming* pp. 420–431.

Rouwette, E., Größler, A. and Vennix, J. (2004), 'Exploring influencing factors on rationality: A literature review of dynamic decision-making studies in system dynamics', *Systems Research and Behavioral Science* **21**(4), 351–370.

Rouwette, E., Korzilius, H., Vennix, J. and E, J. (2010), 'Modeling as persuasion: the impact of group model building on attitudes and behaviour', *System Dynamics Review* p. [forthcoming].

Sargent, R. G. (1996), Some subjective validation methods using graphical displays of data, *in* J. M. Charnes, D. J. Morrice, D. T. Brunner and J. J. Swain, eds, 'Proceedings of the 1996 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 345–351.

Senge, P. (1990), *The Fifth Discipline. The Art and Practice of the Learning Organisation*, Random House, London.

Shields, M. (2001), An experimental investigation comparing the effects of case study, management flight simulator and facilitation of these methods on mental model development in a group setting, *in* 'Proceedings of the 19th Annual International System Dynamics Conference'.

Shields, M. (2002), The role of group dynamics in mental model development: An experimental comparison of the effect of case study and management flight simulator under two levels of faciliation, *in* 'Proceedings of the 20th Annual International System Dynamics Conference'.

Simul8 (2009), 'Simul8 software', Available from `http://www.Simul8.com` [Accessed October 2009].

Sinreich, D. and Marmor, Y. N. (2004), A simple and intuative simulation tool for analyzing emergency department operations, *in* R. G. Ingalls, M. D. Rossetti, J. S. Smith and B. A. Peters, eds, 'Proceedings of the 2004 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 1994 – 2002.

Spiegel, M., Reynolds, P. F. and Brogan, D. C. (2005), A case study of model context for simulation composability and reusability, *in* M. E. Kuhl, N. M. Steiger, F. B. Armstrong and J. A. Joines, eds, 'Proceedings of the 2005 Winter Simulation Conference', Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 437–444.

Sterman, J. (1989*a*), 'Misperceptions of feedback in dynamic decision making', *Organizational Behavior and Human Decision Processes* pp. 301–335.

Sterman, J. (1989*b*), 'Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment', *Management Science* pp. 321–339.

Suri, R. (1998), *Quick Response Manufacturing. A Companywide approach to reducing lead times*, Productivity Press, Portland, Oregon.

Sutcliffe, A. (2003), *The Domain Theory: Patterns for Knowledge and Software Reuse*, 1st edn, CRC Press, London.

Tako, A. (2009), Development and use of simulation models in Operational Research: a comparison of discrete-event simulation and system dynamics, PhD thesis, Warwick Business School.

Tako, A. and Robinson, S. (2009), 'Comparing discrete-event simulation and system dynamics: users' perceptions', *Journal of the Operational Research Society* **60**, 296–312.

Thomke, S, H. (1998), 'Simulation, learning and r&d performance: Evidence from automotive development', *Research Policy* **27**(1), 55–74.

Tocher, K. and Owen, D. (1960), The automatic programming of simulations, *in* 'Second International Conference on Operational Research', pp. 50–68.

Tolk, A. (1999), Non-monotonicities in hla-federations, *in* 'Proceedings of the Spring Simulation Interoperability Workshop', IEEE CS Press.

Tormala, Z., Brinol, P. and Petty, R. (2006), 'When credibility attacksthe reverse impact of source credibility on persuasion', *Journal of Experimental Social Pyschology* pp. 684–691.

Tseng, S. and Fogg, B. (1999), 'Credibility and computing technology', *Communications of the ACM* **42**, 39–44.

Tulving, E. and Kroll, N. (1995), 'Novelty assessment in the brain and long-term memory encoding', *Psychonomic Bulletin and Review* **2**, 387–390.

Urban, G. (1974), 'Building models for decision makers', *Interfaces* pp. 1–11.

Van der Zee, D.-J. and Slomp, J. (2009), 'Simulation as a tool for gaming and training in operations management - a case study', *Journal of Simulation* **3**, 17–28.

Vennix, J., Scheper, W. and Willems, R. (1993), Group model building. what does the client think of it?, *in* 'Proceedings of the 1993 System Dynamics Conference'.

Waern, Y. and Ramberg, R. (1996), 'Peoples perceptions of humand and computer advise', *Computers in Human Behaviour* pp. 17–27.

Wang, W. and Brooks, R. (2011), Improving the understanding of conceptual modelling, *in* S. Robinson, R. Brooks, K. Kotiadis and D. Van der Zee, eds, 'Conceptual modelling for discrete-event simulation', CRC Press, Boca Raton.

Ward, S. (1989), 'Arguments for constructively simple models', *The Journal of the Operational Research Soceity* pp. 141–153.

Wilcox, R. (2005), *Introduction to Robust Estimation and Hypothesis Testing*, 2 edn, Elsevier Academic Press, London.

Willemain, T. R. (1995), 'Model formulation: What experts think about and when', *Operations Research* **43**(6), 916–932.

Yaniv, L. (2004), 'Recieving other people's advice: Influence and benefit', *Organisational Science and Human Decision Processes* **93**, 1–13.

Young, S., Yang and Wang (1992), Enhancing the learning effects of dynamic decision game on systems thinking, *in* 'Proceedings of the 1992 System Dynamics Conference', pp. 847–856.

Appendix A

Design of Experiment Appendix

A.1 Case Study

Saint Specific's Hospital: Accident and Emergency Performance Problems

A.1.1 Introduction

St. Specific's is a busy hospital with around 45,000 patients passing through its Accident and Emergency (A&E) department each year. This high demand means that setting policy and operations to maintain an efficient service level for patients whilst not overloading staff is difficult.

In recent years performance has come under scrutiny at St. Specific's. All A&E departments are required to measure the percentage of patients whose total time in A&E exceeds 4 hours. Hospitals should target that only 2% of patients exceed this time. In the last six months 15% of patients that came to St. Specific's A&E breached this target.

A.1.2 The Problem: How can A&E performance be improved?

Due to current performance, St. Specific's is now under pressure to improve. Management must make some decisions on how A&E operations should be changed to improve their ability to meet the target. A recent consultation with staff has provided several ideas on how this might be achieved. Before these are presented, the following section provides an overview of A&E and its processes.

A.1.3 The A&E Process

Emergency Patients

Patients arrive at A&E either by walking in (75%) or by Ambulance. Patient arrivals are highly unpredictable and demand varies considerably throughout the 24 hour day (see figure 2). Peaks occur in the morning and early evening.

Walk-In patients register at reception and wait for evaluation by a nurse (see figure 3 for a floor plan). Walk-in patients are evaluated as either major emergencies or minor emergencies depending on their condition and prioritised accordingly. In contrast, ambulance patients are registered and evaluated on route to A&E by the attending paramedic. However, ambulance patients may also require resuscitation on arrival. See figure 1 for the process flow and table 4 for percentages of minor, major and resuscitation emergencies.

Patients are then allocated a treatment cubicle, attending nurse, and a doctor. The patient is then treated. If appropriate an x-ray is taken followed by a second treatment session. Once the emergency is over the patient is either admitted to the hospital for further treatment or discharged.

Doctors

Doctors are classed at either a senior or junior level. Both levels of doctor can treat more than one patient at once. That is, the doctors split their time between patients. Senior doctors estimate that they can cope with up to 5 patients at once. Junior doctors estimate they can treat up to 3 patients at once. The length of time that it takes to treat a patient is related to the doctor's level and the classification of the emergency (see Table 1). Table 2 provides details of the doctor staffing rota.

Nurses

Nurses have three responsibilities in A&E:

- Register patients upon their arrival to the A&E.
- Evaluate the level of emergency that a walk-in patient represents.
- Tend to emergency patient care; e.g. cleaning/bandaging injuries or administering prescribed drugs.

It is estimated that on average nurses can perform up to two tasks at any one time. For example they can care for more than one patient at once. Table 2 provides details of the nurse staffing rota.

Treatment Cubicles

Patients require a cubicle before they can be treated by a doctor and a nurse. St. Specifics has 11 cubicles in total. These are only used for minor and major emergencies as patients requiring resuscitation have two additional cubicles outfitted for their level of emergency.

Radiology

A&E makes use of the three X-Ray machines in the radiology ward for minor and major emergencies (see figure 3). However, all three are shared with the rest of the hospital. Non-A&E patient arrival times are highly unpredictable and can vary significantly between 7am and 10pm. Currently it is estimated that on average x-ray machines are utilized 30% of the time between 7am and 10pm (4.5 hours) by non-A&E patients. That is, on average patients from other parts of the hospital arrive at radiology 110 minutes apart during the day (at night x-ray is only used by A&E). If all x-ray machines are in use (by A&E and/or other patients) then the A&E patient remains in their cubicle and 'queues' until it is their turn. Table 3 provides details of x-ray usage.

In addition the radiologists suggest that a percentage of minor patients that junior doctors send to radiology do not require an x-ray.

A.1.4 Your Task

You are a manager at the hospital. You must make a decision about how current performance against the above target can be improved.

