A Quest for a Better Simulation-Based Knowledge Elicitation Tool

by

Poh Khoon Ernie Lee

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

The University of Warwick, Warwick Business School

October 2007

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have happened without the various forms of support from the Engineering and Physical Sciences Research Council through the Warwick Innovative Manufacturing Research Centre initiative, Ford Motor Company (Ford), Lanner Group Limited (Lanner), Warwick Business School, and the people below.

First and foremost, many thanks to Professor Stewart Robinson (my main supervisor) for his guidance, encouragement, occasional reassurance, and most importantly, for taking his chance with a fresh postgraduate a couple of years ago.

In addition, I would like to express my gratitude to Professor John S Edwards (Aston Business School), John Ladbrook (Ford), Tony Waller (Lanner), Professor Ruth Davies (my second supervisor), Justice Akpan and Thanos Alifantis for their timely intervention and help. Also, I would like to thank Professor Mike Pidd (Lancaster University Management School) and Kathy Kotiadis (Warwick Business School) for examining this thesis and suggesting ways to make it even better.

Furthermore, I am deeply indebted to the personnel from Ford for their willingness to participate in my experiments. They are Gian Singh, Jerry Kilonda, Malkit Singh, Premjit Kerai, Raj Katechia, Raj Matharu, Ranjit Dhesi, Tom Brew and Ranta Varsani.

At the risk of this becoming an Oscar-ceremony-type thing, I am also very grateful to the following friends for making my seemingly never-ending Ph.D. years more bearable and enjoyable: Antuela Tako, Ge Guo, Kahwai Fok, Martin Liu, Menesh Patel, Suchi Patel and Wenzhi Yan.

Last but not least, I would like to extend my heartfelt appreciation to Lennon Teng (my soul mate) for always being there for me, as well as my family for their patience, understanding and firm belief in me.

Thank you. ☺

PS: I would like to dedicate this thesis to the memory of my grandmother, who is sadly missed.

ABSTRACT

Knowledge elicitation is a well-known bottleneck in the development of Knowledge-Based Systems (KBS). This is mainly due to the tacit property of knowledge, which renders it unfriendly for explication and therefore, analysis. Previous research shows that Visual Interactive Simulation (VIS) can be used to elicit episodic knowledge in the form of example cases of decisions from the decision makers for machine learning purposes, with a view to building a KBS subsequently. Notwithstanding, there are still issues that need to be explored; these include how to make a better use of existing commercial off-the-shelf VIS packages in order to improve the knowledge elicitation process' effectiveness and efficiency.

Based in a Ford Motor Company (Ford) engine assembly plant in Dagenham (East London), an experiment was planned and performed to investigate the effects of using various VIS models with different levels of visual fidelity and settings on the elicitation process. The empirical work that was carried out can be grouped broadly into eight activities, which began with gaining an *understanding* of the case study. Next, it was followed by four concurrent activities of *designing* the experiment, *adapting* a current VIS model provided by Ford to support a gaming mode and then *assessing* it, and *devising* the measures for evaluating the elicitation process. Following these, eight Ford personnel, who are proficient decision makers in the simulated operations system, were organised to play with the game models in 48 knowledge elicitation sessions over 19 weeks. In so doing, example cases were *collected* during the personnel's interactions with the game models. Lastly, the example cases were processed and *analysed*, and the findings were *discussed*.

Eventually, it seems that the decisions elicited through a 2-Dimensional (2D) VIS model are probably more realistic than those elicited through other equivalent models with a higher level of visual fidelity. Moreover, the former also emerges to be a more efficient knowledge elicitation tool. In addition, it appears that the decisions elicited through a VIS model that is adjusted to simulate more uncommon and extreme scenes are made for a wider range of situations. Consequently, it can be concluded that using a 2D VIS model that has been adjusted to simulate more uncommon and extreme situations is the optimal VIS-based means for eliciting episodic knowledge.

LIST OF ABBREVIATIONS

The abbreviations that are used throughout this thesis are listed below:

2D	2-Dimensional
2*l*	2 litres (engine capacity)
2.4*l*	2.4 litres (engine capacity)
2½D	2½-Dimensional
3D	3-Dimensional
A-D	Anderson-Darling test
AI	Artificial Intelligence
ANOVA	ANalysis Of VAriance
ATD	After Test Dress
CA	Cluster Analysis
CBR	Case-Based Reasoning
COTS	Commercial Off-The-Shelf
DES	Discrete-Event Simulation
EPSRC	Engineering and Physical Sciences Research Council
Ford	Ford Motor Company
KBI	Knowledge-Based Improvement
KBS	Knowledge-Based System
KBSDLC	Knowledge-Based System Development Life-Cycle
K-S	Kolmogorov-Smirnov test
MDS	Multi-Dimensional Scaling

OR	Operational Research
SDLC	Systems Development Life Cycle
SSD	Sum of Squared Distances
UV	Ultra-Violet
VIS	Visual Interactive Simulation
VR	Virtual Reality

1 Introduction

1.1 BACKGROUND

Developing useful models of complex systems is inherently difficult. It is made worse when the systems interplay with human intent and action. Whilst many authors such as Willemain (1994), Powell (1995) and Pidd (2003) argue that it is desirable to model simply, it is also widely conceded that such parsimony should be exercised with an eye on the models' purposes. In short, model fidelity should match model needs. In this respect, if a model is intended for examining the effects of or even to be used as a means for improving human interaction with an operations system, then it should mimic the human decision makers' behaviour in the system as closely as possible.

Human decision-making and intervention is a significant element in most manufacturing systems. Baines and Kay (2002) comment that a manufacturing system may involve any number of manual processes and many aspects of its operation such as scheduling of maintenance works and allocation of resources may require human decision-making. They also add that human decisions and interventions may have a great impact on the systems' performances. As such, manufacturing systems provide a legitimate context for investigating how to model human-operations system interaction appropriately.

Towards this end, Robinson *et al*. (2005) undertook a project (Grant reference: GR/M72876) sponsored by the Engineering and Physical Sciences Research Council (EPSRC), which aims to identify and improve human decision-making in an operations system. Facilitated by a real-world case study set in a Ford Motor Company (Ford) engine assembly plant in Bridgend (Wales), the project ultimately led to the development and application of the Knowledge-Based Improvement (KBI) methodology. Broadly, the KBI methodology is based on Visual Interactive Simulation (VIS) and Artificial Intelligence (AI). It starts by eliciting episodic knowledge in the form of example cases of decisions from the human decision makers via a VIS model. Next, AI methods are used on these example cases to learn and represent the decision makers' strategies for decision-making. Then, the AI models are linked with the VIS model to predict the operations system's performance under different strategies. Lastly, the methodology ends with attempts to improve existing strategies.

In their conclusion, Robinson *et al.* (2005) discover from the knowledge elicitation phase that human decision makers may make less realistic decisions in a simulated environment. It is because they are likely to assume greater risks when there are no real consequences from their decisions. In addition, the authors also recognise that the decision makers may find the experience of providing a full set of data that comprises of a very large number of useful example cases to be a very laborious and time-consuming one. Consequently, these issues provide the impetus for another EPSRC-sponsored project (Grant reference: GR/R64841) that seeks to answer the following questions:

i. Is VIS a valid tool for eliciting knowledge? If there are successful demonstrations of using VIS to collect example cases for machine learning purposes such as rule induction, case-based reasoning, or neural network computing, then;

ii. How can VIS be adapted to make for a better knowledge elicitation tool?

Similarly, this project was carried out with the help of a real-world case study set in a Ford engine assembly plant. It enlisted real human decision makers to solve a real-world case problem using a VIS model that mimicked the real-world operations system as closely as possible. Unlike Bell and O'Keefe's (1995) reservation on conducting an experiment in a laboratory setting that is detached from reality, all the decision makers employed in this project have a good understanding of the real-world system and the decision-making that takes place regularly in it.

Eventually, the investigation that this project embarked on culminated in this thesis.

1.2 AN OUTLINE OF THE THESIS

This thesis begins by exploring the world of the Knowledge-Based Systems (KBS), a well-established domain where knowledge elicitation plays an integral and crucial role (Chapter 2). It includes looking into the terminology that is commonly used in the KBS literature, as well as the KBS development life-cycle and its associated problems. Meanwhile, the subject of knowledge elicitation and the various techniques that can be used to support it are also reviewed. Following this, the working relationship between KBS and VIS is explored (Chapter 3). In so doing, the evidence of using VIS to collect data for building KBS is collated; this establishes VIS as a valid knowledge elicitation tool, and hence contributes to answering the *first research question*.

Next, the scene is set for carrying out an investigation to find out if and how VIS can be improved as a knowledge elicitation tool (Chapter 4); which essentially aims to answer the *second research question*. These include explicating the constructs for assessing ‘elicitation improvement’, and then using them as a basis for forming the research propositions and specifying the research hypotheses. Furthermore, a methodology for executing the investigation and the following hypothesis tests is described briefly; it also serves to provide a structure for organising the subsequent chapters.

The methodology is comprised of a series of processes. Since some of them are independent of the others and can be activated in parallel with them, they are not entirely sequential. The details and outcomes of all work carried out in each of these processes make up the rest of this thesis. They are:

i. Understanding the case study (Chapter 5);

ii. Designing the experiment (Chapter 6);

iii. Building and assessing the VIS model (Chapter 7);

iv. Devising the measures for evaluating the four constructs (Chapter 8); and last but not least

v. Collecting and analysing the data (Chapter 9 and 10).

Finally, the thesis concludes with a summary and discussion of the results from the data analysis (Chapter 11). In addition, the limitations that were encountered throughout the investigation are reflected upon. Also, the opportunities that were identified for probable future research are discussed.

2

Knowledge-Based Systems and Knowledge Elicitation

This chapter provides a context, within which the research questions in Section 1.1 can be addressed. In essence, the what, where, why and how of knowledge elicitation are explored and explicated. It first begins with a background of knowledge-based systems, a well-established domain where knowledge elicitation plays an integral and crucial role. Next, it endeavours to propose working definitions for the basic terms used in the knowledge-based systems literature; these include knowledge engineering, knowledge acquisition and knowledge elicitation. Then, these terms are put into perspective through a basic knowledge-based systems development life-cycle model. Later, the problems at each process of the life-cycle model are discussed briefly, with an emphasis on the knowledge elicitation process – the focus of this thesis. Finally, a concise overview explaining how various techniques have been used to elicit knowledge is provided. As well, the area wherein this thesis makes a positive contribution is also unveiled in the overview.

2.1 KNOWLEDGE-BASED SYSTEMS

Knowledge-Based Systems (KBS) or expert systems originated from a field of study known as Artificial Intelligence (AI). The phase of the computer revolution that spawned KBS actually began in the early seventies, under the guise of computer hardware advances destined to send the price of computers plummeting below even the most optimistic scientist's prediction (Waterman, 1986). Whilst computer hardware specialists were developing microchip technology, software specialists were laying the groundwork for a conceptual breakthrough in a fledging field of Computer Science known as AI.

The goal of AI scientists has always been to develop computer programs that can solve problems in a way that is considered intelligent if done by a human. The first period of AI research is dominated by a naïve belief that a few general laws of reasoning coupled with powerful computers would produce expert performance. As experience accrued, the limited power of programs with general-purpose problem-solving strategies led to the conclusion that they were too weak to solve most complex problems (Newell, 1969). It seemed that the more classes of problems a single program could handle, the more poorly it did so on any individual problem (Waterman, 1986).

In response, the AI scientists then decided to reduce the scope of application by developing programs with general-purpose problem-solving strategies for narrowly defined problems. This new direction produced some successes but still no breakthroughs. Later, it dawned upon the AI scientists that the problem-solving power of a program came from the knowledge it possessed. That is, to make a program intelligent, it should be provided with lots of high quality knowledge that are specific to the problem area (Waterman, 1986). This realisation (a conceptual breakthrough) led to the development of special-purpose programs that were expert in some narrow problem areas. As these programs were meant to *solve problems* and *explain solutions* that would otherwise require an expert, they became known as expert systems. Also, as these programs possessed knowledge about some particular domains, they were also known as knowledge-based systems (Darlington, 2000).

2.2 KNOWLEDGE-BASED SYSTEMS TERMINOLOGY

In the KBS literature, knowledge engineering, knowledge acquisition and knowledge elicitation are three terms that are used frequently. Cordingley (1989), and Johannsen and Alty (1991) comment that they are usually not well defined and often appear to overlap, whilst Firlej and Hellens (1991) even claim that these three terms are used interchangeably throughout the KBS literature.

In most literature, knowledge engineering is a term used to describe the whole process of building a KBS: from the original investigation of the problem through to implementation (Edwards, 1991; Moody *et al.*, 1998; Turban *et al.*, 2005). In other words, it is to KBS what software/systems engineering is to conventional systems. A principle of knowledge engineering holds that whilst expert performance rarely conforms to some rigorous algorithmic process, it lends itself to computerisation. Hence, it follows that the essential tasks in knowledge engineering are expected to include those of ‘extracting, articulating and computerising’ the expert’s knowledge (Hayes-Roth *et al.*, 1983).

Unlike knowledge engineering, the definition of knowledge acquisition is more contentious. Buchanan’s *et al.* (1983) original definition of knowledge acquisition as ‘the transfer and transformation of problem-solving expertise from some knowledge source to a problem’ has lent itself to several interpretations. Firstly, Cordingley

(1989), and Johannsen and Alty (1991) interpret it as sharing the same breadth as knowledge engineering to cover the whole process. It includes the identification of the problem, its conceptualisation, formalisation, implementation, testing and prototype revision. Secondly, Liang (1992) and Jackson (1999) restrict their interpretation to include eliciting knowledge from experts, storing it in some intermediate representation and compiling it into some machine executable format. Thirdly, Edwards (1991) provides the narrowest interpretation by deeming knowledge acquisition as just the act of acquiring basic knowledge from the human expert.

Likewise, the definition of knowledge elicitation is also disputable. Firstly, Cordingley (1989) and Darlington (2000) define it simply as the process of obtaining knowledge about a domain from an expert; this is similar to Edward's (1991) interpretation of knowledge acquisition. Secondly, Johannsen and Alty (1991), and Moody *et al.* (1998) define it as one-half of a dichotomy of knowledge acquisition techniques that includes both manual (human-to-human) and semi-automatic (human-to-machine) means, with the other half being the automatic technique of rule induction. Rule induction is a special case of autonomous machine learning techniques that encompasses heuristics for generalising data types, candidate elimination algorithms, methods for generating decision trees and rule sets, function induction and procedure synthesis. It is described in more detail in Section 2.4.4.

As such, it is evident that there is a grey area when it comes to making a distinction between knowledge engineering and knowledge acquisition, and between knowledge acquisition and knowledge elicitation. Thus, taking advantage of the fact that KBS terminology is not cast in stone, a working definition for each of knowledge engineering, knowledge acquisition and knowledge elicitation is proposed here for the purpose of this thesis. Here, the general definition of knowledge engineering is adopted, where it is taken to mean the entire process of developing a KBS. For knowledge acquisition, Jackson's (1999) interpretation is adopted, where it is deemed to encompass knowledge elicitation, knowledge representation and knowledge execution. Finally, Darlington's (2000) definition for knowledge elicitation is adopted, where it is the process of obtaining domain knowledge from an expert. These definitions are illustrated more clearly through a basic KBS development life-cycle model described later in Section 2.3.2.

2.3 KNOWLEDGE-BASED SYSTEMS DEVELOPMENT LIFE-CYCLE

There are a few essential activities that have to take place when a KBS is being developed. These activities provide the basis for phases that collectively form the KBS' development life-cycle. To help establish the life-cycle of a KBS, Weitzel and Kerschberg (1989a and b), and Edwards (1991) suggest adopting a traditional Systems Development Life-Cycle (SDLC) model as a base first, on which modifications are then made to cater for the significant differences between the KBS and the conventional systems. In this respect, a waterfall model of SDLC is introduced initially in the next section. Then, some modifications are suggested, which later leads to the proposal of a basic KBS Development Life-Cycle (KBSDLC) model. Also, the potential problems that may crop up in the KBSDLC are reviewed.

2.3.1 SYSTEMS DEVELOPMENT LIFE-CYCLE

In the past, software development consisted of a programmer writing code to solve a problem or automate a procedure. Nowadays, systems are so big and complex that teams of architects, analysts, programmers, testers and users are required to work together to create millions of lines of code to drive the enterprises (Computerworld, 2007). As a result, a number of SDLC models were created to manage such mammoth undertakings. Dennis and Wixom (2003) observe that all SDLC models invariably have four fundamental phases: planning, analysis, design and implementation. Different systems development projects may emphasise different parts of the SDLC or approach the SDLC phases in different ways, but all projects' life-cycle will have elements of these phases. Royce's (1970) waterfall model is the oldest and the best known SDLC model, and a simplified version is shown in Figure 2.1. The model shows a sequence of phases where the output of each phase becomes the input for the next. In general, there are six phases in the model:

i. Feasibility and requirements definition

   This planning phase establishes a high-level view of the intended project and determines its goals. A feasibility study is next undertaken to determine whether the project should get the go-ahead. If the project is to proceed, then a project plan with budgeted estimates for the future stages of development is produced;

ii. Analysis

   This phase refines the project goals into defined functions and operations of the intended application. Requirements for the system is gathered via detailed study of the organisation's business needs, and analysis of end-users' information needs;

iii. Design

This phase describes the desired features and operations in detail, focusing on high-level design (what programs will be needed, and how will they interact), low-level design (how will the individual programs work), interface design (how will the interfaces look like) and data design (what data will be needed);

iv. Implementation

This phase translates the design into code, using whatever computer languages that are appropriate. Provisional versions of documentation, manuals and training materials will also be produced in this phase;

v. Testing

Normally, programs are written as a series of individual modules. This phase will bring all the modules together as a system, to check for errors, bugs and interoperability in a special testing environment. The system needs to be tested to ensure that interfaces between modules work (integration testing), the system works on the intended platform and with the expected volume of data (volume testing), and that the system does what the user requires (acceptance/beta testing); and

vi. Maintenance

This phase consists of making sure that the system runs in operational use and continues to do so for as long as is required. It includes correcting any undetected errors, enhancing the functionality of the system, and even moving the system to a different computing platform.

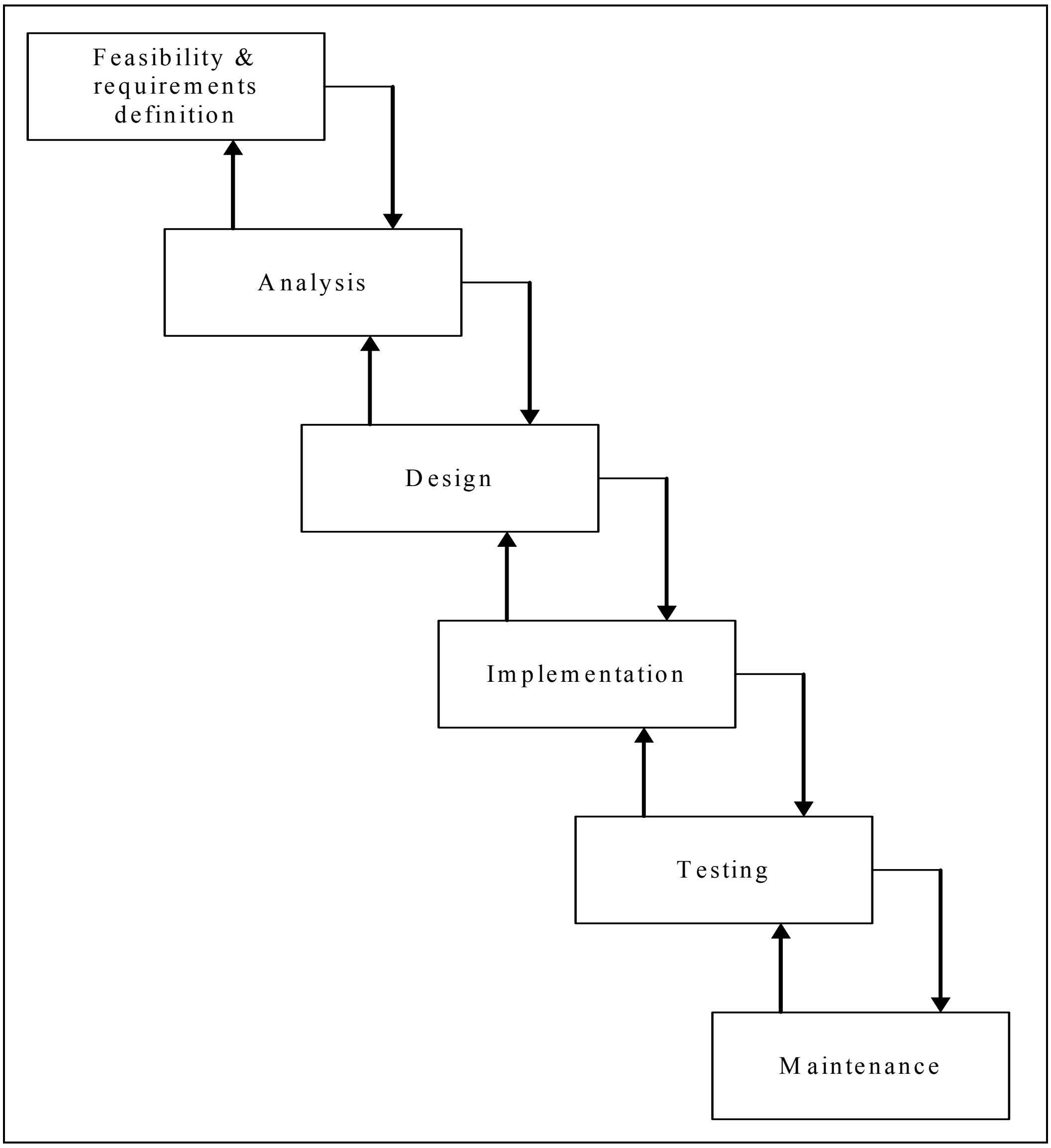


Figure 2.1: A simplified waterfall model of SDLC (Royce, 1970)

However, the waterfall model is not perfect and has its fair share of drawbacks. Mainly, the model assumes that the only role for users is in specifying requirements, and that all requirements can be specified in advance. It also assumes that system design is straightforward, and implementation is the real problem (Weitzel and Kerschberg, 1989a and b; Computerworld, 2007). Unfortunately, requirements do grow and change throughout the process and beyond, and a straightforward system design is rare.