A specialist consultancy firm has been brought in to aid you in making your decisions. A consultant will work with you to build a computer simulation model of the A&E process. It is proposed that the options, presented below, are simulated to evaluate their impact on A&E performance. You and the rest of the management team will work with the consultant to do this and draw conclusions on the options. *You do not require any knowledge of computer simulation. The consultant will use the model as you see fit and where appropriate offer advice on how it might be used.*

Improvement Options

A recent consultation with staff has suggested the following options may improve performance:

1. Reallocate a nurse from a quite period to a busy period. Specifically:

Night	Morning	Day	Evening
-1	0	+1	0
0	-1	+1	0
-1	-1	+2	0

2. Add an extra three cubicles, an extra junior doctor on the morning and day shifts and two extra nurses on the day shift.
3. Add an extra three cubicles extra, an extra senior doctor in the day, an extra junior doctor in the morning and two extra nurses on the day shift.
4. Time table x-ray usage with the rest of the hospital. Currently other parts of the hospital have asked for a total of 6 hours (40%) per day: 7 to 9am, 2:30 to 4:30pm and 8 to 10pm. This would mean that A&E would have 2 x-ray rooms at these times and all 3 outside of these times.

A.1.5 MR Task Description

A specialist consultancy firm has been brought in to aid you in making your decisions. Last year the consultancy developed a simulation model of 'St James' A&E'; a hospital that has very similar processes to St Specific's A&E. It is proposed that the model is reused to simulate the options, presented below, to evaluate their impact on A&E performance. You will work with the consultant to do this and draw conclusions on the options. *You do not require any knowledge of computer simulation. The consultant will use the model as you see fit and where appropriate offer advice on how it might be used.*

A.1.6 MBL Predefined Scenarios

A recent consultation with staff has suggested the following options may improve performance:

1. Reallocate a nurse from a quite period to a busy period. Specifically:

Night	Morning	Day	Evening
-1	0	+1	0

2. Add an extra three cubicles, an extra junior doctor on the morning and day shifts and two extra nurses on the day shift.
3. Time table x-ray usage with the rest of the hospital. Currently other parts of the hospital have asked for a total of 6 hours (40%) per day: 7 to 9am, 2:30 to 4:30pm and 8 to 10pm.

This would mean that A&E would have 2 x-ray rooms at these times and all 3 outside of these times.

A.1.7 St Specific's Hospital: Data Sheet

This sheet provides additional data to complement the description of A&E found in the case study.

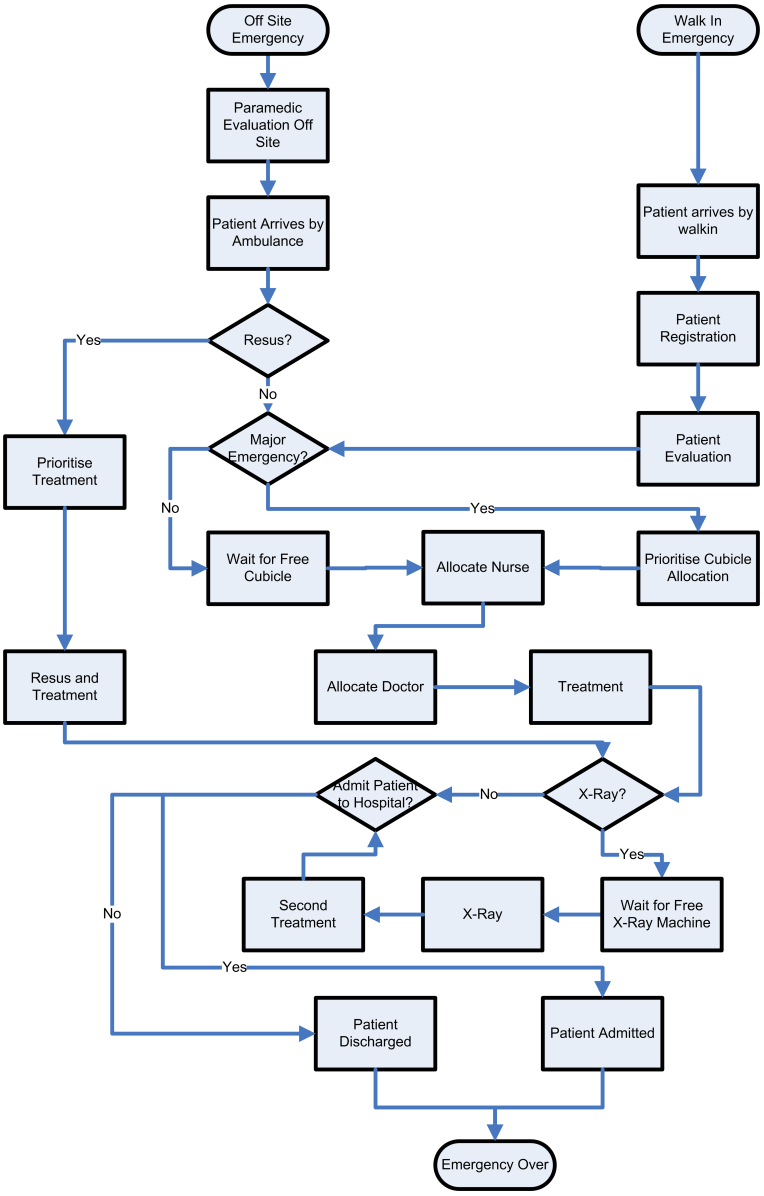


Figure 1: Process Flow for A&E treatment

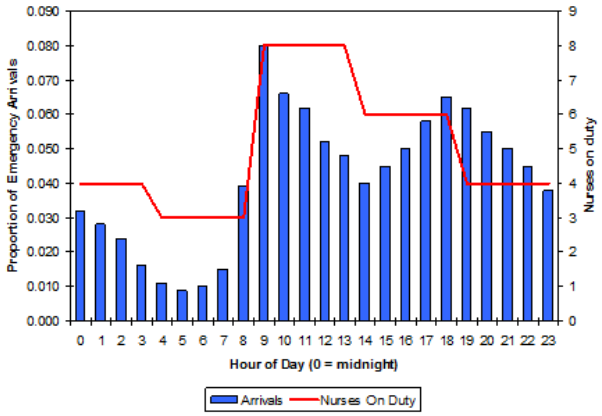


Figure 2: Daily Pattern of Arrivals of Emergency Patients contrasted with Nurse Rota

	Junior Doctor		Senior Doctor	
Emergency	**Average**	**Std Dev**	**Average**	**Std Dev**
Major	50 mins	45 mins	40 mins	25 mins
Minor	35 mins	30 mins	25 mins	10 mins

Table 1: Patient Treatment Times

Shift	**Start**	**End**	**Nurses**	**Sr. Doctors**	**Jr. Doctors**
Night	4am	9am	3	0	2
Morning	9am	2pm	8	1	2
Day	2pm	7pm	6	1	2
Evening	7pm	4am	4	1	1

Table 2: St Specifics Staff Rota

	X-ray %	
Emergency	Junior Drs	Senior Drs
Major	70	70
Minor	40	30

Table 3: Percentage of Emergencies Requiring X-rays

Note: Only general data is available for X-ray times. Average = 33 minutes. Std Dev = 15 minutes.

	Arrival Type	
Emergency	Walk in	Ambulance
Major	40	66
Minor	60	30
Resus	0	4

Table 4: Percentage of Emergencies by Arrival Type

Figure 3: A&E Ward Floor Plan

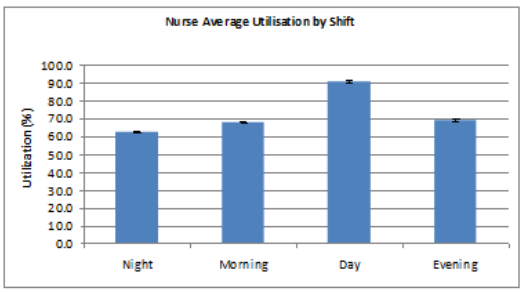


Figure 4: Estimated Nurse Utilisation by Shift Over the last 6 months

A.2 Model Documentation

This section provides documentation for the St. Specifics case study model using the conceptual modelling framework set out by Robinson (2004). Note that it is largely based on the generic models found in Günal and Pidd (2009) and Fletcher et al. (2007); albeit in a slightly simplified form to meet the general project objectives. The information in this section can be used with the written case study and case study data sheet to recreate the model in any DES package.

A.2.1 Objectives

This section details the general project objectives, such as general time scale and analysis requirements, and modelling objectives.

General Project Objectives

- The model must be able to be built within a suitable time frame. A reasonable choice seems to be around 1 hour to 1 hour and 15 minutes.
- The model must be easy to modify for scenarios. If a difficult modification is often suggested by participants then have a prebuilt model on standby.
- Participants can easily differentiate a scenario that has improved performance from a scenario that has reduced performance.
- Participants can compare scenarios using statistics
- Participants can compare scenario results using graphical approaches. E.g. time series, histograms and bar charts.
- Participants can watch the model running.

Modelling Objectives

1. Objective:

- Reduce the percentage of patients breaching the four hour target to 2% over six months.

2. Constraints:

    - No budget figure given, but participants are made aware that resources cost money.

3. Responses to Determine Success:

    - Percentage of patients breaching the four hour target over six months

4. Responses to Determine Reasons for Failure:

    - Time series and histogram of minor and major patient target breaches over six months
    - Bar chart of average nurse utilisation by shift
    - Comparison bar charts for resource utilisation of all resources across scenarios
    - Comparison time series and histogram charts for target breaches across scenarios

A.2.2 Level of Detail

Component	Detail	Include/Exclude	Comment
Patients	Minor/Major/Resus	Include	Minor and major patients are timed as they travel through the system. Resus patient use resources.
	Injury description	Exclude	Not necessary for timings.
	Inter-arrivals	Include	Modelled as time dependent distribution
	Non-A&E	Include	Inter-arrivals to radiology modelled as a distribution.
Doctors	Seniority	Include	The type of doctor a patient gets affects treatment time and the variation in process time. Modelled as resource.
	Multi-tasking	Include	Necessary otherwise large queues build up. Used mini-doctors approach from Günal and Pidd (2006).
	X-ray decision making	Include	Included as a possibility for students to explore.
	Admission decisions	Exclude	Simplify so that this is incorporated into treatment time.
	Fatigue	Exclude	No studies (found) show any impact of workload/fatigue on treatment time.
Nurses	Seniority	Exclude	Assumed largest impact on treatment time was doctors and emergency level.
	Multi-tasking	Include	Same justification as doctors
	Decision making	Exclude	Assume that nurses can correctly triage between minor and major
Diagnostics	Radiology	Include	Modelled as resources and time delay
	Blood tests	Exclude	Cannot include all diagnostic tests due to time constraints.
	Resus x-ray	Exclude	Modelled as time delay
Porters	Trolley resources	Exclude	Simplify problem for students and reduce time of experiment. Note that porters are not included in Günal and Pidd (2006), but they are included in Fletcher et al. (2007).
Cubicles	Resource	Include	Included as pooled resource.
	Prioritisation	Include	Adds to variation in treatment times of minor patients.

A.2.3 Simplifications and Assumptions

Simplifications

- Decision to admit or discharge is aggregated into treatment times.
- Resuscitation treatment including x-rays and additional treatments is aggregated and modelled as a time delay.
- Nurse and doctor multi-tasking are not modelled in detail. For example, if a junior doctor can do 3 things at once then 3 junior doctor 'slots' are included in the model. See Günal and Pidd (2006) for a detailed explanation.
- Only 'level of emergency' is modelled. Detail of the specific injury or emergency is not required to assess system performance.

Assumptions

- Nurses always correctly evaluate the level of emergency a patient represents.
- Nurses can do no more than 2 things at once.
- Senior doctors can do no more than 5 things at once.
- Junior doctors can do no more than 3 things at once.
- The resus mobile x-ray is 100% available and not a constraint.
- Senior doctors do not send any patients to x-ray who do not need to go. I.e. their experience levels make efficient use of x-ray.
- It is more obvious if a major patient needs to go to x-ray. Thus there are minimal mistakes.
- Seniority of nurse does not affect patient treatment times.
- The distribution of x-ray usage by non-A&E patients does not vary throughout the day. I.e. arrival rate is not dependent on time.

A.2.4 Model Building Procedure

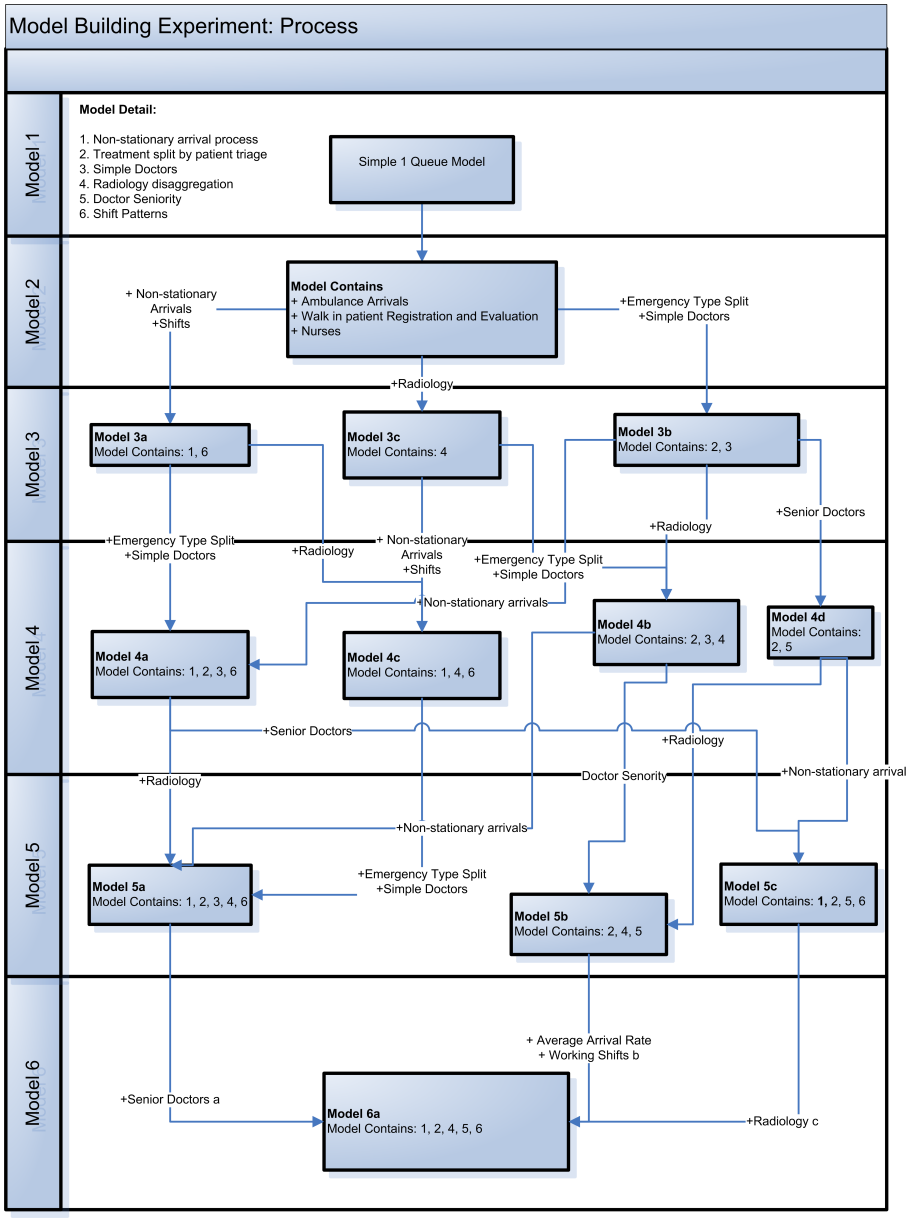


Figure A.1: Model Building Process Flow

A.3 Research Questionairre

Instructions to Participants

Thank you for taking part in this research. Your participation and feedback is greatly appreciated. You can be assured that your data will be treated anonymously.

This document contains three sections.

The first section contains several simple questions to gauge your attitude to certain aspects of the experiment. **It is estimated this should take you no longer than 10 minutes.**

The second section contains questions about the case study. **It is estimated this should take you no longer than 10 minutes.**

The third section contains questions about several simple business scenarios. Some of these questions are multiple choice in nature, but you must provide one or two bullet points on your reasoning behind the answer. Each question asks you to rate the confidence you have in your answer. **It is estimated that this section will take no longer than 35 minutes.**

You are not expected to spend a long time on your reasoning answers. This is not a test so there are no right or wrong answers. I simply wish to know what you believe the correct course of action is and why.

Section 1: St Specific's. Current System and Simulation Model

The following questions related to the resources in the St Specific's case study. Each question provides a pair of adjectives e.g. unlikely and likely. Please indicate your answer by circling the score in an appropriate box between the adjectives.

Please note that utilisation refers to the proportion of time that a resource, i.e., a member of staff spends doing any A&E work because of patient arrivals. E.g. if sufficient patients arrive in an 8 hour shift to result in a nurse spending 5.6 hours working then the nurse has an A&E utilisation of 70%.