Moreover, real projects seldom follow the sequential process illustrated in the model, which explain the feedback and iterative consultation allowed in Royce's (1970) waterfall model. In view of these drawbacks, many other SDLC models were developed later. They are usually variants of Royce's model (Weitzel and Kerschberg, 1989a and b) and include fountain, spiral, build and fix, rapid prototyping, incremental, and synchronise and stabilise (Computerworld, 2007). Nonetheless, in spite of its imperfections, the simpler original SDLC waterfall model will be used as a basic framework for adaptation into a provisional conceptual framework for developing KBS.

2.3.2 A BASIC MODEL OF KNOWLEDGE-BASED SYSTEMS DEVELOPMENT LIFE-CYCLE

Knowledge acquisition is defined earlier (Section 2.2) to encompass knowledge elicitation, knowledge representation and knowledge execution. Edwards (1991) identifies these activities as equivalent to the 'Analysis', 'Design' and 'Implementation' phases in a SDLC respectively. The correspondence between the knowledge acquisition activities and the relevant SDLC phases can be summarised in Table 2.1.

Knowledge acquisition activity	Work involved	Corresponding SDLC phase
· Knowledge elicitation	· Eliciting the basic knowledge from the human expert	· Analysis
· Knowledge representation	· Organising and structuring the knowledge	· Design
· Knowledge execution	· Codifying the knowledge into a machine-executable format	· Implementation

Table 2.1: Knowledge acquisition activities and their corresponding SDLC phases

Moreover, Weitzel and Kerschberg (1989a and b), and Edwards (1991) also suggest infusing the KBSDLC model with regular prototyping, which is characterised by iterative refinement that stresses fast development turnaround. It is because as an expert's conception of his[1] knowledge (such as the intermediate concepts used to monitor the 'state' of the solution, or even the reasoning process) tends to change with the KBS evolvement, such fast development turnaround would allow him to discover any shortcomings more quickly. Further to this, Weitzel and Kerschberg (1989a and b) suggest using the term 'processes' instead of (sequential) 'phases' to describe the KBSDLC model in order to emphasise its flexibility.

Applying these refinements onto the original SDLC waterfall model, a basic broad-brush KBSDLC model may be as adapted in Figure 2.2. The notions of knowledge engineering, knowledge acquisition and knowledge elicitation are illustrated clearly in the model. As well, the iterative refinement that is expected in each process is signified by the ring of arrows that encircles it. Processes in the life-cycle are activated initially by proceeding from the top of the model. A process can be reactivated to correct problems, before other processes have been activated for the first time. Also, the process in which problems are discovered does not necessarily constrain the process that needs to be activated. Therefore, a process can run concurrently with processes that are already activated, or it can be deactivated and reactivated at a later time. In this way, the KBS is actually evolving incrementally (Weitzel and Kerschberg, 1989b).

[1] The author recognises that a knowledge engineer or expert may be a female. However, in light of making this thesis a more pleasant and consistent read, only masculine pronouns are used. Any offence caused is deeply regretted.

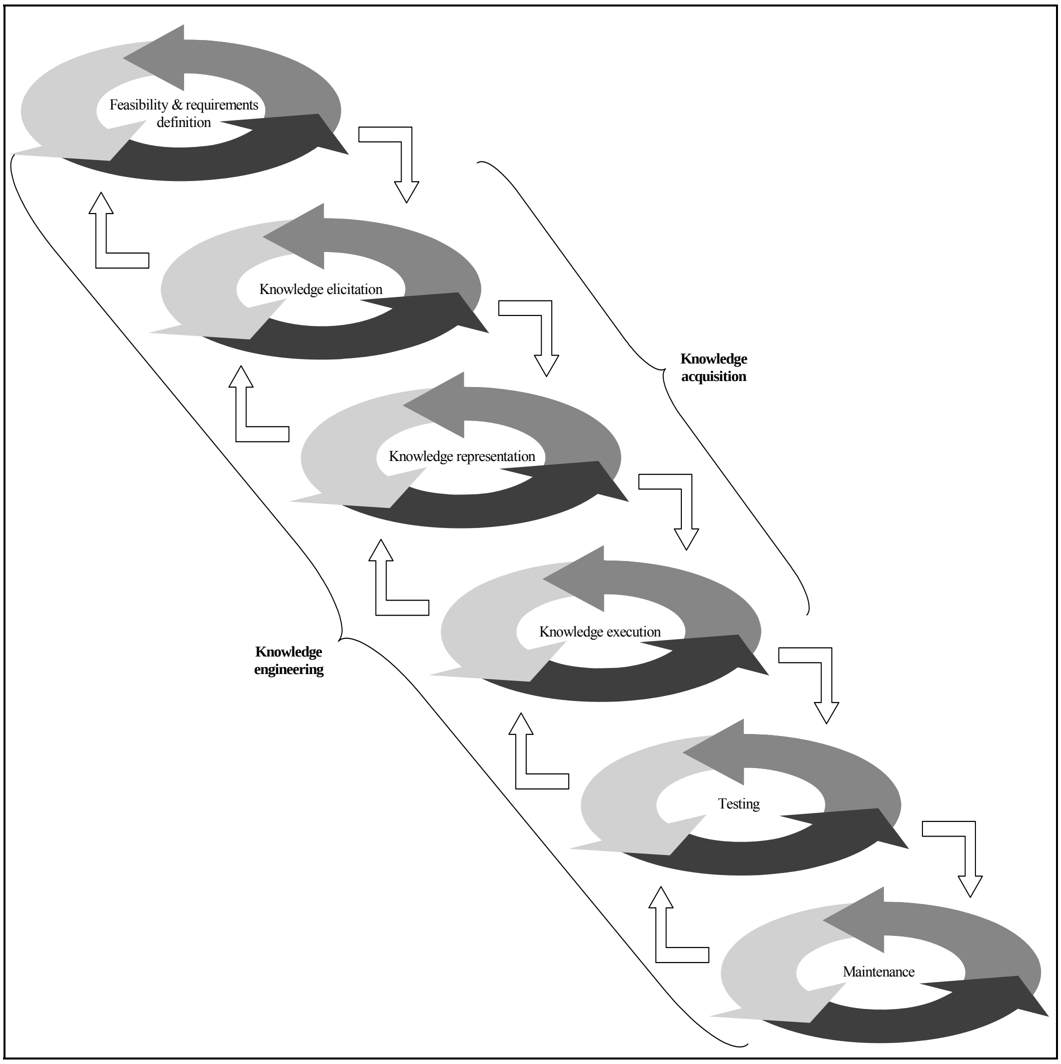


Figure 2.2: A suggested knowledge-based systems development life-cycle model

Briefly, in the feasibility and requirements definition process, the knowledge engineer and expert will work together to identify the problem area and define its scope. They will also determine the resources (human, time and computing facilities) required, as well as finalise the objectives of building the KBS. During the knowledge elicitation process, the knowledge engineer and expert will explicate sufficient key descriptions, relationships and procedures to describe the problem-solving process. In addition, strategies, subtasks, and constraints relating to the problem-solving activity are also specified. In the knowledge representation process, the knowledge elicited above will be organised and mapped into a formal representation. Next, the representation will be used to formulate rules that are then encoded in the knowledge execution process. These coded rules should embody the expert's knowledge and will define a prototype program capable of being executed and tested. Finally, testing involves evaluating the performance of the prototype program and revising it to conform to the standards set in the first process (Hayes-Roth *et al.*, 1983).

Though the KBSDLC model presented in Figure 2.2 is not a definitive version, it does not vary much from other proposed models. For instance, Barrett and Edwards (1995) mention that the BIS KBS methodology (from BIS Information Systems, a company) broadly resembles a waterfall approach to conventional development as it has the stages of feasibility, analysis, design, programming, testing and validation, and review. Like the suggested basic model, the BIS KBS methodology also permits the use of prototyping within many of the stages. In another instance, Madni (1988) suggests six stages in KBS development: knowledge elicitation, cognitive bias filtering, knowledge representation, software development and integration, system evaluation and validation, and advanced prototype expert system. Apart from cognitive bias filtering and advanced prototype expert system, the remaining four stages appear to be in line with the basic model.

2.3.3 POTENTIAL PROBLEMS IN DEVELOPING A KNOWLEDGE-BASED SYSTEM

A number of problems have been uncovered in each process of the life-cycle. McDermott (1983), and Weitzel and Kerschberg (1989b) reflect that *ad hoc* solutions for the problems in early processes seem to create new and even bigger problems in later processes. In other words, problems propagate. Notwithstanding, as the locus of this thesis lies within the knowledge elicitation process, this section (and the rest of the thesis) will concentrate mainly on its issues.

Problems in knowledge acquisition/elicitation process

The best known and most critical bottleneck in a KBS development lies within the knowledge acquisition phase, with particular stress on knowledge elicitation (Buchanan *et al.*, 1983; Breuker and Wielinga, 1987; Byrd, 1995; Moody *et al.*, 1998). It is critical because the power and utility of a KBS depends on the quality of the expert knowledge that is elicited and reproduced. Clancey (1986) points out that the process of eliciting knowledge from an expert entails more than the process of transferring a mental model lying within his brain into the mind of the knowledge engineer. It also includes formalising the expert's domain knowledge for the first time, which is an inherently difficult process due to the latter's tacit nature.

To incubate tacit knowledge, the expert needs to practice to become skilful, using rules of thumb or heuristics, learning which rules work and when they work. Through experience, he then develops judgement, insight, and informed opinions. It is the quality of this undocumented knowledge that is gleaned from his many years of experience in his particular field that determines his level of expertise (Kidd and Welbank, 1984). Unfortunately, when the expert is posed with a problem, he may be able to tell you his decision or diagnosis, but not the details of his thought process. He may even use certain knowledge without being aware that he has it. It is also very likely that he has never been required to formulate his decision-making, and he may have made many assumptions which are not stated explicitly. Furthermore, the expert can be surprised and even alarmed when the simple consequences of these assumptions are pointed out, and consequently he may be reluctant to admit to them (Jackson, 1985). In contrast, when the expert is asked for the factors that he had considered, he may list those which he thinks he ought to use, albeit they will not necessarily be the same as those he had actually used. However, this should not be construed deliberate deception; the expert will have learnt a lot of his knowledge through experience, and he may use it without being consciously aware of the explicit details. As such, tacit knowledge is also often referred to as compiled knowledge, whose elucidation and reproduction is usually much more central and difficult to the knowledge acquisition process.

At present, there is a wide range of techniques that are available to facilitate the knowledge elicitation process. They are discussed in detail in Section 2.4.

Problems in other life-cycle processes

In addition, there are also problems in other KBSDLC processes. They include determining whether the selected domain is appropriate for building a KBS (domain feasibility), and whether the expected costs and efforts are affordable (resource feasibility). Another problem might be finding out why a newly-built KBS fails to be accepted in the intended working environment and even fails to satisfy preset performance criteria. Last but not least, a KBS that requires extensive maintenance might also pose a problem if the system is so opaque and unstructured that it is hard to tell where updates and modifications should be applied (Breuker and Wielinga, 1987).

2.4 KNOWLEDGE ELICITATION TECHNIQUES

The knowledge elicitation process has been identified as a very critical bottleneck in the development of a KBS. This section looks at the ways that the elicitation process can be facilitated. Ideally, a conceptual framework of problem solving behaviour should be established as a prerequisite to the knowledge elicitation process. However, in its absence, the knowledge engineer can only try to use *ad hoc* means to understand in detail the concepts and relations used by the experts in their daily activities. Hopefully, the knowledge engineer is then able to construct a knowledge model whose contents and structure is very similar to that used by the expert, so that it can be used to support clear explanations and be an important part of the interface between the KBS and the expert (Clancey, 1986).

The types of knowledge that can be elicited are introduced first in the following sections. Then, the different techniques that a knowledge engineer may use to elicit an expert's knowledge are explained. They range from manual, semi-automatic to automatic techniques. Finally, these techniques, together with their strengths and weaknesses, are summarised appropriately.

2.4.1 KNOWLEDGE CATEGORISATION

Like the terminology used in the KBS literature (knowledge engineering, knowledge acquisition and knowledge elicitation), defining knowledge, information and data is also a disputable area. On the one hand, Naylor *et al.* (2001) subsume information and data under knowledge as both of them, together with structured information and insight, are considered to be different types of knowledge. On the other hand, Darlington (2000) deems knowledge as a derivative of information, which in turn is deemed a distillate of data. For the purpose of this thesis, the former and broader definition of knowledge is adopted. As such, knowledge ranges from its most factual form (data) to its most abstract form (insight).

Moreover, Turban *et al.* (2005) recognise that there are two major categories of knowledge: declarative and procedural. On the one hand, declarative knowledge can be thought of as ‘knowing that’ type of knowledge, which is essentially a descriptive representation of knowledge. It consists of related facts that can be organised and reorganised according to the occasion’s demands. An operative term for declarative knowledge is description. On the other hand, procedural knowledge can be thought of as ‘knowing how’ type of knowledge, which considers the manner things work under different situations. It includes step-by-step sequences and how-to type of instructions, as well as explanations. An operative term for procedural knowledge is procedure.

2.4.2 MANUAL KNOWLEDGE ELICITATION TECHNIQUES

Manual methods are basically structured around an interview of some kind. These include document analysis, interview, on-site observation, questionnaire and rating scale, teach-back interview, protocol analysis, walkthrough, card-sort, and last but not least, solution-characteristic matrix. As these methods are slow, expensive and sometimes inaccurate, there is a trend towards automating the knowledge elicitation process as far as possible. Semi-automatic and automatic methods are discussed later in Section 2.4.3 and 2.4.4 respectively.

Document analysis

Published documents such as books, papers and reports are good sources for acquiring general knowledge in well-established domains. For instance, Duan and Burrell (1995) remark that using published documents as a major source of knowledge is actually quite common in the marketing area. However, although documented knowledge may cover a wide range and is easy to access, it is limited to generalities. As such, a knowledge engineer cannot expect to rely solely on published documents to build a sufficient knowledge base.

Interview

An interview consists of interactions involving questions and answers between a knowledge engineer and an expert. In general, interviews provide a cheap but effective means of generating concepts, which are then used to produce a rough 'map of the territory' that covers the expert's domain. In addition, initial interviews also serve to develop some rapport between the knowledge engineer and the expert. As it is important to get the expert to communicate fluently, the exact form that an interview may take is not critical. Four possible types of interview are tutorial, unstructured, semi-structured or structured interviews.

In a tutorial interview, the expert will be asked to prepare an introductory talk outlining his domain, and deliver it as a tutorial session to the knowledge engineer. In an unstructured interview, where the control of the interactions lies mainly with the expert, he is given the freedom to cover topics that he deems fit. Here, the knowledge engineer only plays a facilitating role by encouraging the expert with general questions, probes and prompts. As digressions are usually tolerated, any material elicited is usually unpredictable and at times incoherent. Hence, the knowledge engineer has the additional burden of making the outcomes productive (Cordingley, 1989; Johannsen and Alty, 1991). In a semi-structured interview, the knowledge engineer works to a list of topics to be covered in the interview session, which does not specify the precise questions to be asked of the expert. In a structured interview, where the control of the interactions lies mainly with the knowledge engineer, he organises the communication between the expert and himself by working through a list of specific questions that are produced prior to the interview; thereby facilitating a systematic exchange of information. As such, the knowledge engineer's questions and the expert's answers are more restricted here than in less structured interviews (Moody *et al.*, 1998). Normally, no single type of interview is used to the exclusion of the others. It is because each interview type's relative applicability changes as the development process progresses. At the earlier stages of knowledge acquisition, tutorial, unstructured and/or semi-structured interviews are utilised to provide a general overview of the expert's domain. Once the process of knowledge acquisition is more advanced, structured interviews may be introduced to provide more specific focus.

Waterman (1986) mentions that the knowledge engineer may ask the expert to discuss, describe and/or analyse problems pertaining to his area of expertise during an interview. In a 'problem discussion' session, the knowledge engineer may pick a set of representative problems and discusses them with the expert. The goal is to determine how the expert organises his knowledge about each problem, represents concepts and hypotheses, and handles inconsistent, inaccurate, or imprecise data. During this discussion, the expert may introduce new concepts and relations. When this happens, the knowledge engineer will ask the expert to define these new constructs and relate them to the existing body of concepts and relations. In a ‘problem description’ session, the knowledge engineer will require the expert to describe a typical problem for each main category of answer that may arise. This helps the knowledge engineer to define a prototypical problem for each category of answer. This exercise may also suggest ways to organise knowledge hierarchically in the KBS. Finally, in a ‘problem analysis’ session, the knowledge engineer will ask the expert to solve a series of realistic problems and probe for the latter’s reasoning as the problems are solved. Here, the expert is required to describe the solution process and disclose as many intermediate steps as possible. The knowledge engineer will then question each step to determine the underlying rationale, including hypotheses that are entertained, strategies that are used to frame the hypotheses, and goals that are pursued to guide strategy selection.

Moody *et al.* (1998) comment that interviews are a pervasive technique as they can be used to elicit all types of knowledge. Nevertheless, depending on the dynamics of the interviews, the coverage of the expert’s area of expertise through interviews may still be incomplete and arbitrary. Also, Barrett and Edwards (1995) add that the expert may say what they wish to say, or what they think they are expected to say, rather than what they actually do. These suggest that interview aids or other complementary elicitation techniques should be used when possible. They include recording the interviews for subsequent reference, using labelled diagrams to help the expert to construct his talk, or even analysing protocols generated from the interviews.

On-site observation

Waterman (1986) explains that in on-site observation, a knowledge engineer will observe as an expert solves real problems on the job, rather than contrived but realistic problems in a laboratory setting. Here, the knowledge engineer will be observing passively and recording all observed information as accurately as possible. During the observations, the knowledge engineer will neither interfere with the expert's work, nor require much participation from the expert. In this way, the knowledge engineer may gain some insight into the complexity of the expert's domain. However, Barrett and Edwards (1995) warn that this technique is not feasible if the knowledge engineer and expert do not share a 'common ground'. Furthermore, on-site observation may not be practical for some domains, especially when there are time constraints or privacy concerns.

In addition, Johannsen (1989) also suggests a special hybrid of the interview and observation techniques, known as observation interview. In an observation interview session, the knowledge engineer will observe and note down the expert's activities as usual, and then try to clarify with the expert any queries that he has with the observations at the earliest instance. The queries may range from causes and reasons to consequences of the observed activities. In this way, observation interview is a powerful technique as whilst empirical data are being collected through observation, the knowledge engineer is also eliciting decision-making strategies concurrently through his what, how and why questions.

Questionnaire and rating scale

A questionnaire can be used instead of or in addition to an interview (Johannsen and Alty, 1991). It can be standardised in question-answer categories or it can be applied in a more formal way. In effect, a questionnaire is the equivalent of an interview in paper form, though it may not be as expansive or extensive. Similarly, Barrett and Edwards (1995) advise that a requisite for using this technique is that the knowledge engineer and expert need to share a 'common ground'.

A rating scale is a formal technique for evaluating single items of interest by asking the expert to cross-mark a scale. Verbal descriptions along the scale such as from 'very low' to 'very high', or from 'very simple' to 'very difficult' are used as a reference for the expert. A rating scale can either be used alone, or together with an interview and/or questionnaire (Johannsen and Alty, 1991).

Teach-back interview

Teach-back interview is a ready-made checking device by definition (Johnson and Johnson, 1987). It is a technique inspired by Ogborn and Johnson's (1984) conversation theory, which is concerned with the notions of concepts and understanding as entities that are made public by an interaction between participants. The theory posits that there are two levels of analysis to an interaction: Level 0 and Level 1. At Level 0, concepts are explored; whilst at Level 1, Level 0 concepts are reconstructed. For instance, if a Level 0 answer is an explanation of how to do an algorithm, then a Level 1 answer may be an explanation of why the algorithm works. That is, the latter is an explanation of an explanation. To begin, an expert must agree to be a participant in the role of interviewee, before a knowledge engineer can conduct an interview and attempt to find out what the expert knows. Next, both the knowledge engineer and the expert must contract to play the same game. Then, they must decide on the area of discussion (the domain) and on the medium of conversation (verbal, written, or doing something).

At the Level 0 analysis, the expert will first describe a procedure to the knowledge engineer, who will then teach it back to the expert in his terms and to his satisfaction. When the knowledge engineer and the expert agree that the former is doing the procedure the latter's way, it can be said that both of them share the same concept. In this way, this 'teach-back' procedure is a checking device where the expert is the final judge. It should be noted that both the knowledge engineer and the expert do not necessarily have the same thought process; they only agree that the same thing has been done.

At the Level 1 analysis, the knowledge engineer will ask the expert to give an explanation of how he reconstructed that concept and the 'teach-back' process continues until the expert is satisfied with the knowledge engineer's version. By then, the knowledge engineer is said to have understood the expert.

In essence, teach-back interview is an elaborated form of interview that takes place at two levels – generic and specific. At the generic level (Level 0), procedures define concepts, which lead to shared concepts after 'teach-back'. At the specific level (Level 1), reconstructions define 'memories', which lead to understanding after 'teach-back'.

Therefore, teach-back interview is a technique that can be used to produce an expert-authenticated database that is prejudiced minimally by the knowledge engineer's preconceptions. However, it is not a strongly structured technique that involves a lot of transcriptions and hence, consumes a lot of time. Also, teach-back interview is not universally applicable, especially when the expert tries to describe a manual skill or perceptual task on a piece of equipment, and his demonstration is not video-recorded. In such a case, any dialogue will be too context-bound (with meaningless statements like 'you press this and this happens') to make for a clear transcription.

Protocol analysis

Newell and Simon (1972) advocate that only the full complexity of verbal behaviour, as captured in a verbatim transcript, can do justice to the complexity of knowledge. This gives the premise on which (verbal) protocol[2] analysis is based. In essence, protocol analysis requires an expert to think aloud whilst working through a series of either simulated or real examples. Audio-recording is likely to be used in order to facilitate subsequent analysis.

In a 'thinking aloud' exercise, the expert is asked to report what he thinks about as he solves a problem as much as possible. The knowledge engineer intervenes only with non-directive reminders to keep the expert thinking aloud. In this way, the knowledge engineer hopes to conclude that the information reported is actually in the expert's focus of attention at the time and is untainted by any retrospection that would provide the

[2] There are two types of protocols: verbal and motor. Motor protocol analysis involves observing, recording (on a suitable media format) and analysing the physical performance of an expert. Most of the time, motor protocols are only useful when used in conjunction with verbal protocols.

opportunity for the expert to rationalise his thought processes. Notwithstanding, Newell and Simon (1972) warn it is also probable that there is much in the expert's mind that goes unreported as well. Thus, it is not always possible to draw direct conclusions about the limits of the expert's knowledge. As such, the 'thinking aloud' session is occasionally complemented with a cross-examination, where the knowledge engineer will ask probing questions about the expert's knowledge of particular topics. Cross-examination is particularly effective if the expert is highly articulate.

Eventually, the aim of a 'thinking aloud' exercise is to produce a verbatim transcript of the expert's explanation, from which knowledge is then elicited. As the expert encounters decision points in the task, he would have perceived certain conditions that resulted in him taking an action, thereby performing in an if-then manner. Hence, protocol analysis provides one method for capturing procedural knowledge (Moody *et al.*, 1998).

Nonetheless, protocol analysis does have its share of setbacks. The expert, with his many years of experience, may have compiled his knowledge such that a long chain of inferences is reduced to a single association. This feature makes it difficult for an expert to verbalise information that he actually uses to solve a problem. Further examples of knowledge that the expert may not think to mention include 'common sense' knowledge and general problem-solving strategies (Fox *et al.*, 1987). To worsen matters, Waterman (1986) cautions that if the expert is pushed to be more explicit, either during or after the problem-solving session, he may construct a line of plausible reasoning to explain his behaviour. This line may or may not reflect the actual problem-solving techniques used, hence incurring the risk of 'tainting' the knowledge elicited.

In addition, analysing verbatim transcripts is a laborious process that is both time-intensive and expertise-intensive. Last but not least, as protocol analysis is inherently an analysis of an individual expert, it is vulnerable to biases derived from his idiosyncrasies. In this light, the knowledge engineer should not use protocol analysis as his primary tool to elicit knowledge. Instead, he can use protocol analysis as a tool to elicit the broad structure of the expert's knowledge and the way it is applied, and supplement his findings with other techniques to fill in the details.

Walkthrough

A walkthrough is a term that describes the consideration of a process that is carried out in the actual environment at an abstract level. Johannsen and Alty (1991) comment that walkthrough is more detailed and often better than protocol analysis, as better memory cues are provided by being in the actual environment. Nevertheless, a walkthrough needs not necessarily be carried out in real time; indeed, it is even more useful in a simulated environment where states of the system can be frozen and additional questions pursued.

Card-sort

Card-sort is a technique used for eliciting the structural criteria that an expert uses to organise domain elements. It begins by typing elements of the problem domain (*e.g.* words, phrases, diagrams or pictures) on small individual index cards and spreading them randomly on a large table. Next, an expert is asked to sort the elements into as many small, mutually exclusive groups as possible. As most experts' spontaneous strategy in a card-sort exercise is to form slightly larger groups first before splitting them further, the expert will then be asked to split up the newly-formed groups into smaller sub-groups if there is a rationale to do so. Following this, the expert is asked to label each group, before amalgamating them back into slightly larger groups and re-labelling them.

During the entire exercise, the expert will be encouraged to think aloud the rationale that he uses to (re)group the elements. The expert's 'thinking aloud' should be recorded if he is agreeable, as the knowledge engineer will find these recordings a helpful *aide-memoire* for understanding the domain later.

In practice, Gammack (1987) advises that this technique will require the expert to make repeated sorts, whilst the knowledge engineer tries to derive the rules and classification relationships from these sorts. Gammack further adds that since the structural criteria are assumed to be derived from the expert's familiarity with the domain elements, they should reflect groupings that he finds convenient. Eventually, the elicited criteria are expected to be represented by a network of individual and/or groups of elements, and their relationships with each other.

Solution-characteristic matrix

Barrett and Edwards (1995) explain that a solution-characteristic matrix is used for each problem that may be faced by an expert. In a typical matrix, each row corresponds to a potential solution, whilst the columns show various characteristics of the solution that might make it appropriate for solving the problem. Also, these characteristics will be rated by the expert using a suitable measurement scale. In this way, a series of solution-characteristic matrices will be collected for a range of problems, so that a network of many-to-many relationships between problems and solutions can be created. As these matrices can be self-administered, the expert may be given the 'homework' of rating them to complement other elicitation efforts.

2.4.3 SEMI-AUTOMATIC KNOWLEDGE ELICITATION TECHNIQUES

Semi-automatic methods can be divided into two categories. On the one hand, there are those that intend to support the experts by allowing them to build knowledge bases with little help from knowledge engineers. On the other hand, there are those that intend to help the knowledge engineers by allowing them to execute the necessary tasks in a more effective or efficient manner. Two examples of semi-automatic methods are multi-dimensional scaling and repertory grid.

Multi-dimensional scaling

Multi-Dimensional Scaling (MDS) refers to a class of procedures that is used for extracting structure from a matrix of data. Gammack (1987) explains that these data are typically measures of relatedness among a set of objects which are presumed to vary along a number of unknown but interpretable dimensions. Since declarative knowledge (Section 2.4.1) often takes the form of related facts that can be organised and reorganised again as required by the situation, MDS is a suitable technique for eliciting it.

Broadly, an expert begins by comparing the objects with each other and provides some estimates of their perceived similarity. Next, the objects are scaled in a chosen number of dimensions to deliver a global picture of the space in which they reside. Then, using a spatial metaphor to represent similarity, the objects that are closer together are perceived to be more similar to each other than those that are further apart. Lastly, the nature of the dimensions is interpreted using information on the objects and their locations in the pre-specified space.

This technique works best when the objects are preselected on the basis that it is meaningful to rate their similarity. Also, these objects should be representative of the larger domain from which they are selected, and should form a fairly uniform set without including obviously anomalous items.

Repertory grid

The theoretical foundations of the repertory grid rest upon Kelly's (1955) Personal Construct Theory, which maintains that all human activity is a process of anticipating and interacting with events based on the framework of how one construes his past experiences. Moody *et al.* (1998) note that the constructs which a person uses to ascribe meaning to his experience facilitate his ability to distinguish between elements in his world, and these constructs are being adjusted continually contingent on whether they match what really occur. As the way a person interprets his experience determines how he sees the future, one needs to know the construct framework that supports the person's behaviour in order to know him. In this light, a repertory grid can be used to represent such a construct framework.

In essence, Shaw and Gaines (1987) explain that a repertory grid is a two-way classification of data which expresses part of a person's system of cross-references between his personal observations or experience of the world and his personal classification of these observations/experience. The repertory grid is composed of elements and constructs. On the one hand, elements are the things that are used to define the area of the topic, and they can be concrete or abstract entities. For example, in the context of interpersonal relations, the elements might be people. Before choosing a set of elements, a knowledge engineer must think carefully about the area of the topic and relate the elements to his purpose. In addition, these elements should be of the same type and level of complexity, and should span the topic as fully as possible. Also, care should be taken to ensure that each element is well known and personally meaningful to the expert; that is, each element must be central to him in the context of the particular problem. It is usual to start with about 6 to 12 elements. On the other hand, constructs are the terms used by the expert for describing how the elements are similar to or different from each other, and can be organised in contrasting pairs. They originate from various sources such as thoughts and feelings, objective and subjective descriptions, attitudes, and rules of thumb. As these terms only serve as memory aids for eliciting the expert's construct framework, their validity, label and description do not need public concurrence as long as they make sense to the expert. Hence, a two-dimensional grid of relationships can be produced by mapping the elements onto the constructs.

The most common method used for eliciting a pair of contrasting constructs is the minimal context form or triad method. The elements are first presented in groups of three; three being the lowest number that will produce both a similarity and a difference. Next, using a triad of elements, the expert is asked to state how two of them are alike and therefore different from the third. This is the emergent pole of the pair of constructs. The implicit pole may be elicited by the difference method where the expert is asked to state how the singleton is different from the pair, or by the opposite method where the expert is asked what the opposite description of the pair would be. Then, the remaining objects are rated along this dimension. This chain of activities is repeated until the expert can think of no other constructs, after which the resultant rating grid is analysed using cluster analysis.

In a way, a repertory grid encodes information about a person's way of looking at the world. Since the grid can be an aid for remembering the basis for decisions and actions, it is possible to use it in its own right for some purposes. Also, it can be analysed in a variety of ways to bring out possible underlying structures of a person's worldview and its relationship to those of others. Thus, repertory grids are a good means of eliciting declarative knowledge (Section 2.4.1).

Nonetheless, Cordingley (1989) warns that using repertory grid manually is tedious and time-consuming. Fortunately, semi-automatic computer-based versions had been developed and these have proved to be very effective at eliciting knowledge for KBS. In addition, as the repertory grid technique models the constructs of an individual, it is very personal and subject to change as the individual's experiences changes. Furthermore, Shaw and Gaines (1987) add that repertory grid appears to be more suitable for analysis (*e.g.* debugging, diagnosis, interpretation and classification) than for synthesis problems (*e.g.* design and planning).

2.4.4 AUTOMATIC KNOWLEDGE ELICITATION TECHNIQUES

Automatic methods are those intended to minimise or eliminate both the knowledge engineers' and the experts' contributions. In essence, these methods use computers to elicit or learn knowledge from existing data that preserve some historical decisions or experiences. As such, they are also known as knowledge discovery or machine learning methods (Turban *et al.*, 2005). Compared with manual or semi-automatic techniques, automatic techniques will expect to use less time to generate more consistent knowledge bases (Liang, 1992; Raghunathan and Tadikamalla, 1992).

Typical machine learning methods include rule induction, pattern matching, neural network computing and genetic algorithms. However, as neural network computing is an opaque technique where a neural network is essentially a black-box that does not allow rules to be extracted easily from it, it cannot support the explanation facility in a KBS. In addition, genetic algorithms are generally used for solving optimisation problems as opposed to the usual problems handled by a KBS (diagnostic, prediction or classification). Therefore, neural network computing and genetic algorithms are actually outside the ambit of KBS and are not discussed further in this thesis.

Rule induction

Buchanan *et al.* (1983) observe that human experts normally have more difficulty in stating procedural knowledge than stating declarative knowledge (Section 2.4.1). This varying difficulty may be attributed to at least two reasons. Firstly, an expert is likely to be less conscious of problem-solving strategies in their domain than they are of factual knowledge in the domain. Secondly, since an expert needs to have a detailed understanding of the problem-solving framework embodied in the domain in order to express procedural knowledge, he will find it harder to deal with the details involved in understanding the effects of even a small change to procedural knowledge. Hence, in a way, acquiring in-depth procedural knowledge may push the limits of the human expert's cognitive abilities. These limitations advocate the idea of using machine-based induction engines to develop procedural knowledge. This is also known as rule induction.

In principle, rule induction involves the use of an algorithm (induction engine) on a training set of example cases to induce a hierarchy of task-specific rules, which will constitute the knowledge base of a rule-based KBS. This set of rules may be either production rules that take the form of 'if-then' statements, or classification rules that take the form of a decision tree. Negnevitsky (2005) and Turban *et al.* (2005) define a decision tree as a map of the reasoning process; it describes a data set by a tree-like structure, and is composed of nodes representing goals and branches representing decisions.

Generally, relative to other knowledge elicitation techniques, the expert will find it easier to give example cases of different types of decisions, or to describe example cases of decisions that are already documented. These example cases are composed of decisions made by him and the characteristics or measurements (attributes) associated with them. However, the example cases will not state any assumptions and beliefs that the expert has made, nor any details of how he has assessed different evidence and resolved conflicts in order to reach his decisions (Hart, 1987).

Depending on both the algorithm that is used and the training set of example cases obtained from the expert, the induced rules may or may not be correct. Usually, if the algorithm is ‘efficient’, and the training set is ‘informative’, then the rules induced are expected to be ‘good’. Hence, great attention should be paid to the training set of example cases’ composition and use. Breiman *et al.* (1984) advise that an inadequate training set will produce results that are very sensitive to changes in the training set. As a reference, 50 example cases are required for a simple problem that has three decision classes and ten associated attributes, and 215 example cases are required for a more complex problem that has two decision classes and 19 associated attributes.

After the production or classification rules are generated, they are evaluated with both documented examples and expert interviews. The purpose of the interviews is to compare the findings of the induction with the expert’s version, to discuss the way in which attributes are used in the rules, to explain why certain attributes do not feature in the rules, and to question him about the areas of interest raised by the induction. All interviews should be recorded and a transcript drawn up. A flow chart based on the content of the transcript is then produced for approval by the expert. However, in practice, the expert might find it easier to test the rules by assessing their accuracy on actual examples rather than by examining them.

Pattern matching

Traditional KBS are predominantly rule-based systems. These systems are based on production rules ('if-then' statements) or classification rules (decision trees) that are either formulated directly from interactions with an expert, or induced from a collection of example cases provided by an expert. Their popularity stems from the view that an expert relies on a system of rules to solve problems.

Notwithstanding, it has also been observed that when an expert is posed with a problem, he may find it easier to provide a decision than to explain it. This is because the expert may have internalised the entire decision-making process after many years of experience (Section 2.3.3), such that he is able to reach a decision without assessing any circumstantial evidence or resolving any conflicts when confronted with similar problems. Hence, there is an alternative view that the expert may be using his experience instead of a system of rules to solve problems. This gives rise to Case-Based Reasoning (CBR) systems, whose knowledge base is made up of example cases of decisions. Like the example cases used for rule induction, an example case here also consists of a description (attributes) of a problem together with the decision that was taken in response to it.

CBR systems use a method of inference that is fundamentally different from traditional KBS. Instead of relying on a knowledge base composed of an intermediary representation of knowledge (*e.g.* the rules in rule-based systems), CBR systems are able to utilise the specific knowledge stored inside the example cases directly. In essence, a new problem is solved in CBR systems by using an appropriate pattern matching algorithm (inference engine) to retrieve an example case from a database of historical cases that resembles it most closely, and reusing the prior solution recorded in the retrieved example case. Admittedly, there will be some situations where the retrieved example case is vaguely similar to the problem at hand, and the prior solution is not sufficient. In these cases, the CBR systems will modify the prior solution appropriately and then put forward a proposed solution. If the proposed solution manages to solve the problem, then they are stored into the database for future use (Turban *et al.*, 2005). In this way, the knowledge base in CBR systems is dynamic, since it is updated whenever it is used.

Darlington (2000) comments that as the primary source of knowledge in CBR systems is experience (in the form of example cases) instead of theory, they are most useful in domains where the knowledge cannot be underpinned easily by any theoretical understanding. In addition, Cunningham (1998) suggests that CBR systems are considered effective only when their solutions are reusable, rather than being unique to each situation. Lastly, CBR systems are likely to be effective when the objective is to look for the best solution available, rather than a guaranteed exact solution.

2.4.5 A SUMMARY OF KNOWLEDGE ELICITATION TECHNIQUES

The various knowledge elicitation techniques discussed in the preceding sections show that there is no single technique which is able to elicit all types of knowledge by itself. Each technique has both advantages and disadvantages. Therefore, a good knowledge elicitation exercise should always use a few techniques that complement each other. The strengths and weaknesses of the manual, semi-automatic and automatic elicitation techniques discussed in the previous sections are summarised in Table 2.4, Table 2.3 and Table 2.4 respectively.

Manual technique (Rule of thumb)	Strength	Weakness
Document analysis	· Easy to access	· May not be specific and detailed enough to build a good knowledge base · May not be applicable to ill-defined domains
Interview · Use early to get terms of reference and possible framework) · If interviewing comes naturally to both parties, then interview techniques may be fruitful	· Gives knowledge engineer orientation to domain · Generates a lot of relevant material cheaply and in a natural manner · Little demand on expert other than time · Different types of interviews (tutorial, unstructured, semi-structured and structured) to suit needs of occasion	· Incomplete and arbitrary coverage · Requires training and/or social skills to be done properly · Burden of representation and interpretation on knowledge engineer
On-site observation · Can be improved by including interview in an observational session to form observation interview	· Provides first-hand insight into complexity of expert's domain	· Require 'common ground' with expert · May not be suitable in cases with time or privacy concerns
Questionnaire and rating scale · Used separately or with interview	· Completion of questionnaire does not require presence of knowledge engineer (asynchronous efforts)	· Require 'common ground' with expert · Does not cover as much depth or breadth as an interview can

Manual technique (Rule of thumb)	Strength	Weakness
Teach-back interview · Use to elicit global and specific structures	· Produces an expert's conception minimally prejudiced with respect to the knowledge engineer's preconceptions about the domain · Produces an expert authenticated fund of data that can be analysed and represented in several ways · A non-psychological, non-judgemental technique	· Not a strongly structured technique, so requires general interview training · Not appropriate when conversational approach becomes too long-winded · Heavy cognitive load on the investigator, so not recommended for the faint-hearted · Interviews are cumulative with transcription in between, and therefore time-consuming · Not universally applicable; especially when describing a manual skill or perceptual task on a piece of equipment
Protocol analysis (featuring 'thinking aloud' and cross-examination) · Use as an exploratory technique to build a knowledge base	· Yields a detailed picture of the representation of the expert's knowledge · Efficient way to elicit broad structure of knowledge and the way it is applied · Control remains with the expert	· Difficult to verbalise higher-order knowledge that is used · Time-intensive and expert-intensive · Subject to biases derived from individual expert's idiosyncrasies
Walkthrough · More detailed than protocol analysis	· May be carried out in actual environment, hence providing better memory cues; or · May use simulated environment, hence no need to be carried out in real time	· May be resource-intensive

Manual technique (Rule of thumb)	Strength	Weakness
Card-sort · Use to reveal possible hierarchical organisation and to reveal principles of that organisation	· Gives clusters of concepts meaningful to expert · Indicates possible uniting principles across abstraction levels · Provides hierarchical organisation, useful in indexing and placing new concepts · Splits large domain into manageable sub-areas · Easy for people to do, wide range of application	· Strict hierarchy may be too restrictive · Permits only one view per sort · Some aspects may become distributed and lost by technique
Solution-characteristic matrix	· May be self-administered · Facilitates creation of a network between problems and solutions.	· Does not cover as much depth or breadth as an interview can

Table 2.2: A summary of manual knowledge elicitation techniques

Semi-automatic technique (Rule of thumb)	Strength	Weakness
Multi-dimensional scaling (on relatedness measures) · Principled sets of objects should be used when trying to elicit criteria for differentiation · Good in (sub)domains when words may be inadequate for describing distinctions · Gives overall picture giving handle on domain, thus may be a useful alternative to rapid prototyping, *e.g.*, for feasibility	· Provides global picture of similarity of domain concepts · Indicates dimensions for distinguishing objects · Knowledge engineer's involvement unnecessary if suitable data already exist · Many computerised analysis techniques available · Allows comparison/ averaging across expert sources	· Results may be uninterpretable or not very useful · Supplementary analysis may be required to represent local information faithfully · Better at delivering 'structure' than 'content'
Repertory grid · Use with single expert in small set of closely related concepts, especially where no agreed vocabulary already exists	· Captures distinctions among closely related concepts useful to the expert · Elicits expert's personal concepts in absence of public vocabulary · Few, if any, constraints on subject matter, *e.g.*, can be done on perceptual and non-verbal data · Commercial software that enables repertory grid to be self-administered is available	· Distinctions may not be widely agreed · Manual elicitation is tedious · Larger concept sets require more expert time · Very personal technique, and model subject to change as one's experience changes

Table 2.3: A summary of semi-automatic knowledge elicitation techniques

Automatic technique (Rule of thumb)	Strength	Weakness
Rule induction · Use when an expert finds it hard to give detailed descriptions of his tacit knowledge and how he uses it	· Consistent and unbiased · Makes few assumptions about the underlying distributions in the data · Repeatable and indefatigable · May suggest or discover rules omitted by the expert · Identifies difficult, interesting, or contradictory example cases · Discovers knowledge away from the expert, providing the knowledge engineer with results, questions and hypotheses to form the basis of a consultation with the expert	· Uses only one form of reasoning · Produces rules without explanations · Cannot distinguish between necessary and confirmatory attributes · Assumes training set to be complete and correct · Cannot guarantee that induced rules are valid outside training set · Knowledge representation schema is pre-selected by the software used for induction
Pattern matching · Use where experience, not theory, is the primary source of knowledge · Use where solutions are reusable, rather than being unique to each situation · Use where the objective is best solution available	· Supports problem-solving based only on experience, without recourse to any theoretical understanding · Removes need for knowledge representation process in KBSDLC, as CBR systems do not rely on an intermediary representation of knowledge to work · Provides rapid response when an exact solution is not required · Supports an enhanced maintenance capability as knowledge base is dynamic	· Provides best solution available, which may be an approximated solution and therefore, sub-optimal

Table 2.4: A summary of automatic knowledge elicitation techniques

2.5 CONCLUSION

Knowledge elicitation is defined as the process of obtaining domain knowledge from an expert. It is part of a wider process known as knowledge engineering, which includes feasibility and requirements definition, knowledge representation, knowledge execution, testing and maintenance. In the knowledge engineering literature, the best known and most critical bottleneck in building a KBS is the knowledge elicitation process. This is largely due to the tacit nature of knowledge that resides within the experts.

At present, there are various techniques that are used to support the knowledge elicitation process. They can be classified into manual (document analysis, interview, on-site observation, questionnaire and rating scale, teach-back interview, protocol analysis, walkthrough, card-sort, and solution-characteristic matrix), semi-automatic (multi-dimensional scaling and repertory grid) and automatic (rule induction and pattern matching) techniques. These elicitation techniques should be regarded as complementary to each other, and not be used to the exclusion of others. The eventual combination of techniques employed in the elicitation process depends on the actual resources that are available to the knowledge engineer, and the type of knowledge to be elicited.

When automatic techniques are used, both the experts' and the knowledge engineers' participation in the knowledge elicitation process are minimised or even eliminated. It is because computers are used to learn knowledge directly from existing historical data in the form of example cases. If the example cases are not already available, then they have to be collected either manually, or via some computer-aided means. In the next chapter, the potential of using visual interactive simulation as a computer-aided means to collect example cases for machine learning is explored; this serves to answer the first research question in Section 1.1.

3 Visual Interactive Simulation, Virtual Reality and Knowledge Elicitation

Feigenbaum (1980), as paraphrased in Hayes-Roth *et al.* (1983), makes an empirical observation that a KBS derives its power from the knowledge it possesses, not from the particular formalisms and inference schemes it employs. The formalisms and inference schemes only provide the mechanisms to use the power. In other words, an expert's knowledge *per se* seems both necessary and nearly sufficient for developing a KBS. In the context of using automatic knowledge elicitation techniques (rule induction and pattern matching) to build a KBS, this empirical observation stresses the importance of collecting complete and accurate example cases to develop a powerful knowledge base[3]. However, research on collecting such an informative set of example cases appears to be limited. Liang *et al.* (1992) explain that this could be attributed to the assumption that the training data are readily available. Unfortunately, this is often not true, as some expert knowledge is difficult to obtain or needs to be collected on a real-time basis.