1.	Increasing the utilisation of the staff at St Specific's over the next 6 months will increase A&E efficiency.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
2.	Increasing the number of staff on duty over the next 6 months will allow more patients to be treated concurrently.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
3.	Increasing the time that doctors, nurses and cubicles are available for patient treatment over the next 6 months will increase St Specific's A&E costs.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
4.	The bigger the work load of staff over the next 6 months the longer it takes to see and treat patients at St Specific's	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
5.	Maximising resource utilisation reduces queues and reduces patient waiting time	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
6.	Increasing utilisation of A&E staff will save money on hiring new staff and purchasing new equipment to meet the performance target.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
7.	The busier a member of staff over the next 6 months the higher their stress levels.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
8.	Increasing staffing and cubicle levels at St Specific's increases the time they are idle/empty.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
9.	Switching to a time tabled policy on critical shared resources such as x-rays will increase availability to A&E over the next 6 months.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
10.	Dropping to only two x-ray rooms for 6 hours of the day will reduce waiting times for emergency patients requiring x-rays over the next 6 months.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely

11.	Switching to a fixed 40% daily loss of x-ray machines from an average 30% daily loss of x-ray machines will increase performance against the government's target over 6 months.	Extremely unlikely	1	2	3	4	5	6	7	Extremely likely
12.	Increasing efficiency of staff usage at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
13.	Allowing more patients to be treated concurrently at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
14.	Increasing A&E costs by hiring new staff at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
15.	Longer times to see and treat patients at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
16.	Reducing queue sizes at St Specifics is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
17.	Saving money on hiring staff and purchasing resources at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
18.	High stress levels for staff is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
19.	Increasing the idle/empty time of staff and cubicles at St Specific's over 6 months is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
]20.	Time tabling the availability of x-ray machines for A&E over 6 months is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
21.	Increasing waiting times for emergency patients requiring x-rays at St Specific's is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
22.	Switching to a fixed 40% daily loss of x-ray machines from an average 30% daily loss of x-ray machines over 6 months is	Extremely bad	-3	-2	-1	0	+1	+2	+3	Extremely good
23.	For me to balance the utilisation of resources to the speed at which patients can be treated and leave A&E over the next 6 months is	Harmful	1	2	3	4	5	6	7	Beneficial
		The wrong thing to do	1	2	3	4	5	6	7	The right thing to do
		Important	1	2	3	4	5	6	7	Unimportant
		Pleasant	1	2	3	4	5	6	7	Unpleasant

24.	For me to maximise utilisation of resources (i.e. nurses, doctors, x-ray, cubicles) at St Specific's is	Harmful	1	2	3	4	5	6	7	Beneficial
		The wrong thing to do	1	2	3	4	5	6	7	The right thing to do
		Important	1	2	3	4	5	6	7	Unimportant
		Pleasant	1	2	3	4	5	6	7	Unpleasant
25.	For me to consider the behaviour of the A&E process (e.g. changes in patient arrivals, resource availability and queues) over the full 6 months period is	Harmful	1	2	3	4	5	6	7	Beneficial
		The wrong thing to do	1	2	3	4	5	6	7	The right thing to do
		Important	1	2	3	4	5	6	7	Unimportant
		Pleasant	1	2	3	4	5	6	7	Unpleasant

26. The simulation model of the A&E department and its results are

	Completely Disagree	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Completely Agree
Believable							
Accurate							
Biased							
Representative							
Reliable							
Surprising							

27. To what extent do you have confidence in the validity of your above assessment of the

Believability	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident
Accuracy	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident
Bias	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident
Representativeness	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident
Reliability	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

of the simulation model and its results?

Section 2: St Specific's A&E Current Setup.

The following questions relate to the A&E department at St Specific's hospital. Please provide a short answer for each question.

1. If a proportion of the total cubicles, nurses and doctors were dedicated to major emergencies and the remaining cubicles, doctors and nurses were dedicated to minor emergencies, what would happen to the number of performance target breaches and why?

Section 3: Reasoning Questions.

Please read the following scenarios and their corresponding questions. If a multiple choice answer is available please only choose one. In all cases please indicate your reasoning behind the answer. I.e. the reasons that it would improve performance of the system.

Scenario 1

Prior to opening a new doctor's surgery the General Practitioners (GPs) who work there must decide upon the type of surgery they run for everyday patients. To avoid a large number of complaints it is very important that patients do not have to wait for long periods of time to see the doctor. GPs are looking for the best way to minimise the time a patient spends waiting at the surgery to see them.

They can see two options:

a. Run as a 'drop in' clinic where patients can arrive at any time to see the doctor. Patients do not need to ring ahead.

b. Establish a number of appointment slots. Patients would need book an available slot in advance to see the doctor. An example slot length would be 5 minutes. Although once the patient has arrived this can of course take slightly longer or shorter depending on the specific requirements.

At other surgeries it is common for a patient to be with a doctor for an average of 5 minutes. Although some minor variance around this number is expected.

What option do you recommend (please circle one option). a, b
Please indicate why you recommend this option:

How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 2

Holby City A&E department has an evening shift:

- Nurses are occupied with direct patient care (e.g. registration, evaluation and treatment of patients) on average 77% of the time.
- Patient treatment takes, on average, 74% of the senior doctor's time.
- Patient treatment takes, on average, 75% of the junior doctor's time.
- Cubicles are on average utilised 60% of the time.

Over the next 6 months another nearby hospital will close down. It is expected that this will increase the number of emergency patients who must be treated in the evening. Although some demand will be taken by another nearby hospital - St James'. Like other hospitals Holby must report the percentage of patient treatments that breach 4 hours.

At a recent management meeting it was suggested more staff are needed on the shift.

Do you agree with this suggestion?

Answer: Yes/No
The reasons I agree/disagree with this suggestion are:

How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 3

St. Specific's hospital has 3 surgical operating rooms (surgery includes things like appendix removal, heart transplants etc.). Patient operations are booked in advance by a doctor to an operation slot. Each room has the same number of operation slots available each day.

Surgical operations do not always fit into a single slot. This is because the time of operations can vary significantly. That is, an operation may turn out to be more (or less) complex than expected and overrun. An operation may of course also be booked across multiple slots.

If an operation overruns its slot - for example if complications arise - the operation scheduled for the next slot has to wait until the operation ends.

- St Specific's has performance targets for operating room use. Firstly, no patients must wait longer than 15 minutes for an operating room to free up. Secondly, the scheduled operations must end no later than 15 minutes past the end of the day.
- On average 2% of patients have to wait over 15 minutes for an operating room to free up and an average of 4% of operations overrun the end of day per week
- Currently the operation slots of the rooms have an average weekly utilization of 82%.

Performance of the operating rooms was reviewed at a recent management team meeting. When asked how performance can be improved, the following arguments were presented by three different members of the team:

a.) The operating rooms are not being used enough and any additional spare slots should be used up.
b.) The hospital should consider adding extra operating rooms.
c.) The hospital should consider extending the duration of operation slots.

Which team member do you agree with? (Please circle one). a, b, c
Please indicate why you agree this team member and why their option would work:

How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 4

A small UK county supplies first line medical care to the public via two NHS services. A traditional doctor's surgery and an NHS Walk-in centre (a nurse led treatment facility).

- Patients do not need appointments for the doctor's surgery or the NHS walk-in centre.
- At both the surgery and the walk-in centre patients report to a reception as they enter. Each reception can have several receptionists working on them at a time.
- It is believed that the doctor's surgery is busier than the walk-in centre.

In an attempt to reduce costs it is planned that in the near future the doctor's surgery and the NHS walk-in centre will be housed in the same building. However, several questions remain over how the operations should be run.

One problem concerns patient reception. Management are keen that patients are seen and registered quickly. Especially since some major illnesses can be spotted as patients register. At the moment management are faced with one of four choices:

a.) Build a shared reception: doctor's surgery and NHS walk-in patients queue together and report to a single reception containing the combined number of receptionists.

b.) Have separate entrances to the building: doctor's surgery patients use one entrance and NHS walk-in patients use a second entrance. Each will queue for a separate reception and receptionists.

c.) Have separate entrances to the building and hire additional receptionists for the doctor's surgery.

d.) Have separate entrances to the building and hire additional receptionists for both the doctor's surgery and the walk-in centre.

What option do you recommend to minimise patient waiting time for receptionists? (please circle one). a, b, c, d
Please indicate the reasons why this option would minimise patient waiting time:

How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 5

A factory that manufactures steak and kidney pies has three connected machines: M1, M2 and M3.

- M1 assembles the pie. It currently has an average utilization of 65%.
- Ingredients are loaded manually onto M1. The rate at which this happens can be controlled.
- M2 conveys the assembled pies through a cold spiral and freezes them. It currently has a utilization of 86%.
- M3 packages the frozen pies. It currently has a utilization of 79%.

To help win business from supermarkets the factory quotes a lead time for an order (the time in takes to assemble, freeze, package and ship the order). As soon as the order is ready it is shipped. Recently a rival company has been able to ship orders a lot quicker than this factory. This has led to a slight drop in the orders received as supermarkets value fast delivery.

The factory is considering four options to increase the speed at which an order is shipped:

a.) Purchase an additional freezing spiral.

b.) Purchase an additional pie assembler

c.) Purchase an additional packaging machine.

d.) Increase M1s utilization to 100% (i.e. increase the rate that ingredients are loaded)

What option do you recommend to increase the speed at which an order can be shipped? (please circle one). a, b, c, d
Please indicate why your answer increases the speed at which an order can be shipped

How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 6

A police service operates across four different geographic regions of the UK. Police assistance is typically sort through '999' emergency calls. Figure 6a illustrates the call handling process.

- 999 calls are picked up immediately by BT operators who determine if the call is an emergency or a non-emergency. If the call is an emergency it is routed to a police emergency call centre in the appropriate geographic region (i.e. the closest). Non-emergencies are routed to inquiries at a local police station.
- Once an emergency call has been routed to a geographic location it then waits in a queue to be answered by a call taker. Once answered the call taker takes details and prioritises the call for police assistance.
- The number of calls arriving per hour varies across the day and across the week. I.e. expect very busy and variable demand in an evening and extremely busy and variable demand on a Friday night.
- Each call centre has its own call takers. The number of call takers on duty at a geographic region changes throughout the day in order to try and meet customer demand.

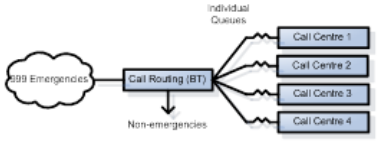


Figure 6a: Emergency Calls and Routing to Police Call Centres

The UK government requires that police call centres must answer 90% of all incoming emergency calls in 15 seconds. That is, it is highly desirable that an emergency does not have to wait longer than 15 seconds in the queue to speak to call takers.

The police service are worried as that they are not currently meeting this target. After some internal consideration and a consultation with some prominent academics at a UK University there appears to be a total of three options:

a.) Split one or more geographic regions. I.e. introduce a new call centre with a new queue and allocate a proportion of the call takers from the original call centre.

b.) Increase the number of lines available at each geographic call centre. I.e. hire additional call takers.

c.) Scrap geographic call centres and combine all operators into a single police call centre with a single queue for the four regions.

I would recommend (please circle) option to improve performance: a, b, c											
The reason(s) that this option would improve performance is											
How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 7

A pie factory has a small, but very successful, high street shop - Grugs - where cooked hot pies are sold. Any pies that are not sold are scrapped at the end of the day. Pies are sold as fast as they can be made!

Before being sold the pies must be assembled and cooked. They are produced in batches of 20 (i.e. a batch contains 20 pies) on two machines: a pie assembly machine and an oven for cooking the pies. Once pies are assembled they are stored in a big fridge waiting to be cooked. This can take a maximum of 10 batches. If the storage area reaches its maximum the pie assembly shuts down and waits for the some space to free up. Figure 7a illustrates this process:

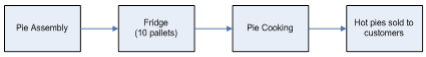


Figure 7a: The pies production process

- The pie assembly process takes on average 20 minutes for one batch of pies.
- The pie cooking process takes on average 40 minutes (depending on the ingredients of pie used) for one batch of pies.
- Pie assembly machines are complex equipment! The time between breakdowns varies a lot, but on average they break down on average once every 12 hours. On average it takes 4 hours to get the assembly machine back up and running. However the time to fix the machine also varies a lot.
- The oven never breaks down.

In the next month a new textiles factory is opening over the road. Factory workers mean extra demand for pies! As customer demand is currently only just met Grugs needs extra pies and quickly!
What would you recommend as the most effective and easiest way to meet the extra demand?

a.) Purchase an identical additional oven so that two palettes can be cooked at the same time.

b.) Hire permanent mechanics to prevent the assembly machine from breaking down. This would mean that the assembly machine is stopped every 30 minutes for a 10 minute maintenance job (i.e., the machine would be undergoing maintenance and not running for 2.7 hours in an 8 hour day or alternatively 4 hours out of 12 hours).

c.) Invest in an extension to the food preparation area and install an additional fridge with the same capacity.

I would recommend (please circle) option: a, b, c:											
The reason(s) that this option would improve performance is											
How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

Scenario 8

ShatterProof Ltd. provide a car windscreen maintenance and replacement service. Customers can book a qualified windscreen engineer to travel to wherever their car is located and either replace or deliver essential maintenance to their car windscreen.

ShatterProof's business relies on a call centre that receives customer calls. Operators take details of the customer's problem and location, organises an appointment, and importantly takes the customer insurance or credit card details.

ShatterProof feel it is essential 90% of customer calls are answered by a call operator in 3 minutes. This helps reduce the problem of customers becoming tired of waiting, hanging up and ringing a rival company - SafetyScreen.

As customer call demand changes throughout the day the call centre uses shifts. The following table provides utilization and performance information for the shifts:

Shift	Operators Average Utilization (per day)	Average Calls Answered in 3 Mins (per day)
Morning (8am to 11am)	88%	81%
Lunchtime (11am to 1:30pm)	73%	91%
Afternoon (1:30pm to 4:30pm)	71%	93%
Evening (4:30pm to 8:00pm)	92%	75%

ShatterProof are keen improve on their current performance and prevent business from being lost to SafetyScreen. At a recent management meeting the following options were suggested:

a.) Hire additional call operators for the 'Morning' and 'Evening' shifts.

b.) Move a proportion of call takers from the 'Lunchtime' and 'Afternoon' shifts to the 'Morning' and 'Evening' shifts

c.) Split the call centre and hence operators into two. That is, have one phone number for windscreen replacement calls and a second for maintenance. Each would take a proportion of the total operators.

Which option do you choose? (Please circle one option): a, b, c											
Please give the reasons why your choice would improve performance											
How confident are you in the validity of your answer?	Not at all confident	1	2	3	4	5	6	7	8	9	Extremely confident

A.4 Answers to Reasoning Questions

Scenario 1

Transfer is achieved by recognising the effect of variable arrivals on queuing time (and resource utilisation). The drop in clinic will produce peaks and lulls of demand whereas an appointment system removes the arrival variability (although there will still be some variation in treatment time).

Scenario 2

Transfer is achieved by recognising the relationship between resource utilisation and waiting time. Any increase in demand will increase the utilisation of nurses / doctors. Resource may still be able to 'cope' with demand, but the waiting time of patients will increase and so will the number of target breaches.

Scenario 3

Transfer is achieved by recognising the relationship between resource utilisation and waiting time. Higher utilisation simply means that there is a greater chance of overruns and more target breaches. If the utilisation was slightly lower there would less chance of overruns. Alternative answers are:

a.) This increases the chance of overruns.

c.) An increase in duration would reduce overruns to the next slot - utilisation of slots would therefore be increased. The problem is then are there enough operating slots per day/week? This approach is likely to increase the waiting time for an operation.

Scenario 4

Transfer success is achieved by recognising that pooling the resources (receptionists) is beneficial for performance. If the two groups of receptionists are pooled then this should reduce the likelihood that a patient is queuing at one entrance while there are free receptionists at another. Alternative answers are:

b.) This is likely to result in patients queuing at entrance, while receptionists are free at the other.

c.) Performance would be improved, but this does not represent transfer and is not cost efficient.

d.) Same as c.

Scenario 5

A queue of work is typically found for machine M2; hence increasing the utilisation of M1 makes no difference as work is likely to become blocked at M2. Transfer is achieved by recognising the reasons for high utilisation at M2. As there is no way to remove the variation in processing time - in this scenario - choosing to purchase an additional M2 is the best option. Alternative answers are:

b.) This would increase the WIP at M2.

c.) This is not the bottleneck machine.

d.) This would increase the WIP at M2.

Scenario 6

Calls to the call centre arrive at varying unpredictable intervals from the different regions resulting in bursts of activity and very quiet periods. Calls also take a

variable amount of time to deal with by operators which can also result in queues. The current call centre setup suffers due to this variation.

A highly busy period in one call centre will result in queues. However, at the same moment in time one or more of the other 3 call centres may have some free operators. This results in breaches of the targets. Note that this only happens because of the variation.

If the call centres were combined then this situation cannot occur. That is, a call cannot be waiting while an operator is at their desk unoccupied with a call. Hence those calls that would breach the target in the regional call centre model are less likely to in the pooled resource call centre. Performance would improve. Alternative answers are:

a.) Splitting a geographic region makes the problem of variation in arrivals and call duration even worse. Callers in that region are more likely to find a caller busy while another operator is free in the other half of the region. Hence service level will plummet in the region.

b.) Increasing the number of operators available in each region will improve performance. However, participants have again failed to understand the effect of variability on the system. The increase in operators will at some point yield the same performance as the combined call centre. However, this solution costs money long term - in that additional employees are permanently needed. They will also have lower utilization than in the combined call centre.

Participants may also suggest an additional scenario where a merge results in two call centres. One may be used for redundancy if the other call centre goes down for some reason.

Scenario 7

We are not looking for a massive increase in throughput.

The large source of variability in the system is the combination breakdowns of the assembly machine and the bound queue. The breakdowns mean the arrivals to the slower oven are highly variable. The bound queue means that at times the oven is starved of work 5% of the time because it has processed all of the batches from the store while the assembly machine is down.

a.) If a second oven was purchased at considerable expense then throughput target would be achieved. However as only an additional two batches are needed it would be overkill.

b.) The maintenance program means that in an 8 hour day the machine will be down for 2.7 hours. Whereas the variable break downs result in significantly more time loss in the long term. Although the oven has a very high utilization it is idle while there is work in the system. Since the breakdown variability is eliminated by the maintenance the utilization can now go up to higher levels. Transfer is achieved here by recognising that the elimination of this variation improves performance (and means that high utilization is possible).

c.) This choice would also work as there is a bigger buffer for pies, but it does not represent transfer from the experiment and (may) not be a cost effective choice. On average the oven has only 5% available capacity. This of course will be filled, but there will be a lot more work in progress in the system while the two assemblers wait for the oven to finish.

Scenario 8

As average utilization of call operator's increase on a shift so does the time a caller has to wait for their call to be answered. You can see this when you compare operator utilization to the percentage of calls answered in the table provided. Higher utilizations have lower performance.

Reducing the operator's utilization in the morning and afternoon by hiring more staff would improve performance. Alternative answers are:

b.) Reallocation of lunchtime and afternoon employees reduces the performance of these shifts (as less operators have to work harder - maybe 80 to 95%). The improvement in Morning and Evening shifts performance due to increased staffing levels may be balanced out or perhaps have a surpassed by the loss in performance in the lunchtime and evening shifts.

c.) This implies firstly that the participant does not understand the effects of variability on a process.