[3] In the case of using rule induction as a knowledge elicitation technique, there is an additional requirement of using an efficient learning algorithm to develop a powerful knowledge base. At present, most existing research about rule induction concentrates on developing and selecting induction methods (Liang *et al.*, 1992). Recent inductive learning algorithms developed include CLS, ID3, ACLS, C4.5 (based on ID3) and last but not least, C5.0 (an improved version of C4.5) (Jackson, 1999).

Traditionally, a training set of example cases is assembled manually, an ongoing task that is both laborious and time-consuming. Here, the expert may be asked for their decisions using the information that is presented in a physical document, such as an application form (Hart, 1987). Alternatively, the training set can also be assembled interactively with the help of computers (Davis, 1985). Although not more effective, this progress from manual to computer-aided assembly does provide some respite for the knowledge engineer by improving overall efficiency.

This chapter continues the search for a better computer-aided means of collecting a training set of example cases. It begins by turning the spotlight onto Simulation, a regular working partner with AI/KBS. The different terms of simulation used in this thesis are first explained. In addition, a brief section on virtual reality is also included to clarify any future references that are made to this area. Following this, past collaborations between AI/KBS and Simulation are briefly reviewed, in an attempt to draw preliminary conclusions on the potential of using the latter as a knowledge elicitation tool. These consequently contribute to answering the first research question in Section 1.1: *Is visual interactive simulation a valid tool for eliciting knowledge?*

3.1 VISUAL INTERACTIVE (DISCRETE-EVENT) SIMULATION

Computer simulation is defined as the simplified imitation (on a computer) of the operation of a real-world process or system over time, whose main objective is to facilitate experimentation for the purpose of better understanding and/or improving that system (Robinson, 2004; Banks *et al.*, 2005). As such, a simulation model is used as a vehicle for experimentation, through which the likely effects of various policies on a real-world system are demonstrated. Subsequently, the policy with the best results is implemented (Pidd, 2005).

In the field of Operational Research (OR), Pidd (2003) mentions that there are three different approaches to dynamic simulation modelling – Discrete-Event Simulation (DES), continuous simulation and mixed discrete/continuous simulation. Lately, agent-based simulation has emerged as the fourth approach. Notwithstanding, DES still appears to be the most popular approach as the majority of dynamic simulation applications uses it. Indeed, since its emergence in the late 1950s, DES has grown steadily in popularity to be recognised as one of the classical OR techniques that is used most frequently across a range of industries, such as manufacturing, travel, finance and health (Jeffrey and Seaton, 1995; Fildes and Ranyard 1997; Robinson, 2005; Hollocks, 2006). DES is so-named because the models built using this approach consist of discrete entities which occupy discrete states that only change at pre-determined discrete points in time. These discrete points in time are events that are decided upon when some conditions are fulfilled (Pidd, 2003; Banks *et al.*, 2005).

Hurrion (1976) introduces a new concept of DES known as Visual Interactive Simulation (VIS). It is the method whereby a DES model drives a display that represents the dynamic workings of the simulation. In addition, VIS also allows a user to interact with the model to view statistics and not least, carry out different experiments (O'Keefe and Pitt, 1991). Proponents of VIS cited some of its advantages as a decision-aiding tool to include better validation, increased credibility and model acceptance, better communication between the modeller and the client, incorporation of the decision maker into the model via interaction, and learning via ‘playing’ with the VIS model (Hurrion, 1980 and 1986; O’Keefe and Pitt, 1991; Chau and Bell, 1995).

Thus, it is imagined that if VIS makes the cut as a knowledge elicitation tool, these advantages would promote a reasonable level of good decision-making in the example cases collected. This provides an impetus to investigate VIS’ potential as a knowledge elicitation tool. At this juncture, a logical first port of call for this investigation is to look at past collaborations between AI/KBS and Simulation. However, before then, it is helpful to digress momentarily from the main plot and introduce some basic terms related to virtual reality. The latter is a subject that is associated closely with recent development in the visual aspect of VIS, and referenced regularly in the rest of this thesis.

3.2 VIRTUAL REALITY SYSTEMS

Virtual Reality (VR) is defined as a computer-generated three-dimensional environment created using virtual environment systems, and can be interactively experienced and manipulated by the participants (Barfield and Furness, 1995). Stuart (2001) explains that a virtual environment system is a human-computer interface capable of providing ‘interactive immersive multi-sensory three-dimensional synthetic environments’. It is supported by ‘interactive computer simulations that sense the participants’ position and actions, and replace or augment the feedback to one or more senses, giving the feeling of being mentally immersed or present in the simulation’ (Sherman and Craig, 2003). Barnes (1996) terms this virtual environment that represents an existing or planned environment realistically, wherein some or all of the objects are animated with behaviour controlled by a simulation engine, a VR world.

To immerse these participants in a VR world, specialised VR equipment is used. For instance, a head-based display known as the Head Mounted Display is used to provide stereoscopic VR views, and gloves that contain flexible fibre optic cabling plus sensors are used to register complex combinations of locational information in the VR world (Barnes, 1996). Nowadays, the concept of a VR system is broadened to include non-immersive VR. Vince (1998) describes a non-immersive VR system as one that uses the desktop system, and does not require any specialised VR equipment. As a result, the participants will not have any sense of immersion, nor any perception of scale. Interaction with the non-immersive VR world is facilitated by conventional means such as the keyboard, mouse and trackball.

3.3 WORKING RELATIONSHIP BETWEEN AI/KBS AND SIMULATION

There has always been considerable interest given to AI working with(in) Simulation, and *vice versa*. To date, several taxonomies depicting such interest have been published, two of which are O’Keefe (1986) and Ören (1994). Ören (1994) lists two types of activities that use Simulation in AI: use of simulation for applications of AI, like evaluating a KBS (Flitman and Hurrion, 1987; Shaw, 1989; Chryssolouris *et al.*, 1991; Liang *et al.*, 1992); and cognitive simulation, where systems with cognitive abilities are simulated. Such systems include humans and autonomous robots. Also, Ören lists two types of activities that use AI in Simulation: AI-assisted simulation and

AI-based simulation. On the one hand, in AI-assisted simulation, AI techniques are used to provide computer assistance in areas such as formulating models and designing simulation experiments. On the other hand, in AI-based simulation, AI techniques are used to generate model behaviour in simulation runs (Flitman and Hurrion, 1987; O'Keefe, 1989; Williams, 1996; Lyu and Gunasekaran, 1997; Robinson *et al.*, 1998; Kunnathur *et al.*, 2004).

In a similar vein, O'Keefe (1986) develops a taxonomy for combining KBS and Simulation, as shown in Figure 3.1. In it, seven combinations are proposed:

a. Embedding a KBS within a simulation model (Caprihan *et al.*, 2006);

b. Embedding a simulation within a KBS model;

c. Separate KBS and simulation model working interactively in parallel, with the user having access to the simulation model (Flitman and Hurrion, 1987; O'Keefe, 1989; Williams, 1996; Lyu and Gunasekaran, 1997; Robinson *et al.*, 1998; Kunnathur *et al.*, 2004);

d. Separate KBS and simulation model working interactively in parallel, with the user having access to the KBS (Shaw, 1989; Jeong and Kim, 1998; Mak *et al.*, 2002);

e. A KBS and a simulation model working in a cooperative manner, where the user has access to both KBS and simulation model (Flitman and Hurrion, 1987; Wu and Wysk, 1990);

f. A cooperative KBS and simulation model sub-system that is embedded in a larger system, where the user has access to both KBS and simulation model (Jeong, 2000); and finally

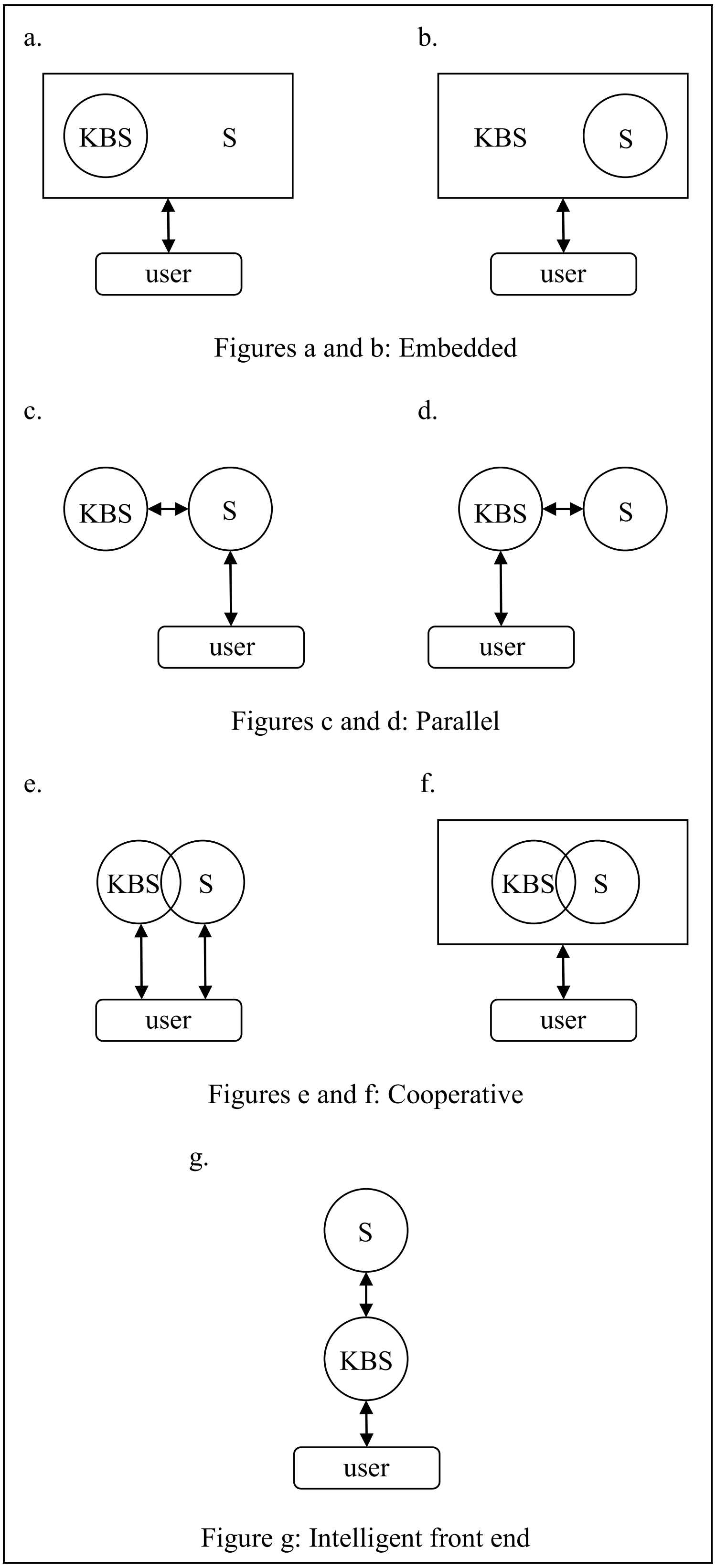
g. Using KBS as an intelligent front end that sits between the user and a simulation package (Hurrion, 1991). In this combination, the KBS serves to generate the necessary instructions to use the simulation package following a dialogue with the user, and interprets and explains results returned from the package to the user.

3.4 EVIDENCE OF SIMULATION AS A KNOWLEDGE ELICITATION TOOL

The evidence to support the use of simulation for knowledge elicitation purpose stems mainly from two collaborations coined by Ören (1994) in the last section: (1) Using simulation for applications of AI; and (2) AI-based simulation. They are discussed and reflected upon further below.

3.4.1 EVIDENCE FROM USING SIMULATION FOR APPLICATIONS OF AI

Anecdotal evidence of using Simulation to elicit expert knowledge are found in Flitman and Hurrion (1987), Shaw (1989), Pierreval and Ralambondrainy (1990), Chryssolouris *et al.* (1991), Hurrion (1991), Liang *et al.* (1992), Tan *et al.* (2000), and last but not least Tan (2003). These are a few examples of what Ören (1994) classifies as using Simulation for applications of AI.



Figures a and b: Embedded

Figures c and d: Parallel

Figures e and f: Cooperative

Figure g: Intelligent front end

Figure 3.1: A taxonomy for combining knowledge-based systems (KBS) and simulation (S) (O'Keefe, 1986)

In Flitman and Hurrion (1987), and Hurrion (1991), the authors built a VIS model for the operations of a simple coal-yard depot. Embracing a gaming mode, the user took on the role of a depot manager and controlled the depot's operations in the model. In the meantime, a KBS linked to the VIS model would monitor and record all user actions. The data thus obtained were then used for machine learning to develop the KBS' knowledge base. Also, Liang *et al.* (1992) employ VIS in a gaming mode to collect real-time scheduling decisions. These decisions were then used to facilitate learning in an automated knowledge acquisition process which integrated semi-Markov processes with neural network computing.

Shaw (1989) describes and demonstrates a 'learning by experimentation' methodology. Following the methodology, a flexible manufacturing system was simulated and alternative scenarios employing different scheduling rules were tested for each selected hypothetical state of the system. The scheduling rule that produced the best performance for a state would become the rule to be deployed whenever the system assumed that state. As such, a collection of state-rule pairs could be generated as a training set for learning scheduling knowledge. A similar methodology is also applied by Pierreval and Ralambondrainy (1990) on a simplified flow shop example.

Chryssolouris *et al.* (1991) begin the learning process for building a neural network by running several simulations of a job shop. In these simulations, the operational policy (weights of the decision-making criteria) and the workload (mix and volume of job types) were varied, and performance measures were collected at the end of each run. The performance measures plus workload parameter values would constitute the input component of an input-output pair, whilst the policy parameter values would constitute the output component. In this way, a set of input-output pairs could be collected from these simulations to learn the knowledge of selecting operational policy.

Lastly, Tan *et al.* (2000) planned to investigate the validity and reliability of using an interactive, simulation-driven immersive VR system for collecting data, which would then be used for learning human behavioural rules. Subsequently, Tan (2003) performed some method-comparison studies between the data collected using the VR system with the data from direct observation, and concludes that there is some evidence supporting VR as a suitable technology for collecting data. However, the original plan to learn human behavioural rules was not carried out.

3.4.2 EVIDENCE FROM AI-BASED SIMULATION

Further circumstantial evidence is found in a string of research that belongs to what Ören (1994) classifies as AI-based Simulation (Section 3.3). In spite of Simulation's success and its increasing use in a large number of application areas, modelling complex systems that include some elements of human decision-making has proved to be problematic (O'Keefe, 1989; Robinson, 2007). It is this shortcoming that has fuelled research in AI-based simulation.

AI-based simulation as defined by Ören (1994) can be regarded as an equivalent of the third combination (Figure 3.1c) proposed by O'Keefe (1986). Under this combination, the author explains that the KBS and simulation model are designed, developed and implemented separately in parallel. In addition, there is a facility between both of them that supports interaction, enabling the simulation model to interrogate the KBS. This form of collaboration is useful where a simulation model is developed for a complex system, and a KBS already exists for part of the decision-making within this system. Therefore, the simulation model can avoid the need to encode the decision rules from scratch and simply access the KBS to simulate the decisions.

To date, Robinson (2003) remarks that much research has been carried out to link a bespoke simulation model with a bespoke KBS in various application areas (Flitman and Hurrion, 1987; O'Keefe, 1989; Hurrion, 1991; Williams, 1996; Lyu and Gunasekaran, 1997, Kunnathur *et al.*, 2004). However, none of the research seems to involve the use of standard Commercial Off-The-Shell (COTS) software packages. This view is also shared in Williams' (1996) comment that there appears to be little work done in linking specialised software from disciplines like VIS and KBS. In response, Robinson *et al.* (1998) then successfully linked a COTS VIS package (Witness) to another COTS KBS package (Xpertrule) for the purpose of representing human decision-making in a fictional truck loading bay example. This research later set the stage to develop the Knowledge-Based Improvement (KBI) methodology (Robinson *et al.*, 2001 and 2005; Alifantis, 2006), which applied Robinson's *et al.* (1998) findings in a real-world setting. Whilst the KBI research's objectives are not to prove VIS' knowledge elicitation capability, it did incidentally demonstrate that VIS can be used to collect example cases which are good enough for knowledge acquisition purposes.

3.4.3 A RETROSPECTION OF EVIDENCE

The above examples illustrate that Simulation/VIS have been used to elicit episodic knowledge in the form of example cases from the experts. They also show that the collected example cases have been used successfully for machine learning. Therefore, using Simulation/VIS for knowledge elicitation/acquisition is a tried and tested concept, which affirmatively answered the first research question duplicated at the beginning of this chapter.

In particular, Robinson's *et al.* (1998, 2001 and 2005) and Alifantis' (2006) work also point out that with appropriate adaptation, a COTS VIS package can be used for knowledge elicitation. Moreover, Tan's (2003) work on using an interactive, simulation-driven VR system to collect data suggests the idea of investigating the value of using VIS supported with an appropriate visual representation to create a quasi-VR system for eliciting expert knowledge. Finally, Flitman and Hurrion (1987), Hurrion (1991) and Liang *et al.* (1992) demonstrate the idea of using VIS in a gaming mode to elicit expert knowledge.

3.5 CONCLUSION

VIS is a widely-used variant of DES, which is recognised as the most popular approach to dynamic simulation modelling in the field of OR. In the former, a DES model drives a display that represents the dynamic workings of the simulation. In addition, VIS also allows a user to interact with the model in order to view statistics and not least, carry out different experiments to better understand and/or improve the real-world system that is simulated.

Due in no small part to the many credits given by VIS proponents, there have been several attempts to widen the scope wherein VIS can be applied. An area where VIS research has taken an interest, and *vice versa*, is AI/KBS. A review of AI/KBS collaborations with Simulation/VIS shows that the latter has been used to collect data in the form of example cases to facilitate machine learning. This finding establishes VIS as a valid knowledge elicitation tool, and hence contributes to answering the first research question (Section 1.1). Also, it provides an important basis to support further research on improving VIS as such a means. Moreover, the review identifies a few notions on where or how it might be conducted.

In the next chapter, the scene is being set for carrying out a study to find out how VIS can be improved as a knowledge elicitation tool, which essentially aims to answer the second research question in Section 1.1. To begin, some basic terms along with the constructs that will be used for assessing 'elicitation improvement' are explicated. Then, the two ways in which VIS might be improved are suggested. Together, these provide the premise for laying down the research propositions and framing the hypotheses. Lastly, the methodology that will be used to test the hypotheses is outlined.

4 Research Propositions, Hypotheses and Methodology

A review of published literature on AI/KBS and Simulation/VIS collaborations in the last chapter shows that Simulation/VIS has been employed to collect example cases, with an intention to use them for building a knowledge base. Notwithstanding, the research on using Simulation/VIS to elicit expert knowledge is still considerably sparse. This apparent paucity of literature suggests that there is still a lot of margin for developing its use. In view of this void, a next step would be to find out how VIS can enhance its ability to elicit expert knowledge.

A lead is found in the 'V' and 'I' of VIS. However, before this lead is explored in more depth, it is helpful to first explain the elements that make up an example case collected from using a VIS model as a knowledge elicitation tool. As well, the constructs that will be used for assessing 'elicitation improvement' in the set of example cases collected are identified. Together, these provide the basis for laying down the research propositions and framing the hypotheses for testing in this thesis. Finally, the methodology used to test the hypotheses is introduced. Essentially, the scene is being set for answering the second research question in Section 1.1: *How can VIS be adapted to make for a better knowledge elicitation tool?*

4.1 COMPOSITION OF AN EXAMPLE CASE

An example case collected during a knowledge elicitation session describes the *scene* that was recorded when the expert interacted with the VIS model. Each example case thus collected is made up of two parts: a decision element and an attribute element. On the one hand, the decision element is a set of decisions made by the expert when he interacts with the VIS model. On the other hand, the attribute element is a corresponding set of attributes that describes the state in the VIS model when the interaction takes place.

4.2 CONSTRUCTS FOR ASSESSING 'ELICITATION IMPROVEMENT'

There are two aspects in which a knowledge elicitation process can be improved: effectiveness and efficiency. Whilst there are three distinct views of elicitation effectiveness: decision fidelity, state space and case quantity; there is only one view of elicitation efficiency: collection rate. These views are elaborated below.

4.2.1 CONSTRUCT ONE: DECISION FIDELITY

In their work on the KBI methodology, Robinson *et al.* (2001 and 2005) and Alifantis (2006) carried out a collaborative study with Ford Motor Company to devise a VIS-based means for identifying and improving human decision-making. In their conclusions, the authors recognise that the experts may take less realistic decisions in a simulated environment, as they are quite likely to take greater risks when there are no real consequences to their decisions. This observation relates to the proximity to reality of the decisions elicited from experts, and provides the basis for the first view of elicitation effectiveness. Therefore, an example case is considered to have a high degree of *decision fidelity* if its decision element bears close resemblance to the decision that the expert would have made in a reality described by the corresponding attribute element.

4.2.2 CONSTRUCT TWO: STATE SPACE

Negnevitsky (2005) states that 'a training set (of example cases) must cover the full range of values for all inputs'. This criterion relates to the adequacy of the range of situations from which the example cases are collected for training a knowledge base, and provides the basis for the second view of elicitation effectiveness. Therefore, given the definition for attribute element in Section 4.1, a set of example cases is considered to have a large *state space* if their attribute elements collectively cover a wide range of values for all attributes.

4.2.3 CONSTRUCT THREE: CASE QUANTITY

Moreover, Negnevitsky (2005) also states that 'the training set (of example cases) has to be sufficiently large'. This criterion relates to the size of the set of example cases collected for training a knowledge base, and provides the basis for the last view of elicitation effectiveness. Therefore, *case quantity* refers to the total number of example cases recorded in a knowledge elicitation session.

4.2.4 CONSTRUCT FOUR: COLLECTION RATE

Finally, in Robinson *et al.* (2001 and 2005) and Alifantis (2006), the authors also recognise that the experts may find it a very laborious and time-consuming experience to provide a full set of data, comprising of a very large quantity of wide-ranging decisions. As such, besides providing additional support for the second and third views of elicitation effectiveness, this observation relates to the expediency of the elicitation process in real-time, otherwise known as 'elicitation efficiency'. Therefore, *collection rate* is the number of example cases recorded per unit of real-time in a knowledge elicitation session. In this thesis, a unit of real-time is taken to be one minute.

4.3 FACTORS FOR IMPROVING VIS AS A KNOWLEDGE ELICITATION TOOL

As the 'V' and 'I' of VIS suggest, there are two main factors that differentiate VIS from other forms of Simulation: (1) the Visual representation that is used to illustrate the dynamics of the simulation model; and (2) the facility that encourages and enables a user to Interact with the simulation model and experiment with different scenarios (Hurrion, 1986; O'Keefe and Pitt, 1991). This visual interactive approach has ceased making Simulation a black-box technique, and opened up the method for management to look and experiment inside. With this approach, Simulation becomes a transparent-box that has greatly relieved the problems of communication and model credibility (Hurrion, 1980 and 1986). Here, these two factors also provide the initial directions on where the investigation on improving VIS' ability to elicit expert knowledge might begin. These factors, together with the associated research propositions being investigated in this thesis, are discussed further below.

4.3.1 FACTOR ONE: VISUAL REPRESENTATION

'A picture is worth a thousand words' is an adage, which refers to the idea that complex stories can be told with just an image, or that an image may be more influential than a substantial amount of text (for instance, if-then production rules used in DES). Indeed, visualisation is thought to have elevated DES to another level, with Hurrion (1986) extending the adage to 'a colour video sequence is worth a thousand pictures, whilst an animated interactive model is worth a thousand video sequences'. Two perspectives to the concept of visual representation are discussed below: mode and dimension.

Visual representation mode

Hurrion (1986), and O'Keefe and Pitt (1991) have identified two forms of dynamic visual representation commonly used in VIS packages: schematic/iconic animation (iconic) and logical/dynamically changing graphic (graphical). With an iconic representation, operational characteristics of the system under study are mimicked. Here, icons representing entities move through the display as time progresses. This is the mode of visual representation used by most VIS applications. On the other hand, with a graphical visual representation, logical representations such as bar charts, time series and histograms are used to summarise and display graphically the performance measures of the system as time progresses (Liang *et al.*, 1992).

O'Keefe and Pitt (1991), and Bell and O'Keefe (1995) conclude that users have a strong preference for either the iconic or graphical visual representation, over the third mode of visual representation – a listing of performance measures (not mentioned above, and generally not used in VIS packages). O'Keefe and Bell (1992), and Bell and O'Keefe (1994 and 1995) go further to conclude that iconic representation usage had consistently been related to a feasible (as opposed to correct or optimal) solution that was an improvement over that considered prior to using VIS. From their conclusions, it can be inferred that decision makers have more confidence in solutions obtained from using iconic representation.

In addition, Chau and Bell (1995) also conclude that more effective and efficient decision-making takes place when a visual display shows a 'paired systems' iconic VIS model which allows the comparative performance of two systems to be observed, in contrast with one that only shows a 'single system' iconic VIS model at any time. The authors further conclude that in any case, the use of a VIS model promotes better decision-making than a traditional (non-iconic and non-graphical) simulation model embedded in an interactive interface does.

These findings suggest that a simulation model with a visual representation is a superior decision-aiding tool to one without a visual representation in the contexts studied. Moreover, they show that VIS with an iconic representation is a better decision-aiding tool than VIS with other forms of visual representation. As such, it is fair to claim that VIS with an iconic representation would make for a stronger candidate for eliciting expert knowledge than other forms of Simulation. Nonetheless, one might still ask how an iconic representation can be improved further to enhance its ability to aid decisions or elicit expert knowledge. As previously mentioned in Section 3.4, Tan (2003) has reported some evidence in support of VR as a suitable technology for collecting data. This springs up the idea to investigate whether an enhanced visual fidelity is one such means.

Visual representation dimension

Ideally, it seems desirable to present information on the visual display with characteristics similar to the objects that are perceived in a real-world environment. An expert can then use the same processes that he uses when perceiving objects in the real-world environment. However, generating real-time images incurs a high cost. Moreover, such a degree of realism is often unnecessary when considered against the actual needs of an application (Preece, 1994). Using a flight simulator as an example, it is less important to deceive pilots into believing that they are flying through real terrain than it is to provide all the necessary information in the right format to allow them to function as if they are in a plane.

Hence, the research will not be about improving VIS models to look as lifelike as possible. Instead, the focus is on the efficacy of different representational dimensions in terms of the function(s) they are intended to support. With respect to the visual representation dimensions that are currently available in COTS VIS packages, the influence of a low-level 2-Dimensional (2D), mid-level 2½-Dimensional (2½D), and high-level 3-Dimensional (3D) representation will be studied[4]. Here, a 2½D representation consists of three-dimensional icons displayed against a plain, two-dimensional background, with no perspective projection (where more distant objects are drawn smaller relative to those that are closer to the eye) or any other efforts at creating photo-realism (Wenzel *et al.*, 2003; Akpan and Brooks, 2005a). On the other hand, 3D representation consists of three-dimensional photo-realistic icons displayed against a three-dimensional photo-realistic background, with perspective projection. In addition, a user is able to manipulate the view of the virtual environment by using a mouse. In effect, a VIS model supported with a 3D iconic representation (3D-VIS) resembles a non-immersive VR system (Section 3.2).

The empirical evidence showing that a simulation model with a 2D iconic representation is a better decision aid than one without a visual representation, or with a 2D representation of another mode is given in the last section. However, there is a dearth of empirical studies committed to comparing simulation models with 2D, 2½D or 3D representations (Akpan, 2005; Akpan and Brooks, 2005a and b). This apparent lack of empirical evidence is in spite of the many praises that practitioners often heap onto 3D (over 2D) representations (Barnes, 1996 and 1997; Waller and Ladbrook, 2002).

First and foremost, Barnes (1996) mentions that an animated VR world (Section 3.2) can provide a participant with an experience that appears quite realistic. Further to this, the participant would be expected to develop a strong sense of involvement and participation if the VR world is interactive as well. Barnes further adds that if the facility supporting the interactions is able to quickly reflect any changes made to the VR world in the behaviour of the animated VR world, then these interactions would parallel closely with those in the real world. As an extension to Barnes' view, it may be argued

[4] See Section 7.3 for illustrations of equivalent VIS screen shots in these three visual representation

that the decisions made by an expert during his interactions with a 3D-VIS model (which resembles a non-immersive VR system) in a knowledge elicitation session would bear the closest resemblance to those in the real world. This would be followed by a less realistic 2½D-VIS model (VIS model with a 2½D iconic representation) and then the least realistic 2D-VIS model (VIS model with a 2D iconic representation). This argument leads to the first proposition in the thesis:

Proposition 1 – *A higher dimension of iconic representation would demonstrably improve the degree of decision fidelity in the example cases collected in a knowledge elicitation session.*

In addition, Barnes (1996, 1997) mentions that simulation with a 3D iconic representation allows for a better communication and visualisation of ideas and concepts to the users. Moreover, the author argues that the use of high dimension representation makes it easier to understand what the simulation represents. These sentiments are also echoed by other practitioners (Waller and Ladbrook, 2002; Akpan, 2005; Akpan and Brooks, 2005b). In a survey study with 57 usable responses, it is found that a majority of respondents agreed a 3D-VIS model enhances communication between the modeller and user better than a 2D-VIS model can (Akpan, 2005; Akpan and Brooks, 2005b). Likewise, the authors also find that a majority of respondents thought it is easier to understand the system that is mimicked in a 3D-VIS model than in a 2D-VIS model. This ease of understanding might explain another finding, where a majority of respondents opined it is easier to uncover inaccuracies in a 3D-VIS model than in a 2D-VIS model. Along with the survey study, Akpan and Brooks (2005a) also

dimensions.

conclude from an experiment that it is easier and more efficient to uncover inaccuracies in a 2½D-VIS model than in a 2D-VIS model.

From the sentiments and empirical evidence above, it can be argued that the traits which are responsible for a 3D-VIS model's ability to communicate, 'explain' the model and 'highlight' model inaccuracies better might also improve its ability to assist an expert in identifying the occasions in a knowledge elicitation session, during which the expert needs to intervene and interact with the model. Extending this argument, it may be suggested that these traits are strongest in a 3D-VIS model, followed by a 2½D-VIS model and then a 2D-VIS model. As such, *ceteris paribus*, an expert would identify the greatest number of occasions to intervene when shown a 3D-VIS model, followed by a 2½D-VIS model and then a 2D-VIS model. This argument leads to the second proposition:

Proposition 2 – *A higher dimension of iconic representation would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Furthermore, Akpan (2005), and Akpan and Brooks (2005b) find out from their survey study that there is strong support to the claim that a 3D-VIS model increases a user's confidence in the simulation results than a 2D-VIS model does. This has the implication that a 3D-VIS model lends itself to greater model credibility and acceptance than an equivalent 2D-VIS model does. However, the authors also find out that the respondents could not agree if a 3D-VIS model is a better decision aid than a 2D-VIS model. Last but not least, the authors find out that a majority of respondents agreed it is more difficult and takes more time to build a 3D-VIS model than a 2D-VIS model.

A non-significant but nonetheless noteworthy feedback in their survey study also mentions that a 3D-VIS model's run speed is slower than a 2D-VIS model. In fact, the 3D-VIS model's run speed is found to vary inversely with the level of photo-realism and resolution of its graphics (Rehn *et al.*, 2004). This immediately calls the integrity of Proposition 2 into question, if it turns out to be true. In this case, one might challenge that even if a larger quantity of example cases is shown to be collected from using a 3D-VIS model (than a 2½D-VIS or 2D-VIS model) in a knowledge elicitation session, the phenomenon might be due to its slower run speed instead of its higher fidelity. This doubt is predicated on the reasoning that a slower-running model would allow an expert more time to process the information from the simulation run and identify the occasions that require his intervention and interaction. However, the same cannot be said of 2½D-VIS versus 2D-VIS model, as their run speeds are not differentiable. Recognising that it is not possible to determine that one dimension of visual representation is generally more efficient than any other dimensions, this leads to the third proposition:

Proposition 3 – *Different dimension of iconic representation would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

4.3.2 FACTOR TWO: MODEL PARAMETERS

Simulation experimentation is one of the main phases in any simulation study (Robinson, 1994). To reiterate, the main objective of computer simulation is to facilitate experimentation for the purpose of better understanding and/or improving the real-world system that is mimicked in the model (Robinson, 2004; Banks *et al.*, 2005).

During an experiment, alternative *scenarios* of the real-world system are often tested by running its VIS model with different combinations of values from various model parameters (Robinson, 2004; Pidd, 2005). The purpose is to search the solution space, which is made up of all possible combinations of various model parameter values, for a scenario that meets the objective(s) of the simulation undertaking. However, as the authors have also pointed out, not all experiments are carried out with a purpose to look for an optimal solution to some objectives. On some occasions, experiments are conducted to develop a better understanding of the real-world system.

Bearing the second purpose in mind, the scenes developed in an experiment can range from mimicking the real-world system under normal operating conditions to very extreme conditions, contingent on the choice of model parameter values. It should be emphasised that a scene of extreme operating conditions does not necessarily represent a trying state of affairs. Instead, it simply means that the scene has a low chance of occurring in the real world.

Using the same principle applied in the second purpose for simulation experimentation, it is imagined that if the model parameters are adjusted such that the VIS model would develop more uncommon and extreme scenes, then it is very likely that the range of situations from which the set of example cases is collected in a knowledge elicitation session would be larger. This thread of thought leads to the fourth proposition:

Proposition 4 – *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the size of state space occupied by the example cases collected in a knowledge elicitation session.*

Further to this, with more uncommon and extreme scenes being developed in a knowledge elicitation session, the reservation that Robinson *et al*. (2005) and Alifantis (2006) have on inundating and consequently boring the expert with vaguely similar scenes is cleared. It can be argued that these scenes would offer the expert more interesting situations that would retain his attention and induce his intervention. As such, *ceteris paribus*, the expert is expected to identify more occasions to intervene and interact with the VIS model. Following this argument leads to the fifth proposition:

Proposition 5 – *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Finally, the situations described in these uncommon and extreme scenes would presumably require more unconventional and perhaps difficult decision-making. This implies that an expert might need more time to consider the attributes that surrounded and provoked each intervention, before making a decision and interacting with the VIS model. In this way, the knowledge elicitation session would expect to take more time. However, the effect from the additional time required for decision-making on the overall elicitation efficiency might be under or over-compensated by the increase in the quantity of example cases collected, if Proposition 5 turns out to be true. In other words, the overall effect on elicitation efficiency is not clear. Hence, this reasoning leads to the last proposition:

Proposition 6 – *Different sets of model parameters would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

4.4 HYPOTHESES FRAMED FOR INVESTIGATING PROPOSITIONS

To recapitulate, six propositions have been suggested for investigation in the last section. They are based on two factors (visual representation dimension and model parameters) and four constructs (decision fidelity, state space, case quantity and collection rate). The propositions are as reproduced below and used to frame six sets of hypotheses, which are expressed in terms of these research factors and constructs. These hypotheses will be tested using a case study in this thesis.

4.4.1 HYPOTHESES RELATED TO VISUAL REPRESENTATION DIMENSION (FACTOR ONE)

Three propositions have been suggested for investigating the effect of the visual representation dimension on decision fidelity, case quantity and collection rate. They are used to frame the first three hypotheses for testing in this thesis.

Hypothesis One

Proposition 1 is put forward as below:

> *A higher dimension of iconic representation would demonstrably improve the degree of decision fidelity in the example cases collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{1(0)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{1(a)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session improves as a higher dimension of visual representation is used.

Hypothesis Two

Proposition 2 is put forward as below:

> *A higher dimension of iconic representation would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{2(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{2(a)}$: The size of case quantity of a knowledge elicitation session increases as a higher dimension of visual representation is used.

Hypothesis Three

Proposition 3 is put forward as below:

> *Different dimension of iconic representation would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{3(0)}$: The collection rate in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{3(a)}$: The collection rate in a knowledge elicitation session is affected by the visual representation dimension used.

4.4.2 HYPOTHESES RELATED TO MODEL PARAMETERS (FACTOR TWO)

Similarly, three propositions have been suggested for investigating the effect of model parameters on state space, case quantity and collection rate. They are also used to frame the next three hypotheses for testing in this thesis.

Hypothesis Four

Proposition 4 is put forward as below:

> *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the size of state space occupied by the example cases collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{4(0)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session is not affected by the model parameters used;

$H_{4(a)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

Hypothesis Five

Proposition 5 is put forward as below:

> *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{5(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the model parameters used;

$H_{5(a)}$: The size of case quantity of a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

Hypothesis Six

Proposition 6 is put forward as below:

> *Different sets of model parameters would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

Therefore, an overarching set of null and alternative hypotheses that corresponds to the proposition above can be stated as follows:

$H_{6(0)}$: The collection rate in a knowledge elicitation session is not affected by the model parameters used;

$H_{6(a)}$: The collection rate in a knowledge elicitation session is affected by the model parameters used.

4.5 METHODOLOGY USED FOR TESTING HYPOTHESES

At the outset, it is helpful to determine the research approach to be used in the investigation. Following this, a methodology for testing the hypothesis is then worked out. It should be mentioned at this point that a case study based on a real-world system with experts making real decisions will be used to support the investigation.

In a nutshell, the propositions are trying to establish the probable causal links between the factors and the constructs. Ideally, this would entail an investigation where cause and effect are isolated and studied. In this respect, an experimental research approach is chosen to lead the investigation. Field and Hole (2006) explain an experimental research as a mode of research where experiments are designed and executed. In these experiments, the causal variables (*i.e.* the factors, in this case) will be manipulated to study their effects on some variables of interest (*i.e.* the constructs, in this case), whilst all other confounding variables are controlled.

Keeping the research approach to be used in mind, a methodology is then developed. Adopting Weitzel and Kerschberg's (1989a and b) philosophy to use flexible 'processes' instead of sequential 'phases' to describe the methodology, the investigation will be applying one with eight processes. They are:

i. Understanding the case study;

ii. Designing the experiment;

iii. Building the VIS model;

iv. Assessing the VIS model;

v. Devising the measures for evaluating the four constructs;

vi. Collecting data;

vii. Analysing data; and finally

viii. Discussing the results.

These processes are flexible because they can be activated, deactivated and reactivated as necessary (Section 2.3.2). It is stressed that these processes are not entirely sequential. Some processes are independent of the others and can be activated in parallel with them. Moreover, some working-in-progress processes may be deactivated (albeit not always necessary), if additional information is required from their preceding processes. Consequently, the preceding processes will be reactivated and so forth, thus forming an iteration of activities. A framework depicting the relationships between the processes is provided in Figure 4.1.

The methodology begins with gaining an in-depth appreciation of the case study (process i). This includes finding out about the physical and logical design of the real-world system, on which the case study is based upon. In addition, the decisional roles assumed by the experts working in the system are also determined. Further to this, the types of decisions that the experts make regularly in the course of their work, and the information used by them to make these decisions are identified. Last but not least, the experts' commitment to participate in the investigation is also secured.

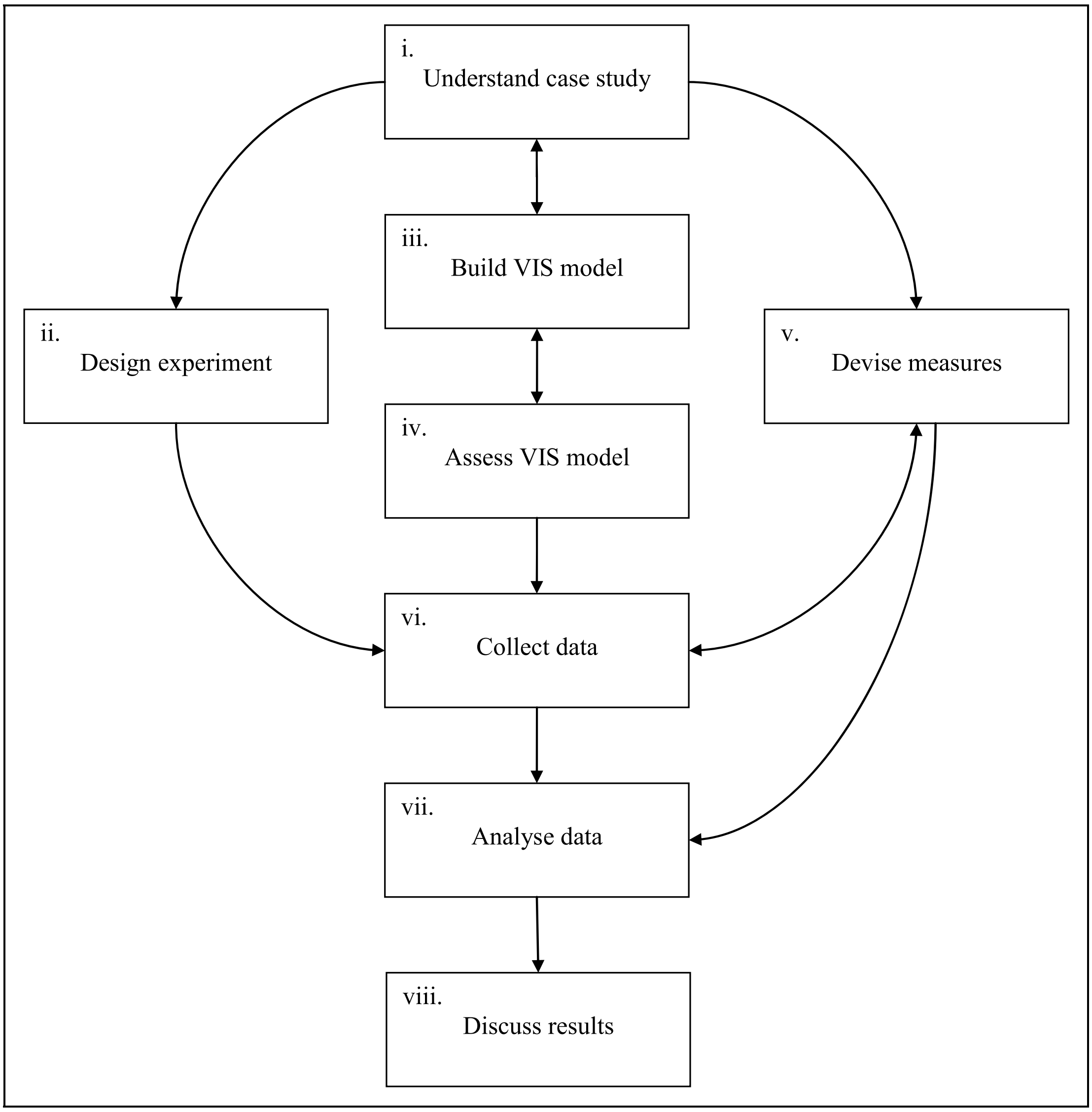


Figure 4.1: A framework depicting the research methodology

With the above information in hand, three independent processes can be activated concurrently. These processes are essentially equivalent to the when, how and what of data collection. Firstly, using the hypotheses set out above and information pertaining to the experts' commitment, an experiment is designed and a tentative schedule for the knowledge elicitation sessions is drawn up in advance (process ii). Secondly, a VIS model of the real-world system is built and adapted to record the experts' decision-making episodes during the knowledge elicitation sessions (process iii). If more information is needed to complete the model, then the previous process of information gathering (*i.e.* process i) is reactivated to fill the gap. This process is only deactivated after the missing information is gathered. In the meantime, the modelling work may stop, if it cannot continue without the missing information, and recommence when the missing information is gathered. Once the model is deemed to be complete, the next process of assessing it with a few experts is activated (process iv). During these assessments, all appropriate observations made by the modeller and feedback furnished by the experts will be noted and used to fine-tune the model, before subjecting it to another session of assessments. In effect, this series of activation, deactivation and reactivation between the three processes of understanding the case study, building and assessing the VIS model serve to refine the latter iteratively. Thirdly, measures are devised to evaluate the four constructs that are used for assessing elicitation effectiveness and efficiency (process v). These measures are conceptualised initially as if all data needs can be met by the real-world system and the knowledge elicitation sessions.

Having designed the experiment, built and assessed the VIS model, and devised the necessary measures, the process of data collection is activated (process vi). Here, two types of data are being collected: historical data from the real-world system and example cases from the experts' interactions with the VIS model during the knowledge elicitation sessions. If it is realised that certain historical data that is required for computing the measures cannot be obtained, or that the experts cannot afford the time to participate in all the knowledge elicitation sessions that are originally planned, then the process of devising measures (*i.e.* process v) will be reactivated to circumvent this problem. Eventually, alternative measures based on data that can be collected are devised to replace those unviable measures. Although not as extensive as above, this series of activation, deactivation and reactivation between the two processes of collecting data and devising measures also serve to refine the measures iteratively. Subsequently, these alternative measures will lead to a new process of data collection being activated, on top of the original process that is activated earlier.