Appendix B

Single-Loop Comparison Appendix

B.1 Analysis Considerations

This appendix contains technical information related to the analysis and comparison of single-loop learning across the three conditions.

B.1.1 Calculating an effect size

Field (2009) details the procedure for calculating and effect size r from the output of a Mann-Whitney test.

$r = Z/\sqrt{N}$

Where Z is the z-score outputted from the Mann-Whitney test and N is the sample size.

B.1.2 The percentile bootstrap method

Wilcox (2005) outlines the percentile bootstrap method for calculating confidence intervals for mean differences.

- Let θ_1, θ_2 be the statistics of interest from a samples of n_1, n_2 values.
- Let $\hat{\theta}_i^*$ be an estimate of θ_i based on resampling n_i values with replacement.
- Let $D^* = \theta_1^* - \theta_2^*$ be the difference between the two resampled estimates.
- Repeat this process B times (here $B = 2000$) yielding $(D_1^*, ..., D_B^*)$
- All resampled D^* are arranged in ascending order. $(D*_1 \leq ... \leq D_B^*)$
- Let $l = \alpha \mathrm{B}/2$ and $u = B - l$
- An approximate 1 - α confidence interval for $\theta_1 - \theta_2$ is $\left(D_{l+1}^*, D_u^*\right)$

B.1.3 Mann-Whitney Test Approach

All non-parametric statistics reported are the results of Mann-Whitney tests. The Mann-Whitney procedure firstly ranks all data. This has the benefit of reducing the impact of outliers on the data set, but loses information on the magnitude of the differences. The statistics reported in this section are a significance value and corresponding estimate of effect size (see Section 6.2 for a definition). Given the small sample size, alpha is set to $\alpha = .1$ to improve power.

Note that reported significance is classed as an exact significance. To calculate this value we firstly assume that none of the experimental conditions have any effect on the dependent variable(s). If this is the case then a significant effect may be simply due to the (random) allocation of participants between conditions. Thus we can reformulate our null and alternative hypotheses as:

- H_0: The test effect is dependent on the ordering of participants between conditions;
- H_1: The test effect is not dependent on the ordering of participants between conditions.

To test this hypothesis it is necessary to generate the distribution of the test statistic (e.g. mean difference in ranks) given different orderings of the participants. For small samples (as is the case here), it is possible to generate every combination of participants and significances are classed as exact. For large samples, it is necessary to perform Monte-Carlo re-randomisation of participants - Lunneborg (2000) recommends at least 10,000. The proportion of the distribution containing a test statistic as or more extreme as the one observed is the exact significance. If this is small (say $p < .1$) then we can conclude the chances of the results being due to the allocation of participants to condition is highly unlikely. I.e. the null hypothesis is rejected.

B.1.4 Multiple Comparison Control

When simulation practitioners perform multiple comparisons of simulation scenarios it is standard practice to apply a Bonferroni Correction to the alpha level used. (See Law, 2007; Robinson, 2004). This is due to the inflation of the probability of making a single Type I error - more commonly known as the Familywise Error Rate (FWER) (Benjamini and Hochberg, 1995) - when making multiple comparisons. A Bonferroni Correction ensures that the overall probability of making a Type I error remains at level alpha. This control of the FWER is especially important when choosing the best system design out of n possible designs; for example, selecting the process design that maximises throughput in a factory. However, there are at times two problems with its application.

If there are M comparisons to be made, then Bonferroni works simply by dividing α by M. For example, the desired alpha level is $\alpha = .05$ and there are eight comparisons to be made then eight comparisons are made at the $\alpha_{Bon} = .00625$ level. The first problem with Bonferroni then occurs when M is large. In fact the conservative nature of Bonferroni means that its application is impractical beyond

15 comparisons (Robinson, 2004). Although the FWER control is applied the approach is so conservative that it is likely several Type II errors (failing to reject the null when it is false) are made.

The second problem relates to how appropriate it is to control the FWER. If a decision maker wishes to understand how (management) decisions or factors of a system affect the performance of a system then it is desirable to have a better trade off between type I and II errors. This is when comparisons are used to explore a problem rather than to choose a system. Strong control of the FWER might mean more type II errors and hence I fail to recognise the influence of some experimental factors on performance.

The first and second problems apply in the analysis of attitude change in this experiment. Firstly, the total of 18 attitude change comparisons and a small sample size mean that use of Bonferroni is much too strict. Secondly, the problem concerns differences in learning as well as disconfirming predictions. Hence we are not concerned with choosing a single best condition for learning, rather exploring differences in outcomes and process for learning.

A more suitable procedure for this type of research - False Discovery Rate (FDR) control - was introduced by Benjamini and Hochberg (1995). Here the problem of multiple comparisons is conceptualised differently. For example, take an experiment where m comparisons between conditions are made. Out of these a total of R significant differences are found. FDR controls the expected proportion of R that are false positives. This results in a less conservative approach to multiple comparison control than a simple Bonferroni correction (Benjamini and Hochberg, 1995) and is more suitable to an exploratory study than control of the FWER. As an insurance policy FDR also controls FWER in the *weak sense* (Benjamini and Hochberg, 1995). If all of the m hypotheses are true (i.e. there is no evidence to suggest differences between samples/conditions) then FDR is equivalent to Bonferroni. These claims

are backed up by several simulation studies (e.g. Benjamini and Hochberg, 1995, 2000).

The basic FDR procedure is sequential and applied to the p-values of the significance tests. The procedure given in Benjamini and Hochberg (1995) is:

1. Conduct the m significance tests as normal at level α
2. Order the resulting p-values from smallest to largest $(p_1 ... p_m)$
3. Starting at the highest p-value $(i = m)$and working downwards compare each p_i to $\alpha i/m$
4. Reject all hypotheses that are above the first $p_i < \alpha i/m$

The single-loop results presented in Section 6.4 and the double-loop results presented in Section 9.4 use the FDR procedure for multiple comparison control.

B.2 Retrospective Comparison of Exam Marks

The primary approach of the research to prevent an imbalance of intelligence across groups is to use randomisation. That is, participants are randomly assigned to a condition. Although this mechanism is largely believed to be effective (Field, 2009) in practice this may be problematic due to the small sample size of the experiment. Therefore it was decided that a retrospective approach to testing for differences in intelligence was appropriate.

The most common and relevant module taken by participants is the Quantitative Analysis Methods (QAM) module in the business school. The module is split over two parts. The first (QAM1) introduces students to basic statistical concepts such as probability trees and simple regression. The second (QAM2) is an introduction to Operational Research e.g. Linear programming and the Economic Order Quantity.

Permission to use the marks was obtained retrospectively. Students who participated in the experiment were contacted by e-mail. To boost response rate a prize draw for 25 was held for all replies regardless of their nature (i.e. permission granted or denied). Students were also told that a 10 donation would be made to Sports Relief if 50 responses were received (again regardless of their nature). In total 45 students responded - all of which were positive; however, it was not possible to obtain marks for all students as some did not take the module and others took QAM1 only. The sample sizes used for each condition are listed in Tables 1 and 2. Both QAM1 and QAM2 marks are out of 100%.

This appendix firstly explores the distribution of marks using boxplots and descriptive statistics. Secondly, an exploration of within group relationships is presented; i.e. correlations between exam marks and single-loop learning variables. This is followed by a formal inference test for differences in exam marks between conditions. Lastly, a conclusion is drawn on the effect of intelligence on performance.

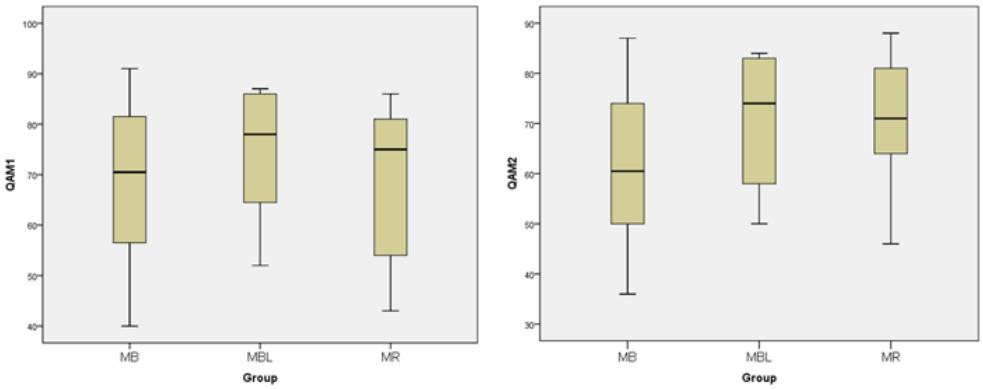
B.2.1 Descriptive Statistics

Figures B.1a and B.1b illustrate the distribution of exam marks using boxplots. The most notable point, out of the two diagrams, is that it appears that the median value of the MB condition is slightly lower than MBL and MR.

This result is reflected in the Tables B.1 and B.2: the mean QAM1 and QAM2 scores by condition. The biggest difference in means is found between the MB condition and others in the QAM2 results. Although note that it has the highest standard error.