Following the completion of the data collection process(es), the historical data and example cases collected are used to compute the measures for testing the hypotheses framed in Section 4.4 (process vii). Finally, the results from the analysis are discussed in the last process (process viii).

4.6 CONCLUSION

Six propositions have been suggested for investigation in this chapter. They are based on two experimental factors (visual representation dimension and model parameters) and four constructs (decision fidelity, state space, case quantity and collection rate). On the one hand, the factors are identified for their potential to improve VIS' ability in eliciting expert knowledge. On the other hand, the constructs are conceived to assess elicitation effectiveness and efficiency. In essence, the propositions call for an investigation into the postulated causal links between the factors (causes) and constructs (effects) as organised in Table 4.1.

Cause	Effect
Visual representation dimension (2D, 2½D and 3D)	· Decision fidelity · Case quantity · Collection rate
Model parameters (Unadjusted and Adjusted)	· State space · Case quantity · Collection rate

Table 4.1: Postulated cause and effect relationships

These propositions are then used to frame the hypotheses, which will be tested with the help of a case study. In addition, the methodology employed to test the hypotheses is also outlined. Adopting Weitzel and Kerschberg's (1989a and b) philosophy to use flexible 'processes' to describe the methodology, it comprises of eight processes that can be activated, deactivated and reactivated as necessary. The details and outcomes of all work carried out in each process are provided in the subsequent chapters. They are: understanding the case study (process i – Chapter 5), designing the experiment (process ii – Chapter 6), building and assessing the VIS model (processes iii and iv – Chapter 7), devising the measures for evaluating the four constructs (process v – Chapter 8), collecting data and analysing the data (processes vi and vii – Chapter 9 and 10), and finally discussing the results (process viii – Chapter 11). Together, the work performed in these processes contributes to answering the second research question (Section 1.1).

5 A Case Study: Ford Puma Diesel Engine Hot-test Operations

The hypotheses in this thesis will be tested using a real-world case study set in a Ford Motor Company (Ford) engine assembly plant located in Dagenham, East London. The case study looks at the hot-test operations of the Puma diesel engine (a named line of engines) assembly line, wherein a team of dedicated experts monitor and manage the stream of engines passing through it. This chapter relates to the first process of the methodology outlined in the last chapter, which is to gain an in-depth understanding of the case study.

Whilst it is known at the beginning that having a clear picture of the case study is an essential step in the investigation, it is also realised later that gaining an appreciation of the decisional role played by the experts is pivotal in prescribing the way that the VIS model should go about eliciting expert knowledge. Further to this, the decisions that the experts make regularly in the course of their work, and the information used by them to make these decisions are also identified. Last but not least, the experts' commitment to participate in the investigation is secured. The work carried out in this process and its findings are described in detail below.

5.1 DISCOVERING THE HOT-TEST OPERATIONS

The purpose of this first process is to lay the groundwork for the investigation. To begin, there is a need to gain a high-level understanding of the entire Puma diesel engine assembly line and the circumstances in which an expert is required to make decisions. Also, the decision and attribute variables that compose the decision-making process are identified. In effect, this process is akin to the 'broad and shallow' phase suggested by Barrett and Edwards (1995), where the priority is to extend the breadth of background information as wide as feasible. As no technique alone is sufficient to elicit all kinds of information (Rugg *et al.*, 2000; Rugg *et al.*, 2002; Coffey and Hoffman, 2003), any techniques that complement each other in gaining such a broad overview will be used. The techniques used in this process are document analysis, unstructured and semi-structured interview, and observation interview.

First and foremost, paper documents like plant layout, versatility charts and log sheets are reviewed to gain a quick introduction to the hot-test operations. The plant layout is used to locate the hot-test operations in the context of the entire engine assembly line. It shows the activities before an engine enters and after the engine leaves the hot-test operations. In addition, it also provides a record of the physical entities that make up the engine assembly line, including the hot-test operations. Versatility charts are used to provide information on manpower status of the hot-test operations. They show the responsibilities that each hot-test person in every shift is qualified to assume. This information is especially useful when the experiment is being designed in the next process (Chapter 6). Last but not least, log sheets are used to provide clues on the types of information that might be used to monitor the hot-test operations. An instance of a useful log sheet is one that records the number of engines that had been tested in each hot-test cell on an hourly basis.

Next, informal interviews are conducted with the experts to collect any other pertinent undocumented information, and then to clarify any queries that are formed after reviewing the documents and preliminary interview materials. Moreover, conducting these interviews in an informal setting also helps to create opportunities for establishing good rapport with the experts, whose co-operation in subsequent knowledge elicitation sessions might be crucial. Initially, unstructured interviews are used where the experts are given the freedom to cover topics that they deem fit. It is because at this early stage, a person who is not familiar with the hot-test operations will not have enough background information to ask specific questions (as in structured interviews), or even work to cover a list of topics (as in semi-structured interviews) in an interview session. After a few sessions of unstructured interviews, sufficient information should have been collated from both documents and interview materials to form a clearer picture of the hot-test operations. At this point, a few queries may have surfaced, and these are answered via another few rounds of semi-structured interviews.

Finally, observation interviews are conducted as a ‘catch-all’ attempt to collect additional information that is neither documented nor conversed in earlier efforts. Here, the experts’ activities in the hot-test operations are observed and recorded. If there are any queries in respect of the observations made, they will be clarified with the experts at the first instance. These queries will range from the reasons behind the observed activities to the consequences as a result of them. In this way, a rough idea of the experts’ decision-making strategies is conceived. In addition, questions regarding the physical and logical layout of the hot-test operations may also be asked. The latter information is especially important in the next process of building a VIS model of the hot-test operations. The aim of this ‘observe-query-observe’ activity sequence is to verify any assumptions that are made during the document analysis and interview efforts, and to reinforce one’s understanding.

5.2 FINDING ONE: PHYSICAL AND LOGICAL LAYOUT OF THE HOT-TEST OPERATIONS

As mentioned above, the case study is set in a Ford engine assembly plant in Dagenham, East London. It is responsible for assembling an assortment of engines, a group of which is labelled as the Puma diesel engines. The operations to assemble Puma diesel engines are carried out along two main assembly lines known simply as Assembly Line A and B. The assembly operations start on Assembly Line A and continue onto Assembly Line B. At the end of Assembly Line B, the assembled engines are loaded onto the hot-test operations, where they are filled with fuel and run to test their build quality. If the engines meet the standards required to pass the hot-test, they are then sent to the Ultra-Violet (UV) booth to inspect for leakages before being dispatched to the After Test Dress (ATD) operations for some final touch-up and shipping out.

The case study used in this thesis is based on the hot-test operations that follow the assembly operations in Assembly Line B. A schema of the hot-test operations is shown in Figure 5.1. The hot-test operations are made up of 20 pairs of hot-test cells and waiting stands (represented by the blue boxes), one path control panel (pink box), ten cell/stand control panels (yellow boxes) and a set of conveyors (black lines with arrows to denote directions of conveyor movement).

Broadly, assembled engines will arrive from Assembly Line B at the top of the schema, and enter the hot-test operations after being loaded onto the hot-test platens (metal pallets). A picture of a (loaded) platen is displayed in Figure 5.2. Upon reaching Junction J, these untested engines will be sent either along the path on the right (via Conveyor B, D, F and E) or straight down (via Conveyor C and E). The actual path taken by them is controlled by the qualified hot-test personnel.

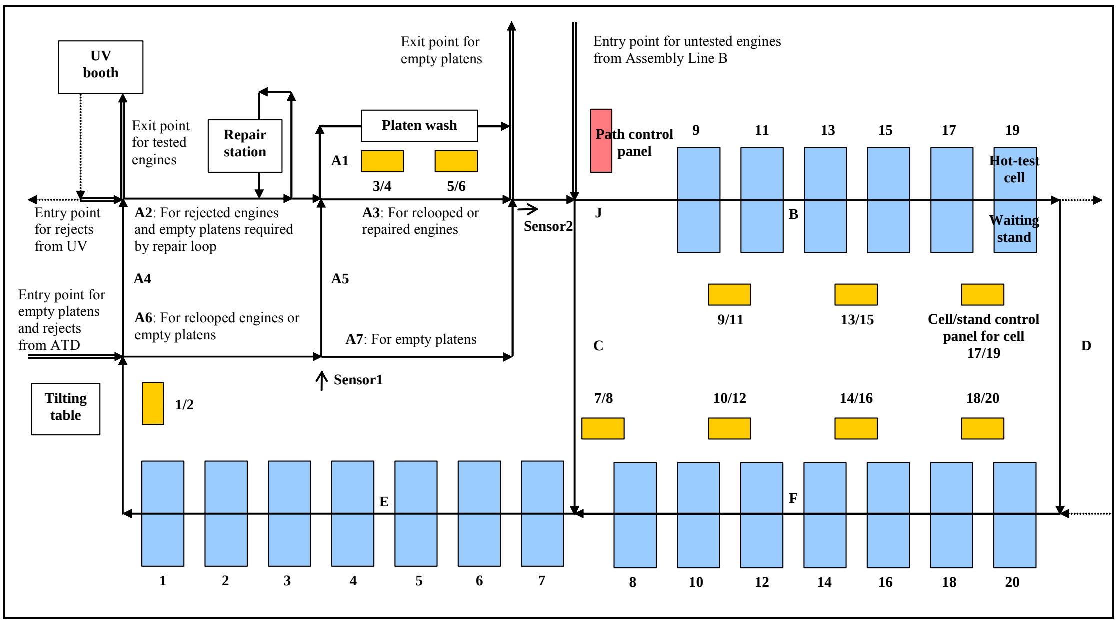

UV booth
Exit point for empty platens
Entry point for untested engines from Assembly Line B
Exit point for tested engines
Repair station
Platen wash
A1
3/4
5/6
Path control panel
9
11
13
15
17
19
Hot-test cell
Waiting stand
Entry point for rejects from UV
A2: For rejected engines and empty platens required by repair loop
A3: For relooped or repaired engines
Sensor2
J
B
A4
A5
Entry point for empty platens and rejects from ATD
A6: For relooped engines or empty platens
A7: For empty platens
9/11
13/15
Cell/stand control panel for cell 17/19
C
D
Sensor1
Tilting table
1/2
7/8
10/12
14/16
18/20
E
F
1
2
3
4
5
6
7
8
10
12
14
16
18
20

Figure 5.1: A schema of the hot-test operations

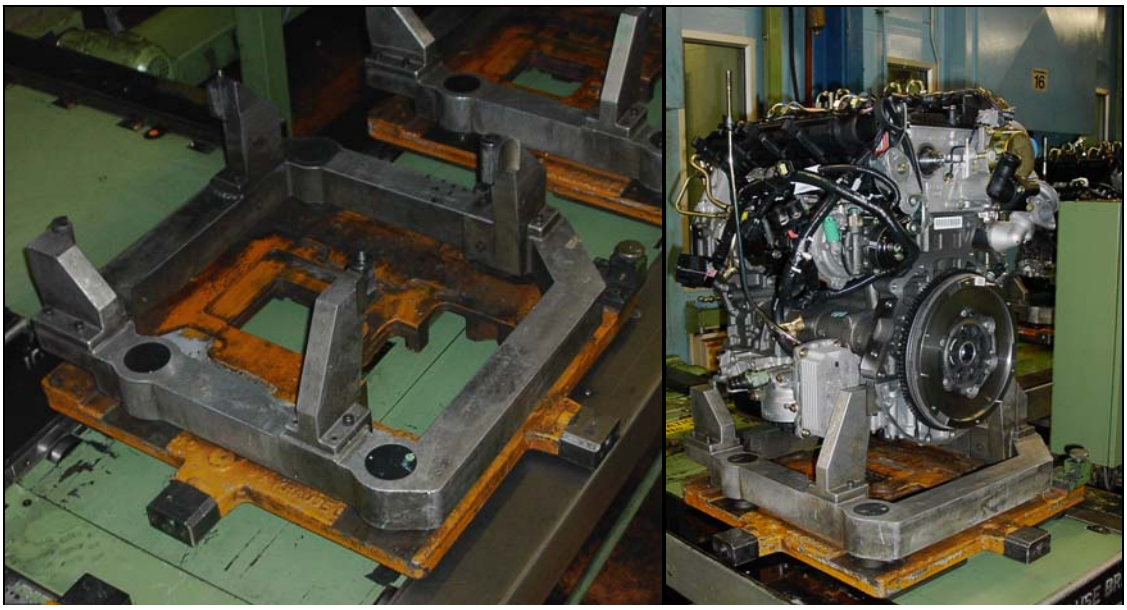


Figure 5.2: A platen (left) and a platen loaded with an engine (right)

Regardless of the path taken, an untested engine will enter the nearest vacant hot-test cell or stand, provided the vacant cell or stand is not switched off. When an occupied hot-test cell has completed hot-testing the engine inside it, the tested engine will exit the cell and an untested engine from the adjacent stand will enter the cell for testing. A snapshot that shows a tested engine preparing to exit a hot-test cell, whilst an untested engine waits to enter from an adjacent waiting stand is displayed in Figure 5.3. If there is no untested engine on the adjacent stand, then the vacant cell will wait for the next untested engine that comes along the conveyor. As an illustration, the order of cell/stand by which an untested engine at Junction J that is sent along the path on the right will try to enter is Cell 9, Stand 9, Cell 11, Stand 11 and so on. The eventual point of entry is the first vacant and switched-on cell/stand in this order that the untested engine comes across as it travels along the conveyor. Albeit this defies commonsense, an untested engine will enter Stand 9 even if Cell 11 is vacant. Similarly, the order of cell/stand considered by an untested engine sent straight down from Junction J is Cell 7, Stand 7, Cell 6, Stand 6 and so on. In this way, the untested engines are assigned to one of the 20 hot-test cells, where they are rigged to a testing machine and run for a few minutes. Following the hot-test, defective engines are sent to a repair station for rectification (via Conveyor A4 and A2), whilst engines that passed the hot-test are sent to the ATD operations after a final inspection at the UV booth. Defective engines that have been repaired are sent along Conveyor A3 to Junction J for re-assigning and re-testing.

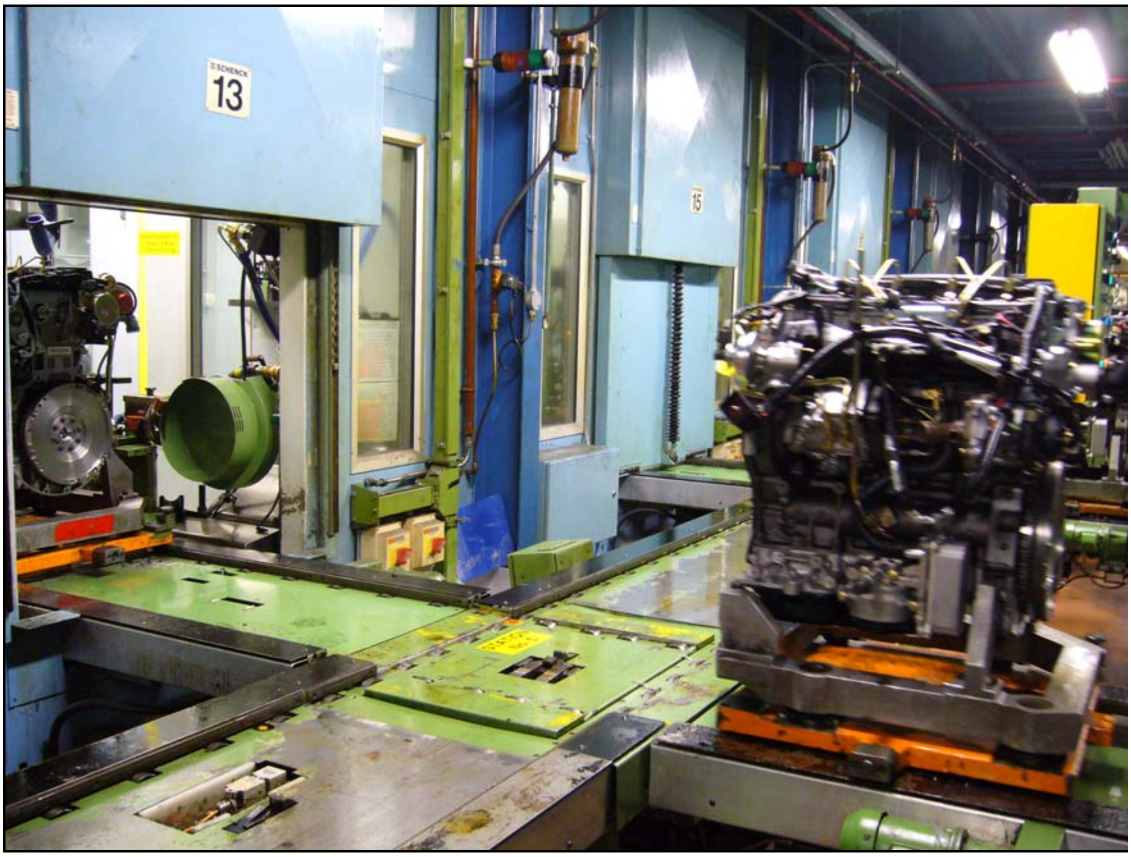


Figure 5.3: A tested engine prepares to exit cell 13, whilst an adjacent untested engine waits to enter it

The key decision maker or expert here is known as the switch operator, who is responsible for assigning untested engines to vacant hot-test cells. His main objective is to maximise the number of engines that are tested and sent to ATD by the end of his shift, through maintaining an efficient and smooth workflow with no bottlenecks. In the meantime, he also aims to distribute the workload equitably among all hot-test cell operators, which entails manual rigging and de-rigging of engines onto/from the hot-test machines. To aid him, the expert has access to a set of 21 basic switches comprising of a path control switch (on the path control panel), and 20 cell/stand control switches (on the cell/stand control panels). The path control switch allows the expert to send incoming untested engines from Assembly Line B either to the path on the right of Junction J or straight down. Alternatively, the expert may also use the same switch to opt for the automated cyclical mode of sending three engines to the right of Junction J followed by one straight down. On the other hand, a cell/stand control switch allows the expert to either switch on or off a hot-test cell and its adjacent stand. As only one cell/stand control switch is used for switching on or off both a hot-test cell and its adjacent stand concurrently, there are a total of 20 cell/stand control switches for the 20 pairs of cells and stands.

5.3 FINDING TWO: DECISIONAL ROLES OF THE HOT-TEST SWITCH OPERATORS (THE EXPERTS)

Mintzberg (1973) breaks down management work into ten roles: monitor, disseminator, spokesperson, figurehead, leader, liaison, entrepreneur, disturbance handler, resource allocator and negotiator. In addition, he realised that these roles can be arranged into three groups: informational, interpersonal and decisional. The activities involved in these roles are as described by Boddy (2005) in Table 5.1.

Category	Role	Activity
Informational	· Monitor	· Seek and receive information, scan papers and reports, and maintain interpersonal contacts
	· Disseminator	· Forward information to others, send memos, and make phone calls
	· Spokesperson	· Represent the unit to outsiders in speeches and reports
Interpersonal	· Figurehead	· Perform ceremonial and symbolic duties, and receive visitors
	· Leader	· Direct and motivate subordinates, train, advise and influence others
	· Liaison	· Maintain information links in and beyond the organisation
Decisional	· Entrepreneur	· Initiate new projects, spot opportunities, and identify areas of business development
	· Disturbance handler	· Take corrective action during crises, resolve conflicts amongst staff, and adapt to external changes
	· Resource allocator	· Decides who get resources, schedule, budget, and set priorities
	· Negotiator	· Represent department during negotiations with unions, suppliers, and generally defend interests

Table 5.1: Mintzberg’s ten management roles (Boddy, 2005)

Under Mintzberg’s (1973) classification, four management roles are recognised to assume a decisional nature. They are those of an entrepreneur, disturbance handler, resource allocator and negotiator. Boddy (2005) describes an entrepreneurial role as one where the managers initiate change within the organisation. They see opportunities or problems and create projects to deal with them. Managers play this role when they introduce a new product or create a major change programme. Managers play the disturbance-handler role when they deal with problems and changes that arise unexpectedly during daily routine. This includes taking corrective actions during operational crises and resolving conflicts among subordinate staff. In the resource allocator role, managers have to choose among competing demands for money, equipment, personnel and perhaps their time. Lastly, managers play the negotiator role when they have to reach agreement with other parties on whom they depend.

The switch operators in the hot-test operations, who are the experts in the case study, play two roles. They are mainly resource allocators, and occasionally disturbance handlers. From the description of their responsibility in the last section, it is quite evident that they are primarily resource allocators. However, they also assume the role of disturbance handlers every now and then. For instance, as the hot-test operations has been around for at least 15 years at the time of this investigation, the conveyor-and-platen system is not as free of problems as it used to be. It is not uncommon for platens to be misaligned and stuck at conveyor intersections several times during a shift. In this case, an expert would expect to stop his regular tasks and attend to the stuck platens as soon as possible. In another instance, there might be a stream of untested engines that has just entered the hot-test operations, which need to be shipped out of the engine assembly plant urgently. In this case, these untested engines will be given the highest priority in the hot-test operations. Hence, the expert would expect to put aside his usual *modus operandi* for assigning engines to hot-test cells, and focus on testing these engines at the earliest opportunity available.

5.4 FINDING THREE: MAKE-UP OF THE EXPERTS' DECISION-MAKING PROCESS

In general, there are two groups of diesel engines of different capacity that are being assembled in the Puma engine assembly line at the time of this investigation. The capacities are 2 litres (2*l*) and 2.4 litres (2.4*l*). As mentioned earlier, when newly assembled and untested engines first enter the hot-test operations at Junction J in Figure 5.1, an expert can decide to send them either to the path on the right of Junction J (via Conveyor B, D, F and E) or straight down (via Conveyor C and E). Alternatively, the expert may also use the same switch to opt for the automated cyclical mode of sending three engines to the right of Junction J followed by one straight down. At this moment, the expert may use information on the quantity and type of engine on the conveyor, the type (2*l* versus 2.4*l*) of engine currently being or last tested in each hot-test cell and operational status of each hot-test cell to aid his decision.

After deciding on the path that the engines will be dispatched, the expert has to decide where to assign each engine. At this stage, the expert will execute his hot-test cell allocation plan by switching on/off various combinations of hot-test cells and their adjacent stands. There are infinite different situations that may motivate the expert to carry out these series of switches. At this moment, the expert may use the same set of information as above, as well as those on the type of engine parked on each waiting stand. Finally, as the expert also aims to distribute the workload equitably among all hot-test cell operators, he may use information on the quantity of engines handled by each hot-test cell (and hence its operator) to bring his attention to any unintended allocation bias.

The decisions that these experts make regularly in the course of their work, and the attributes that influence their decisions are summarised in Table 5.2. The rationale behind why these decisions are made and attributes are used are illustrated further with two different scenes below, with frequent references made to Figure 5.1.

Decision	Attribute
· Set path option at junction J to straight, left or automatic · Switch hot-test cell/stand on or off	· Quantity of engines on each section of conveyor · Type of engine on each section of conveyor · Type of engine currently being/last tested in each hot-test cell · Type of engine parked on each waiting stand · Operational status of each hot-test cell · Quantity of engines tested by each hot-test cell operator

Table 5.2: A summary of decision and attribute variables used in engine assignment

5.4.1 SWITCHING OPERATIONS IN A STANDARD SCENE

A standard scene in the hot-test operations is typified by an irregular flow of untested engines entering the hot-test operations. A common sight from a standard scene is displayed in Figure 5.4. This scene may be due to upstream problems such as a machine breakdown in Assembly Line B. It may also be due to downstream problems such as a bottleneck in the ATD operations that results in a shortage of empty platens (unloaded of tested engines), which should otherwise be released for loading untested engines from Assembly Line B into the hot-test operations. As a result, there are more vacant hot-test cells that are idling than there are untested engines to fill them. Bearing in mind that an expert's main concern is to maximise the number of engines tested by the end of his shift, he will therefore try to keep as many operational hot-test cells occupied as possible. In this case, the expert will try to get an untested engine to a vacant hot-test cell using the quickest means. There are mainly two ways that the expert may achieve this.

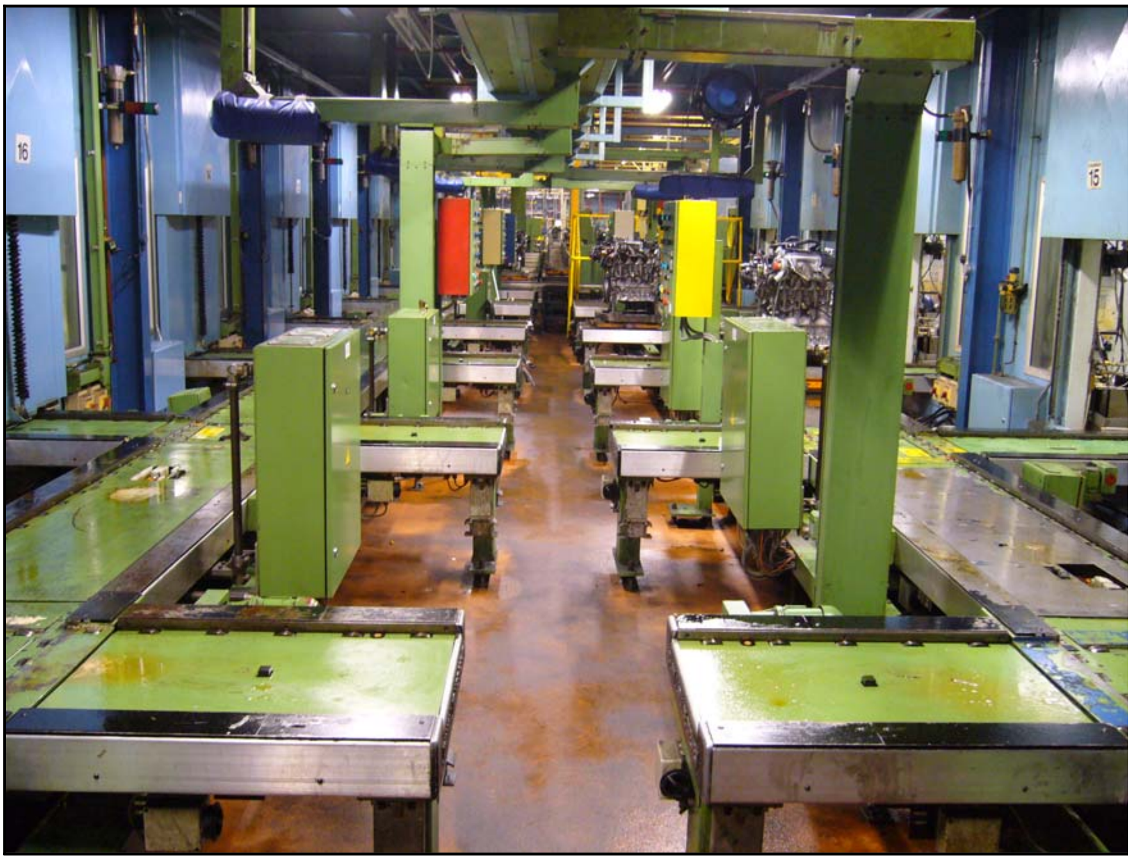


Figure 5.4: A snapshot of a standard scene

Firstly, the expert may decide to send incoming engines at Junction J towards vacant hot-test cells using the shortest route. This will depend on the operational status of each hot-test cell, and the quantity of engines that is currently on the conveyor. For instance, if the operational hot-test cells along Conveyor E are vacant whilst those along Conveyor B and F are fully occupied, and there are no untested engines travelling along Conveyor E, then the expert might decide to send incoming engines at Junction J straight down to Conveyor E via Conveyor C (instead of the longer route on the right via Conveyor B, D and F).

Secondly, the expert may decide to switch off occupied hot-test cells, which precede those that are operational but vacant. This will effectively disengage the waiting stands adjacent to the occupied hot-test cells without affecting the latter's current operations. Such a move serves two purposes: (1) to eject the untested engines parked on these waiting stands so that they can be transferred to succeeding vacant hot-test cells; and (2) to prevent any oncoming untested engines from parking on these waiting stands so that they will bypass to succeeding vacant hot-test cells. Again, this will depend on the operational status of each hot-test cell, and the quantity of engines that is currently on the conveyor.

Extending the example above, assume a case where there are untested engines that are parked on the waiting stands along Conveyor B and F, and incoming engines from Assembly Line B are sparse. Thus, the expert might find it quicker to transfer these untested engines to the vacant hot-test cells along Conveyor E by switching off all the waiting stands along Conveyor B and F.

5.4.2 SWITCHING OPERATIONS IN A NON-STANDARD SCENE

A non-standard scene in the hot-test operations is typified by a regular flow of untested engines entering the hot-test operations. More often than not, this means that the hot-test operations will be swarmed with both tested and untested engines. A common sight from a non-standard scene is displayed in Figure 5.5. Unlike in the previous scene, an expert here will not be worried with keeping operational hot-test cells occupied. It is because there are more untested engines than there are hot-test cells available to test them. Instead, the expert will try to maintain the engine type that is being handled by each operational hot-test cell. For example, the expert would prefer to assign a 2*l* engine to a hot-test cell that is currently testing or has just tested a 2*l* engine. This is to minimise the amount of unproductive changeover time that is lost when a hot-test cell changes from testing one type of engine to another. As such, there will be more time to test the engines, thereby addressing the expert's main concern to maximise the number of engines tested by the end of his shift. Moreover, this also serves to keep the hot-test operators happy as fewer changeovers imply less work. Similarly, there are mainly two ways that the expert may achieve this.

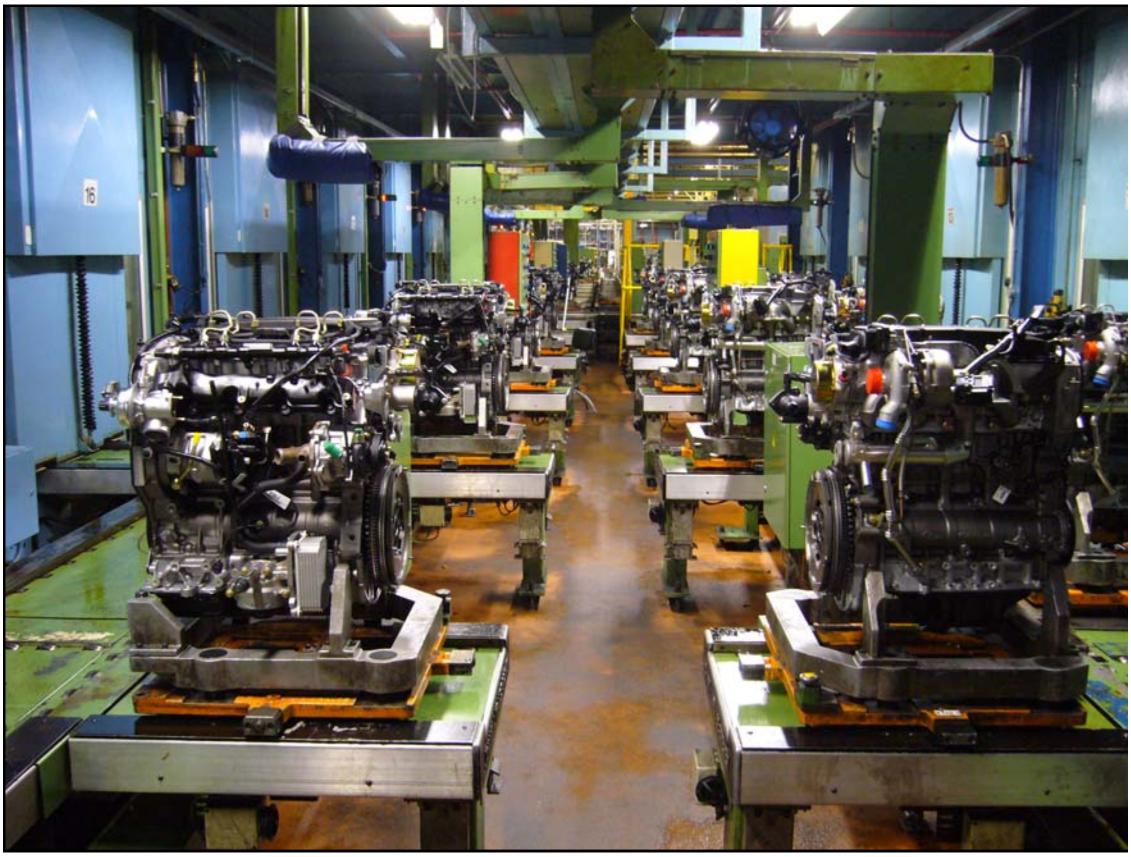


Figure 5.5: A snapshot of a non-standard scene

Firstly, the expert may decide to send an incoming engine at Junction J onwards a path that he thinks is appropriate. This will depend on where the 2*l*/2.4*l* engines are being tested in the hot-test operations, the operational status of each hot-test cell, the quantity and type of engine that are currently on the conveyor, and the type of engine that is arriving at Junction J. For example, if the expert has arranged for 2*l* engines to be tested along Conveyor B only, and 2.4*l* engines to be tested along Conveyor F and E, then contingent on the operational status of hot-test cells and quantity of untested 2.4*l* engines along Conveyor F and E, the next 2.4*l* engine that turns up at Junction J would be sent either to the path on the right (towards Conveyor F), or straight down (towards Conveyor E).

Secondly, the expert may decide to switch on/off an appropriate combination of hot-test cells/stands. Hence, in addition to the information that is taken into account above, the expert will also consider the type of engine that is currently parked on each waiting stand. Extending the example above, assume a case where there is a 2*l* engine that is parked on a waiting stand and is adjacent to a hot-test cell that is currently testing a 2.4*l* engine. Thus, if there is a 2.4*l* engine travelling on the conveyor section that precedes this waiting stand, then the expert might switch it off (without affecting the current operations in the adjacent hot-test cell) to eject the 2*l* engine and switch it on again to receive the oncoming 2.4*l* engine. The untested 2*l* engine that is ejected will then travel along the conveyor and enter the next hot-test cell or waiting stand that is available.

5.5 PROFILES OF THE PARTICIPATING EXPERTS

There was a total of ten hot-test personnel who are qualified to assign engines to hot-test cells for testing. Unfortunately, only eight of them were able to commit themselves to the investigation. Furthermore, their participation was subject to the management's consent on a daily basis. In order to maintain their anonymity, the experts are identified as Subject A, B, C and so on up to H. Their profiles are as summarised in Table 5.3.

These are compiled by administering pre-experiment questionnaires, a copy of which is available in Appendix A.

Subject	Age (years)	Experience in switch operations (years)	Experience (yes or no)		Visual score (%)
			With computers	With game consoles	
A	40-49	6	Yes	No	80
B	40-49	N/A	Yes	Yes	80
C	50-59	8	No	Yes	77
D	50-59	3	No	Yes	72
E	50-59	6	No	No	85
F	40-49	8	Yes	Yes	83
G	40-49	<1	Yes	Yes	72
H	50-59	6	No	No	72

Table 5.3: A summary of the experts' profiles

All eight experts range from 40 to 59 years in age, with half of them (Subject A, B, F and G) in their 40s and the rest (Subject C, D, E and H) in their 50s. Moreover, with the exception of Subject B, all other experts are reasonably experienced in engine assignment and are therefore eligible to help with the investigation. Nonetheless, as Subject B is a team leader whose responsibility is to run the entire hot-test operations, which includes standing in as a switch operator occasionally, he is also deemed fit to help with the investigation.

Further to this, the experts have been asked if they have any prior experience of using a computer or game console. The purpose is to detect the experts' comfort level with using the VIS model. With the exception of Subject E and H, all other experts have claimed to have some form of experience with a computer or game console, and also professed to be competent at executing point-and-click actions using a mouse. On the other hand, Subject E and H have declared to have no experience with a computer or game console, and also expressed their doubts with using a mouse to navigate the cursor on the visual display. In this respect, these experts will be provided with a brief session to orientate them with the mouse in addition to the VIS model to ease any potential awkwardness.

Last but not least, the experts have also been asked to self-assess their learning style by rating themselves on 12 psychological questions adapted from Akpan (2005). Following this, a visual score is computed for each expert to indicate if he is a visual learner. The latter is described as a person who learns better and faster from what he sees than from what he hears. Bearing in mind that visual representation dimension is a factor that is being investigated, it is desirable that the experts are visual learners. It is because a non-visual learner's decision-making might not be affected by a change in visual representation dimension since he does not rely primarily on what he sees to ingest information. Consequently, a bias might be introduced into the investigation unwittingly, and lead to an insignificant finding.

To compute an expert's visual score, his ratings for all the questions are summed and then converted into a percentage. As a guideline, Akpan (2005) mentions that a person with a visual score of 80% and above is considered a major visual learner, whilst a score of between 60% and 80% means that he is a minor visual learner. However, a visual score of less than 60% is regarded to have no significant connotation. The visual scores in Table 5.3 range from 72% to 85%. Although these scores indicate some differences in the eight experts' ability to handle visual information, they are fortunately quite small. More importantly, these scores are all above the 60% benchmark, which show that all of the experts are visual learners. As such, the reservation of the experts biasing the investigation as a result of them not being visual learners is laid to rest[5].

5.6 CONCLUSION

This chapter attempts to gain an understanding of the hot-test operations and its environment by using an array of complementary techniques. These include document analysis, unstructured and semi-structured interview, observation interview, and questionnaire survey. Following this, four main findings are delivered from this series of activities. Firstly, the constituents of the hot-test operations, and how they relate to each other are determined. Secondly, the decisional roles that the experts play in their daily undertakings are established. Thirdly, the decisions that these experts made regularly in their daily undertakings, and the information used by them to make these decisions is also identified. These three findings are essential to design and build a VIS model that is fit-for-purpose. Finally, the eight experts who have committed themselves to participate in the investigation are profiled. In particular, a visual score is computed for each expert to assess their suitability to help with the investigation.

In the next chapter, a quasi-experiment for testing the hypotheses is designed and planned. It belongs to the process of designing the experiment (process ii in Figure 4.1) and is the first of three concurrent processes described in the research methodology in Section 4.5, which follow this initial process of understanding the case study.

[5] As long as all experts are visual learners, the actual differences between their visual scores are not important in light of the experimental design to be adopted in Chapter 6.

6 Experimental Design

An experimental research approach is identified in Section 4.5 as the best way to establish cause and effect directly and unambiguously. In a true experimental research, Field and Hole (2006) explain that the causal variables are manipulated to study their effects on some variables of interest, whilst all other confounding variables are controlled. However, the authors also remark that it is not always possible to conduct true experiments in real-world situations, as the order by which the experiment trials are carried out cannot always be randomised completely. As a result, bias might be introduced systematically into the findings. On these occasions, a quasi-experimental design is used instead. Campbell and Stanley (1963) define a quasi-experimental design as one that shares the logic and many features of the experimental method, but does not have as much control over some, if not all, of the variables. As a result, cause and effect are not isolated as conclusively as with a true experimental design, albeit any subsequent findings can still be reasonably reliable and valid (Kowalski and Westen, 2005; Field and Hole, 2006).

Using the information gathered in the last chapter, the repeated measures experimental design is chosen for the research and is discussed below. Following this, the design of the VIS-based means for eliciting knowledge from the experts is also described briefly. Finally, whilst every care is taken and effort is made to ensure that a true and appropriate experimental design is used, randomisation is still compromised in an aspect of the experiment. This means that a quasi-experiment is carried out instead. As a mitigating measure, some actions are taken to diminish the effects from the compromise and they are presented as well.

6.1 REPEATED MEASURES EXPERIMENTAL DESIGN

In the last chapter, it is identified that there are eight experts who are able and committed to help with the entire experiment. However, their actual participation is subject to management consent on a daily basis. In respect of a sample size as small as this, Field and Hole (2006) recommend that the repeated measures experimental design be used if it is feasible to do so. The latter is also commonly known as the within-subjects experimental design.

In a repeated measures design, each expert will be exposed to all conditions of the experiment, where an experimental condition is defined as a unique combination of the factors being investigated. As the same set of participants is used in all conditions, the issue of random differences between participants exposed to one condition and participants exposed to another condition is eliminated. Since there are fewer sources of random variations that can obscure the effects of experimental manipulations, any differences between an expert's scores measured under different conditions are expected to be due mainly to them. Hence, Field and Hole (2006) comment that this design is generally more sensitive than others such as the between-groups design, and will be more likely to detect any differences that exist between conditions.

As mentioned in an earlier chapter, there are two factors that are being explored in this thesis: visual representation dimension, and model parameters (Section 4.3). On the one hand, there are three treatment levels that are considered under visual representation dimension: 2D, 2½D and 3D. On the other hand, there are two treatment levels that are considered under model parameters: unadjusted parameters and adjusted parameters. Hence, there are altogether six unique combinations that can be formed from these two factors. This implies that each expert will be exposed to six different conditions under the repeated measures design. They are:

I. 2D representation with unadjusted parameters;

II. 2D representation with adjusted parameters;

III. 2½D representation with unadjusted parameters;

IV. 2½D representation with adjusted parameters;

V. 3D representation with unadjusted parameters; and

VI. 3D representation with adjusted parameters.

6.2 VISUAL INTERACTIVE SIMULATION IN GAMING MODE

It is known at the beginning of the experiment that a VIS model will be used as the vehicle for eliciting knowledge from the experts. Hence, there is a need to design one for serving this purpose. Preece (1994) observes from early computer applications that well-designed applications always supported the interaction style that matched the user and task requirements relatively well. For instance, form-fill applications such as the Microsoft Office Access are designed at enabling clerical workers to carry out repetitive data entry tasks with relative ease, by using the same format as actual paper forms and retaining the characteristics of manual data entry as much as possible. Thus, a logical first step to design the VIS model is to reflect on the tasks that are executed regularly by the experts in the hot-test operations.

Using Mintzberg's (1973) definitions, it is recognised in Section 5.3 that the experts in the case study are primarily resource allocators. Their responsibility entails assigning untested engines to vacant hot-test cells for testing, with the main aim to maximise the number of engines that are tested and sent to ATD by the end of their shifts. Further to this, they also aim to distribute the workload equitably among all hot-test cell operators. Also, it is realised that the hot-test operations are essentially a non-active system, in the sense that it does not actively seek the experts' interventions. Instead, most of the interventions performed by the experts are more likely to be initiated by them in order to maintain an efficient and smooth workflow, and prevent bottlenecks from developing. In other words, the experts' interventions are usually not performed in response to actual problems in the hot-test operations, but are performed to pre-empt problems from occurring in the first place. As these interventions are often *ad hoc* in nature, it is inherently difficult to pinpoint and replicate the exact circumstances that would stir the experts to action. Hence, the VIS model used in the experiment needs to be designed in such a way that the experts are given the ability to observe the simulated hot-test operations as they do with the real-world hot-test operations, as well as the freedom to determine when an intervention should take place and what it should be.

In view of the nature of the experts' interventions, Flitman and Hurrion's (1987), Hurrion's (1991) and Liang's *et al*. (1992) idea of using the VIS model in a gaming mode to elicit expert knowledge comes to mind. It is because in doing so, the experts are provided with a quasi-realistic environment that allows them to behave and function in the same manner as in the real-world hot-test operations. That is, the experts are able to monitor the workflow in the simulated hot-test operations and intervene as necessary.

As a consequence, through their interactions with the VIS model, it is imagined that the experts will be conveying their knowledge in the most unadulterated and faithful form.

6.3 MITIGATING MEASURES TO IMPROVE THE QUASI-EXPERIMENTAL DESIGN

A common concern with the repeated measures experimental design is the possibility of carrying over practice and fatigue effects from one experimental condition to another (Field and Hole, 2006). As each expert will be exposed to all of the conditions of the experiment, there is a real chance that his exposure to one condition can affect his performance in another. It is because the expert, being human, can become fatigued, bored, better practised at playing the VIS-based game model, and so on. In this way, a systematic effect might be introduced into the experiment that would interact with the experimental manipulations and confound the findings.

In order to suppress the influence of these carry-over effects, Field and Hole (2006) recommend randomising the order of the different conditions that each expert will be exposed to. The authors further suggest that if complete randomisation is not possible, then mitigating measures should be taken to reduce any systematic differences that are produced between conditions as much as possible. However in the latter case, the experiment will no longer be considered a true experiment and become a quasi-experiment that is described at the start of this chapter. Unfortunately, the experiment encountered time constraint and managerial resistance that limited the order of conditions which each expert is exposed to.