B.2.2 Inferential Results: Between Groups

The distribution of QAM1 and QAM2 marks across the groups was tested using a Kruskall-Wallis test (as there were no specific hypotheses). This found no evidence



(a) QAM1 Results **(b)** QAM2 Results

Figure B.1: Boxplots of QAM Marks

Table B.1: QAM1 Marks

	Mean	S.E	N
MB	68.5	6.1	8
MBL	75.0	3.7	12
MR	69.0	4.1	13

Table B.2: QAM2 Marks

	Mean	S.E	N
MB	61.5	6.1	10
MBL	70.7	3.9	10
MR	71.2	4.0	8

to support a difference between the conditions ($p > .1$).

B.2.3 Conclusions

The sample size obtained for the retrospective analysis is extremely small and all results should be considered with this limitation in mind. The analysis found no substantial evidence for differences in exam performance between the conditions. The descriptive statistics support this view for QAM1, but hint at a potential difference

in QAM2. However, it is difficult to be certain given the data limitation.

B.3 Comparison of results adjusted for regression to the mean and original data

The analysis of attitude change in this research focuses on subgroups i.e. groups that are above or below zero. One problem with this approach is that pre-test data that are extreme (i.e. scoring relatively high or low on the attitude measure) are likely to regress towards the mean on the post-test score. Thus high scores are more likely to be in the reduction in attitude change subgroup by chance alone and vice versa. Furthermore evidence was found that the typical pre-test scores for ElimVar were different across the subgroups. To address these problems the main analysis is complemented with simple adjustment procedure used to reduce the impact of regression to the mean.

This section provides a comparison the results of the adjusted analysis to the original data. Results are presented as tables of inferential results and effect sizes. Differences in interpretation are discussed using effect sizes, Q-Q plots and typical baseline pre-test scores. The section is organised by comparison.

B.3.1 MR versus MB

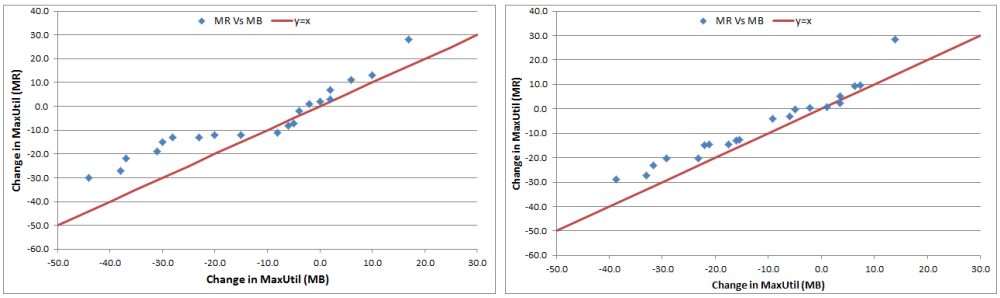
There two notable differences between the original and adjusted analyses of the MR and MB comparison. The first is in the interpretation of the correct attitude change subgroup for $ElimVar$: the original data concludes that there is evidence of a difference, while the adjusted cannot conclude there is a difference and reports a (very) small effect size. This can be seen in correct grouping for $ElimVar$ in Table B.3. This is result is not surprising as section 5.2.2 reported the pre-test scores were unbalanced across the subgroups. In particular section 5.3.1 illustrated that the MB adjusted median change in ElimVar was relatively low compared to the originals.

The second change illustrated by Table B.3 is that correct subgroup for MaxUtil is no longer significant, although the effect size is similar. Figures B.2a and B.2b

Table B.3: MR versus MB Comparison

		Original		Adjusted		
Direction	**Measure**	**Result**	**Effect Size (r)**	**Result**	**Effect Size (r)**	**Agree**
Correct	MaxUtil	$MB > MR$	0.20	No Diff	0.24	False
	TradeUtil	No Diff	0.14	No Diff	0.14	True
	ElimVar	$MB > MR$	0.28	No Diff	0.02	False
Incorrect	MaxUtil	No Diff	0.07	No Diff	0.04	True
	TradeUtil	No Diff	0.07	No Diff	0.21	True
	ElimVar	$MB < MR$	0.39	$MB < MR$	0.49	True

illustrate the original and adjusted distributions using Q-Q plots. It can be seen that the points are slightly closer to the line in the adjusted distribution - hence the difference is smaller and the results are no longer significant - however, they are still consistently above the line (hence a similar effect size). A safe conclusion to draw from the data, therefore, is that the MB condition does have an effect on attitude change of $MaxUtil$ relative to MR; however, the effect is small.



(a) Original **(b)** Adjusted

Figure B.2: MR versus MB MaxUtil Q-Q Plots: Original and Adjusted

B.3.2 MR versus MBL

The first important difference between the original and adjusted inference tests is found in incorrect direction subgroup for $TradeUtil$. Table B.4 illustrates that the effect size is similar in both the original and adjusted data ($r = .23$ and ($r = .23$ respectively); however, the adjusted data result finds evidence that this difference is not due to chance alone, while the results for original data does not. Figure B.3 illustrates the differences in the distribution of the MR and MBL $TradeUtil$ results using a Q-Q plot of the adjusted data. It can be seen that the points below zero (on the x axis) lie below the line $y = x$; there are also a number of points around zero. This, and the similar effect size, supports the view that attitude change in the incorrect direction was greater in MR than MBL.

Table B.4: MR versus MBL Comparison

		Original		**Adjusted**		
Direction	**Measure**	**Result**	**Effect Size (r)**	**Result**	**Effect Size (r)**	**Agree**
Correct	MaxUtil	No Diff	0.22	$MR > MBL$	0.31	False
	TradeUtil	$MR > MBL$	0.40	$MR > MBL$	0.29	True
	ElimVar	No Diff	0.11	No Diff	0.01	True
Incorrect	MaxUtil	No Diff	0.14	No Diff	0.14	True
	TradeUtil	No Diff	0.23	$MR > MBL$	0.21	False
	ElimVar	No Diff	0.10	No Diff	0.17	False

There is also a difference between the conclusions for the correct direction subgroup of $MaxUtil$. Table B.4 illustrates that the original data does not find enough evidence to conclude a difference is present (and reports a small effect size), while the adjusted analyses does find enough evidence and reports a small to medium effect size. Looking at Q-Q plots of the original and adjusted data in Figure B.4 it is clear in both the original and adjusted data the lower quantiles dip consistently below the line $y = x$; however, the difference appears to be slightly greater in the

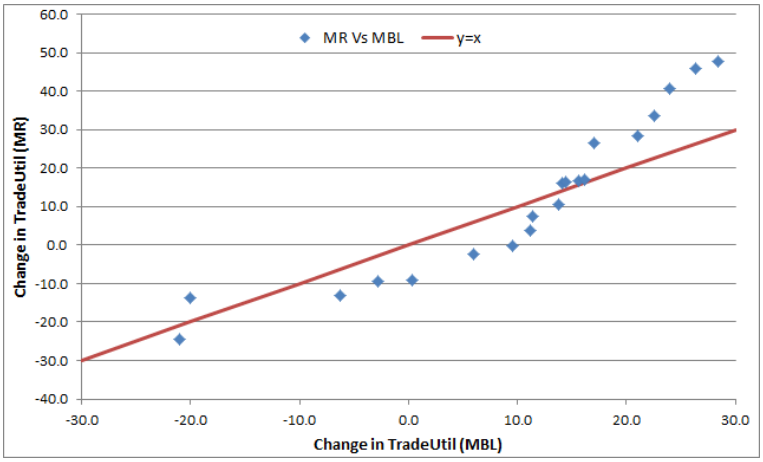
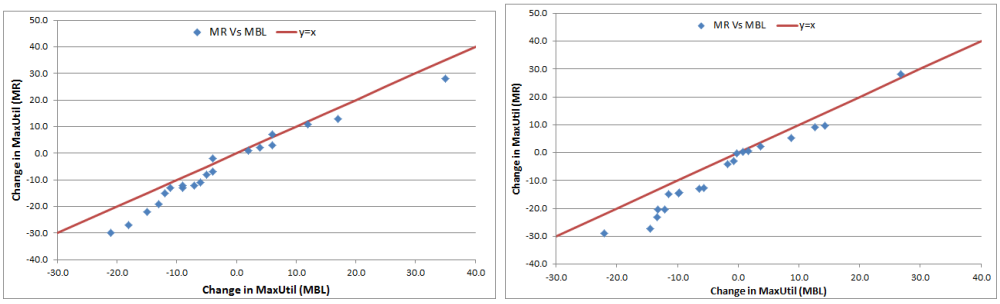


Figure B.3: Q-Q Plot of TradeUtil MR versus MBL (adjusted data)

adjusted data. Again a safe conclusion appears to be that some effect is taking place between MR and MBL, but that it is small in size. One explanation for this difference may be interaction between the complexity of the concept of resource utilisation and the constraint in the MBL condition to run scenarios exploring it. Section 7.1.1 discusses this possible mechanism in more detail.



(a) Original **(b)** Adjusted

Figure B.4: MR versus MBL MaxUtil Q-Q Plots: Original and Adjusted

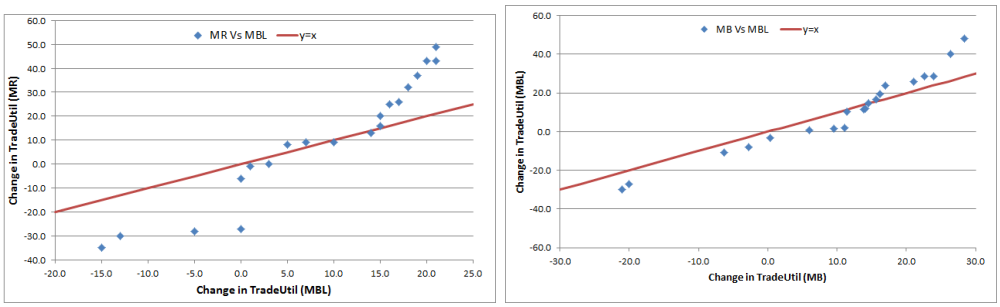
B.3.3 MB versus MBL

The adjusted and original data show a difference between the correct direction of attitude change in $TradeUtil$. Table B.5 shows that the original data concludes there is evidence of a difference and the adjusted data does not. Figure B.5 illustrates, using Q-Q plots, that the effect is slightly reduced after adjustment for regression to the mean (the points below zero are closer to the line $y = x$ after adjustment). This is potentially because the sample mean of the MBL subgroup is slightly higher than the MB group (mean difference = 8.5), although inference tests do not find sufficient evidence to conclude that the difference is due to anything other than chance alone. Thus it might be concluded that there is a small effect ($r = .11$) on the correct direction of attitude change, but that the experiment lacks sufficient statistical power to detect it.

Table B.5: MB versus MBL Comparison

		Original		Adjusted		
Direction	**Measure**	**Result**	**Effect Size (r)**	**Result**	**Effect Size (r)**	**Agree**
Correct	MaxUtil	$MB > MBL$	0.27	$MB > MBL$	0.40	True
	TradeUtil	$MB > MBL$	0.28	No Diff	0.11	True
	ElimVar	No Diff	0.15	No Diff	0.03	True
Incorrect	MaxUtil	No Diff	0.19	$MB < MBL$	0.40	True
	TradeUtil	No Diff	0.21	No Diff	0.17	False
	ElimVar	No Diff	0.42	No Diff	0.33	False

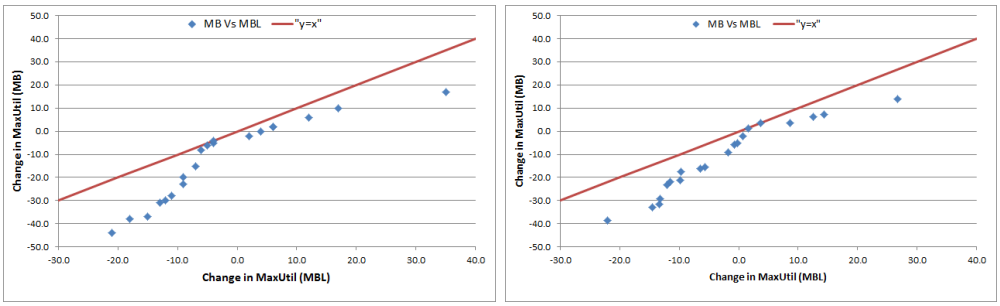
The second difference concerns the result for incorrect change in $MaxUtil$. Table B.5 shows that the original inferential results do not find a difference, but inferential results for the adjusted data do conclude that a difference is present. However, looking at Figures B.6a and B.6b, Q-Q plots of MB versus MBL using the original and adjusted data respectively, it can be seen the majority of the points lie below the line in both plots; however, the number is greater in the data for regression to



(a) Original **(b)** Adjusted

Figure B.5: MB versus MBL TradeUtil Q-Q Plots: Original and Adjusted

the mean; hence the disagreement between the inference tests. Although the results are similar, it is recommended that some caution is taken in simply assuming that the adjusted conclusion is correct. This is because the pre-test means of the two subgroups were very similar (mean difference = 0.3).



(a) Original **(b)** Adjusted

Figure B.6: MB versus MBL MaxUtil Q-Q Plots: Original and Adjusted

Appendix C

Double-Loop Learning Appendix

C.1 Correlations Results

Table C.1: Variables Listed in the Correlation Analysis

Dependent Variable	Description
$HCTranSuccess$	High Confidence Transfer Success - the total transfer success of the participant that have a confidence rating of seven or higher;
$UtilTransferSuccess$	The total transfer success of learning about the relationship between resource utilisation and system performance;
$TradeUtil$	The change in attitude towards trading off some resource utilisation to achieve higher system performance. A positive change represents beneficial attitude change resulting from the simulation, as this indicates that participants recognise that utilisation has a relationship to system time;

Table C.1 – continued from previous page

$ElimVar$	The change in attitude towards reducing the variation in the availability of radiology resources. A positive change represents beneficial attitude change resulting from the simulation, as lower variation in the availability of radiology improves long term system time;
$TotalScenarios$	The total number of scenarios simulated by the participant in the experiment;
$OtherScenarios$	The number of scenarios out of the total that included an experimental factor not listed in the pre-defined case study scenarios;
$StatViewsPerScenario$	The average number of times a participant viewed more detailed results charts per scenario they simulated.
$ModBuildRoute$	The first choice that MB and MBL partipants make during the model building exercise.

Table C.2: MB Significant Correlations

Variable 1	**Variable 2**	r_s	sig
TradeUtil	HCTransferSuccess	.566	$p < .01$
ElimVar	HCTransferSuccess	.465	$p < .05$
ElimVar	UtilTransferSuccess	.459	$p < .05$
TotalScenarios	TransferSuccess	-.477	$p < .05$
TotalScenarios	FarTransferSuccess	-.573	$p < .05$
ModelBuildingRoute	CloseTransferSuccess	-.497	$p < .05$
OtherScenarios	CloseTransferSuccess	-.458	$p < .05$

r_s = Spearman's Rho non-parametric correlation coefficient

Table C.3: MBL Significant Correlations r_s

Variable 1	**Variable 2**	r_s	sig
ModelBuildingRoute	TransferSuccess	.453	$p < .05$
ModelBuildingRoute	FarTransferSuccess	.368	$p < .1$
StatViewsPerScenario	UtilTransferSuccess	.395	$p < .1$

r_s = Spearman's Rho non-parametric correlation coefficient

Table C.4: MR Significant Correlations r_s

Variable 1	**Variable 2**	r_s	sig
TradeUtil	CloseTransferSuccess	.414	$p < .1$
TradeUtil	CloseTransferSuccessVar	.392	$p < .1$
ElimVar	CloseTransferSuccess	.400	$p < .1$
StatViewsPerScenario	Transfer Success	-.520	$p < .05$
StatViewsPerScenario	FarTransferSuccess	-.462	$p < .05$
OtherScenarios	FarTransferSuccess	.590	$p < .05$
StatViewsPerScenario	HCTransferSuccess	-.383	$p < .1$
OtherScenarios	HCTransferSuccess	.534	$p < .05$

r_s = Spearman's Rho non-parametric correlation coefficient

Appendix D

Double-Loop Comparison Appendix

D.1 Analysis Considerations

D.1.1 Calculating an Odds Ratio

Table D.1 is an example contingency table comparing transfer success in scenario x between the MBL and MR conditions. The calculation is as follows:

	MBL(0)	MR(1)	Total
Transfer Success (1)	15	8	23
Transfer Failure (0)	5	12	17
Total	20	20	40

Table D.1: Example Contingency Table for Scenario X

$$Odds_{SuccessMBL} = \frac{\sum MBL_{Success}}{\sum MBL_{Failure}} \tag{D.1}$$

$$Odds_{SuccessMBL} = 15/5 = 3.0 \tag{D.2}$$

$$Odds_{SuccessMR} = \frac{\sum MR_{Success}}{\sum MR_{Failure}} \tag{D.3}$$

$$Odds_{SuccessMR} = 8/12 = 0.7 \tag{D.4}$$

$$Oddsratio = \frac{Odds_{SuccessMBL}}{Odds_{SuccessMR}} \tag{D.5}$$

$$Oddsratio = 3.0/0.7 = 4.3 \tag{D.6}$$

Thus the odds of achieving transfer in scenario x were 4.3 times higher in MBL than MR.

D.2 Chi-Square Test Results

Table D.2 presents the results of two Chi-square tests of association between the MR and MBL conditions. These are tests of differences in the performance on transfer scenarios two and six respectively. The test result for difference in performance of scenario six is significant at the $\alpha = .1$ level. However, there may be a problem with this result. Table D.3 is the 2x2 contingency table for scenario six transfer success in the MBL and MR conditions. When the group has a count below five and the sample size is small then the test will only have an approximate chi-square distribution (Field, 2009). In these circumstances it is necessary to perform an exact test (Field, 2009) - see Appendix B.1.3 for an explanation of the exact test. This yields a revised p-value of .092. Although the result is still significant, it appears the group size issue is having an effect on results. Thus although there appears to be a difference in performance there is not substantial evidence to support this view.

Measure	Chi-Square	Sig
Scenario Two	1.24	.265
Scenario Six	3.40	.067*

Table D.2: MR versus MBL - Chi-Square Test of Association Results

	MBL(0)	MR(1)	Total
Transfer Success (1)	1	5	6
Transfer Failure (0)	20	15	35
Total	21	20	41

Table D.3: MR versus MBL - Contingency Table for Scenario 6