In order to expedite the experiment, the experts are arranged to begin playing with any game model that is completed first. Furthermore, as the management generally resists the notion of losing productivity as a result of the experts absenting themselves to participate in the experiment, steps are taken to avoid incurring the former’s wrath by not removing the experts from their work for an extended period of time. Since the 3D-VIS models are expected to run very slowly relative to the 2D-VIS and 2½D-VIS models, the experts are arranged to play the 3D game models only after they have played with the rest. Consequentially, as the 2D game models are produced first, followed by the 3D and then the 2½D game models, the experts are arranged to start playing the 2D game models (for experimental condition I – 2D representation with unadjusted parameters; and II – 2D representation with adjusted parameters), followed by the 2½D game models (for experimental condition III – 2½D representation with unadjusted parameters; and IV – 2½D representation with adjusted parameters), and finally the 3D game models (for experimental condition V – 3D representation with unadjusted parameters; and VI – 3D representation with adjusted parameters). The actual knowledge elicitation timeline for the experiment is produced in Table 6.1.

Subject	Week																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A					I					II					III	IV			V VI
B		I II		III	IV			V	VI										
C	I		II	III						IV			V			VI			
D	I		II	III				IV					V	VI					
E	I			II				III		IV				V			VI		
F			I II		III				IV							V		VI	
G									I			II			III IV		V	VI	
H										I			II			III IV		V	VI

I to VI denote the six conditions (Section 6.1) in the experiment

Table 6.1: The actual knowledge elicitation timeline for the experiment

Due to the constraints explained above, the experiment is relegated to being a quasi-experiment. Following Field and Hole's (2006) suggestion, two measures are taken to mitigate the situation. Firstly, there is an implicit randomising element at work when the VIS model is used in a gaming mode. The mechanism behind a running VIS model is such that the states of the model are highly interdependent. Whilst each state of the VIS model is related to and reliant on the states that preceded it, it also has a repercussion on the states that follow it. Hence, when an expert decides to intervene with a running VIS model, his interaction at that point will inevitably alter the original course of the running model. This will produce a knock-on effect as a different course will present a different set of situations that evokes a different set of interventions from the expert. In this way, it is unlikely that a state will reappear exactly in another condition, which diminishes the danger that the expert might carry over any practice effects from participating in one condition to the next.

Secondly, the knowledge elicitation schedule has been planned so that an expert will not participate in more than two knowledge elicitation sessions in a fortnight. In addition, there is always a lapse of at least two days between two sessions. Although this arrangement prolonged the entire data collection process, it is necessary in order to minimise any disruptions that would be imposed onto the hot-test operations by not availing the expert to work. Besides, this also allows the expert to have sufficient time to recover between sessions, and perhaps undo any learning gained in the previous session. As a result, the probability that the expert will carry over any practice and fatigue effects from participating in one condition to the next is reduced even further.

6.4 CONCLUSION

This chapter explains the reasons behind choosing a repeated measures experimental design, as well as employing the VIS model in a gaming mode to elicit expert knowledge. In addition, it also justifies how the latter, together with a deliberate knowledge elicitation schedule, could diminish any practice and fatigue effects that might arise from the experiment and undermine its findings.

In the next chapter, the details of how a VIS-based game model is built and assessed are described. It belongs to the process of building and assessing the VIS model (process iii and iv in Figure 4.1) and is the second of three concurrent processes described in the research methodology in Section 4.5, which follow the process of understanding the case study (process i) in Chapter 5.

7 Visual Interactive Simulation Game Model

It has been assumed that if the experts are able to perform their decision-making intuitively in a familiar setup, then they might be able to convey their knowledge in an unadulterated and faithful form. Thus, in light of the experts' decisional role and its nature in the hot-test operations, it is determined in Section 6.2 that employing the VIS model in a gaming mode affords the most congenial setup to elicit their knowledge. It is because in so doing, the experts are provided with a quasi-realistic environment that allows and encourages them to behave and function in the same manner as they do in the real-world hot-test operations.

To give the VIS-based game model building process a kickstart, Ford supplied a current and detailed 2D-VIS model of its entire Puma diesel engine assembly line in Dagenham. A screenshot of Ford's 2D-VIS model is displayed in Figure 7.1. The VIS model was developed using WITNESS, a COTS VIS software from Lanner Group Limited. As such, instead of building a new game model entirely from scratch, Ford's existing model could be adopted as a base model, on which all game model building efforts will concentrate on adapting it for the experiment's needs. Since Ford's model was in use then, it could be assumed that any worries on its currency were unfounded at the time of this research. In the midst of the various adaptations, additional attention was paid to ensure that the information offered in the game models is consistent with those in the real world. In other words, the game model should provide neither more nor less information than what an expert is able to obtain in the real-world working environment.

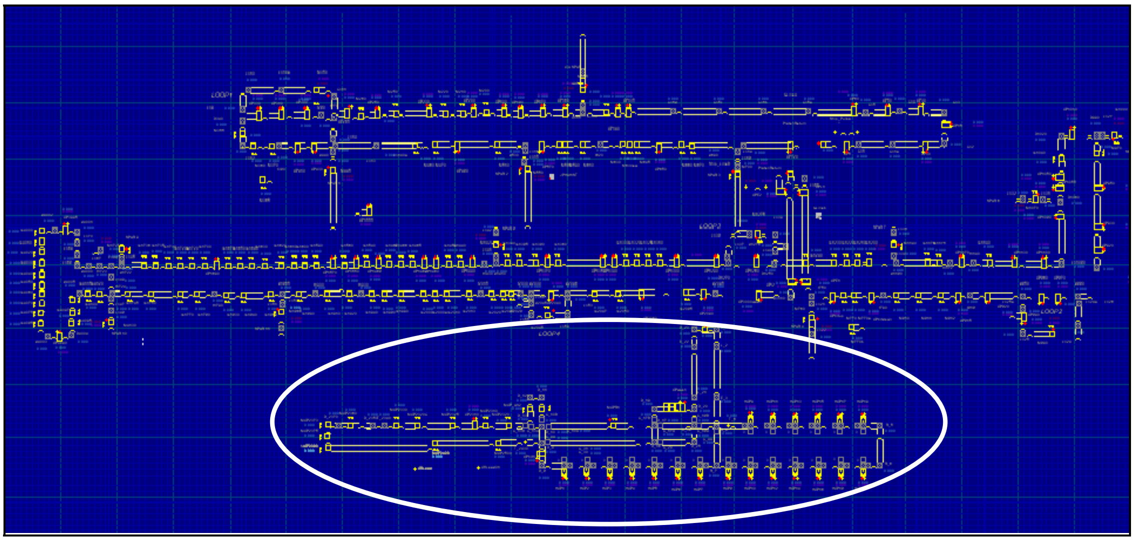


Figure 7.1: The original 2D-VIS model provided by Ford that spans the entire Puma diesel engine assembly line. The hot-test and ATD operations (Model B) are as circled.

Using the information gathered from the fact-finding process in Chapter 5, the base model provided by Ford was first reduced in size and modified to improve its utility. Next, the base model's code was removed of any decision rules that were originally included to represent the experts' presumed decision-making in the hot-test operations. Then, further work was done to incorporate a gaming facility into the base model, thereby transforming it into a game model. Following this, six versions of the game model were produced, with one for each experimental condition defined in Section 6.1. Lastly, the game models were assessed to make sure that they were sufficiently accurate for the experiment's purposes. These procedures are described in the sections that follow.

7.1 GAME MODEL CONSTRUCTION

7.1.1 ADAPTATIONS MADE TO IMPROVE THE BASE MODEL'S UTILITY

The first impression of the base model in Figure 7.1 is its great expanse, as it spans the entire Puma diesel engine assembly line. This makes the base model unwieldy to manage, and causes it to run quite slowly. Consequently, the base model is expected to impede the data collection process, if it is used as the vehicle for eliciting knowledge from the experts. As Robinson *et al.* (2001 and 2005) and Alifantis (2006) have conceded previously that the VIS-based knowledge elicitation experience is a laborious and time-consuming one for participating experts in Section 4.2.4, it is pertinent that every effort is made to expedite the data collection process as much as possible. A means to do so is by improving the base model's running efficiency.

There are two options to improve the base model's efficiency, and both of them resort to simplifying the base model whilst upholding its credibility. The first option entails the removal of a big section of the base model that mimics the pre-hot-test operations (the assembly operations on Assembly Line A and B). Then, a 'black-box' that uses an appropriate time delay to imitate the time spent by an engine on the assembly operations is created as a replacement (Robinson, 2004). In short, model entities like engine parts are made to enter the black-box and leave after some time as assembled engines, instead of entering into a model of the assembly operations. The time duration between entering and leaving the black-box is sampled from a distribution that is built from the actual times spent by model entities in the assembly operations section of the base model. Alternatively, the second option entails the splitting of the original model into two sub-models. The first sub-model (Model A) will mimic the pre-hot-test operations, whilst the second sub-model (Model B) will mimic the rest of the engine assembly line. This includes the hot-test and ATD operations, which are circled in Figure 7.1. Under this option, when Model A is run, its output data such as the times when model entities leave it are collected and written to a data file. The contents of the data file are then used as input data for Model B when it is run, to re-create the engine entities at the times they left Model A.

Out of the two options described above, the latter option was preferred and executed. It is because the experts do not rely on any information from the assembly operations to aid their decision-making, and hence including the black-box into the game model will not serve any particular purpose in subsequent knowledge elicitation sessions. Nonetheless, there is a limitation for using Model A's output data file contents in its original, raw form as input data for Model B. As the arrival rate of assembled engines entering Model B will be used as an instrument for an experimental factor (model parameters), it is better to use the raw values to construct a pseudo-empirical distribution that can be manipulated easily to cater for different experimental conditions. The distribution constructed in this way is not considered a *bona fide* empirical distribution, as the raw values used are output data from another simulation model (Model A) and not historical data collected on the grounds of the engine assembly line.

7.1.2 ADAPTATIONS MADE TO IMPROVE THE BASE MODEL'S LOGIC

Following the changes made to improve the base model's efficiency, the smaller base model's code was scrutinised for its logic. This is necessary because the base model is originally built for a purpose different from the experiment. In the real world, there are several operations that actually require human supervision and intervention in practice, but these had been substituted by some pre-set decision rules in the base model's code. Hence, there is a need to remove these decision rules from the section of the base model that mimics the hot-test operations, and augment the model code with functions to allow for the experts' intervention and interaction.

Further to this, there is also a need to configure the base model to enable it to record the experts' interactions as well as the situations in the base model when the experts intervene during simulation runs. As has been explained in Section 4.1, the set of decisions made in each interaction together with the set of attributes that describes the situation when an intervention takes place will constitute an example case. These decisions and attributes are represented by a mixture of counts, binary and categorical values. In this experiment, each expert will play with six different game models, with one being built for each experimental condition. The example cases that are recorded in these knowledge elicitation sessions will be used subsequently to test the hypotheses in Section 4.4.

7.1.3 ADAPTATIONS MADE TO OPERATIONALISE THE GAME MODEL

The game model is basically a VIS model equipped with a control bar for gaming purpose. In a nutshell, the control bar enables the experts to access the newly-incorporated functions mentioned in the last section. As shown in Figure 7.2, it has a small window to inform on the current simulated time in the game model, as well as four groups of buttons that an expert would expect to use frequently whilst playing with the game model. These buttons are described below with references made to Figure 5.1 when necessary.

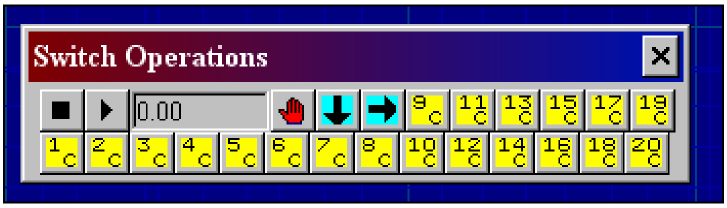


Figure 7.2: The control bar used to facilitate the experts' interventions and interactions

Firstly, there are two grey buttons at the top left corner of the control bar with either a square or triangle on them. These are standard buttons provided in WITNESS, where the button with a square is used to pause the running game model, and the triangle is used to resume running the game model.

Secondly, there is a button next to the small window on the top row, which features a red palm. It mimics the action of blocking Sensor 1 in the hot-test operations (Figure 5.1), which will then allow excess untested engines that are routed to Conveyor A6 to pass through Conveyor A5 and A3 for reassignment at Junction J.

Thirdly, there are two blue buttons with either a downward or rightward arrow drawn on them. These are buttons whose collective function mimics that of the path control switch (Figure 5.1), which the experts can access in the real world to perform their duties. On the one hand, the button with the downward arrow is used for electing to send incoming untested engines from Assembly Line B straight down to Conveyor C. On the other hand, the button with the rightward arrow is used for electing to send the incoming untested engines to the path on the right of Junction J. Alternatively, by depressing both blue buttons, the option for the automated cyclical mode of sending three untested engines to the right of Junction J followed by one straight down is activated.

Lastly, there are 20 yellow buttons labelled from ‘1C’ to ‘20C’. These are buttons whose functions mimic those of the 20 cell/stand control switches described in Section 5.2, which the experts can access in the real world to perform their duties. In essence, these buttons are used for switching on or off the hot-test cells and their respective adjacent waiting stands. The number on a button’s label corresponds to the label value of the hot-test cell that the button purports to control. For instance, the yellow button with label value ‘1C’ is used to switch on or off the hot-test cell with label value ‘1’.

7.1.4 ADAPTATIONS MADE FOR THE EXPERIMENTAL CONDITIONS

After adopting Ford’s 2D-VIS model as a base model and subjecting it to a series of adaptations before transforming it into a game model fit for the experiment’s purposes, it was adapted for the final time to produce six different versions. Each version was tailored to investigate a different experimental condition defined in Section 6.1. These six experimental conditions are:

I. 2D representation with unadjusted parameters;

II. 2D representation with adjusted parameters;

III. 2½D representation with unadjusted parameters;

IV. 2½D representation with adjusted parameters;

V. 3D representation with unadjusted parameters; and

VI. 3D representation with adjusted parameters.

In summary, the 2D game model was first reproduced in two other visual representation dimensions: 2½D and 3D. Next, these three game models were checked for their adherence with the information consistency principle. Then, each game model was duplicated with a different set of model parameters. In this way, six different game models were produced for the experiment to be carried out later. These adaptations are described in more detail below.

Adaptations related to visual representation dimension (Factor One)

At this juncture, it should be pointed out that the game model that was adapted originally from the base model is already in 2D form. It was then decided that the game model was used as-is for the 2D version of the game models. To reproduce the 2D game model into the 2½D and 3D versions, specialist drawing applications were used to first reproduce each icon in the 2D game model in 2½D and 3D forms. Then, the newly-drawn 2½D and 3D icons were used to produce the respective 2½D and 3D versions of the game model. For illustration purpose, the original 2D icon and its 2½D and 3D equivalents for representing a hot-test cell, its adjacent waiting stand and a small section of conveyor are depicted in Figure 7.3.

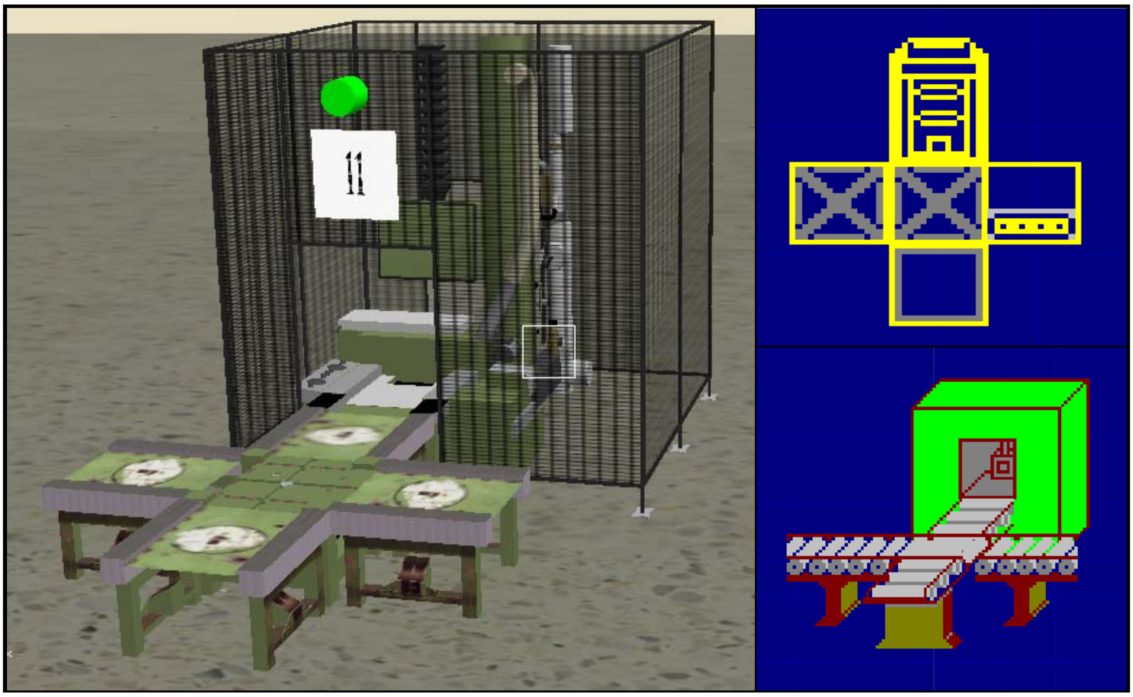


Figure 7.3: (Clockwise, from top-right) The 2D, 2½D and 3D icons used in the game model to represent a hot-test cell, its adjacent waiting stand and a small section of conveyor

On the one hand, CorelDRAW and Microsoft Paint are the two specialist drawing applications that were used to draw the 2½D icons. Drawing these 2½D icons is a relatively straightforward task, as they are basically three-dimensional icons drawn against a plain, two-dimensional background with no perspective projection (where more distant objects are drawn smaller relative to those that are closer to the eye). Also, there is no need to make these 2½D icons look photo-realistic. After these 2½D icons were drawn, they were used to replace the 2D icons in the 2D game model to produce the 2½D game model.

On the other hand, mantra4D (Lanner, 2007) is the specialist drawing application that was used to create the 3D icons. In comparison, creating these 3D icons is a more tedious task, as they are three-dimensional photo-realistic icons created against a three-dimensional photo-realistic background with perspective projection. Briefly, each component of an icon was initially drawn separate from one another, using appropriate materials, textures, and lighting and shading effects to make the component look photo-realistic. Then, the finished components were oriented and aligned for assembly to create the 3D icon. After these 3D icons were created, they were linked with the model entities in the 2D game model using the 'fast build' facility in WITNESS VR (an integrated module in the WITNESS software – Lanner, 2007) to produce a 3D representation. As such, the 3D game model is actually a 2D game model with an alternative 3D display.

Adaptations made to preserve information consistency

It had been determined at the outset that the game models should provide neither more nor less information than what an expert is able to obtain in the real-world working environment. This principle essentially requires the game models to provide information that approximates to those received by the experts in the real world. For instance, due to the physical layout of the hot-test operations and its size in the real world, an expert is not able to capture the entire hot-test operations within his peripheral vision. In reality, it was observed that the expert can only monitor around half of the hot-test operations at a time. As such, the appearances of the game models on the computer's visual display were adjusted to imitate the limited peripheral vision of the human eye. In both 2D and 2½D game models, this limitation was replicated by enlarging them to the extent that the computer's visual display only reveals half of them at a time. However, no adjustment was made to the 3D game models' appearances. It is because they use perspective projection that makes distant objects to appear smaller relative to those that are closer to the eye. In this way, the 3D game models mimic the experts' limited peripheral vision by making objects that are very far away to appear so small that the experts cannot have a clear view of them. The adjustments made to the 2D and 2½D game models, and the visual effect of perspective projection in the 3D game models can be viewed in the catalogue of game model screenshots displayed in Section 7.3.

In another instance, there is a three-colour signal light (green-red-yellow) outside each hot-test cell that is used for indicating its operational state. An example of this signal light can be viewed at the top-middle of Figure 5.3. On the one hand, when the light turns green, this signals that the hot-test cell has finished testing the engine inside it, and it is safe to de-rig the newly-tested engine. On the other hand, when the light turns red, this signals that there is an emergency in the hot-test cell, or it has broken down and is being repaired at the moment. Lastly, when the light turns yellow, this signals that an engine is being currently tested in the hot-test cell. This piece of information was replicated in the game models by applying a colour scheme similar to the signal light on the hot-test cell icons. In both 2D and 2½D game models, the colours of the hot-test cell icons will change in accordance with their respective operational states. Likewise, in the 3D game models, there is a small cylinder attached to each hot-test cell icon (above its label) that mimics the signal light's function. Using Figure 7.3 as an example, the 2D, 2½D and 3D icons' current operational states are signalled by the yellow, green and green colours respectively.

Adaptations related to model parameters (Factor Two)

After adapting for the different visual representation dimensions and adherence with the information consistency principle, there are now three game models for the following experimental conditions:

I. 2D representation with unadjusted parameters;

III. 2½D representation with unadjusted parameters; and

V. 3D representation with unadjusted parameters.

In general, there are two typical scenes in an expert's work shift: standard and non-standard. These have been elaborated earlier in Section 5.4. To recap, a standard shift in the hot-test operations is typified by an irregular flow of untested engines entering the hot-test operations. As a result, there are normally more vacant hot-test cells that are idling than there are untested engines to fill them. Conversely, a non-standard shift in the hot-test operations is typified by a regular flow of untested engines entering the hot-test operations. More often than not, this means that the hot-test operations will be swarmed with both tested and untested engines. Both these scenes had been encapsulated in the base model's original unadjusted parameters and can be mimicked by running it as-is. Accordingly, the game model is expected to simulate these scenes when it is using the unadjusted parameters. In addition, being imitative of the real world, it is also expected to simulate more standard scenes than non-standard ones.

Following the adaptations made in the last two sections, the game models for experimental condition I, III and V were duplicated with the set of adjusted model parameters. Thus, another three game models were produced for the remaining experimental conditions. They are:

II. 2D representation with adjusted parameters;

IV. 2½D representation with adjusted parameters; and

VI. 3D representation with adjusted parameters.

Like the game models with the unadjusted parameters, these game models with the adjusted parameters will also simulate both standard and non-standard scenes in a work shift. However, the latter will simulate more non-standard scenes than standard ones. On the one hand, in the game models with the unadjusted parameters, the inter-arrival times between untested engines entering from Assembly Line B into the hot-test operations will be sampled randomly from the pseudo-empirical distribution constructed in Section 7.1.1. On the other hand, in the game model with the adjusted parameters, the inter-arrival times between untested engines are kept constant for each period[6] in the shift, in order to regulate the engine flow and simulate more non-standard scenes. The distributions of inter-arrival times that were used in the unadjusted and adjusted parameters are plotted in Figure 7.4 and Figure 7.5 respectively.

[6] There are four periods in each simulated work shift.

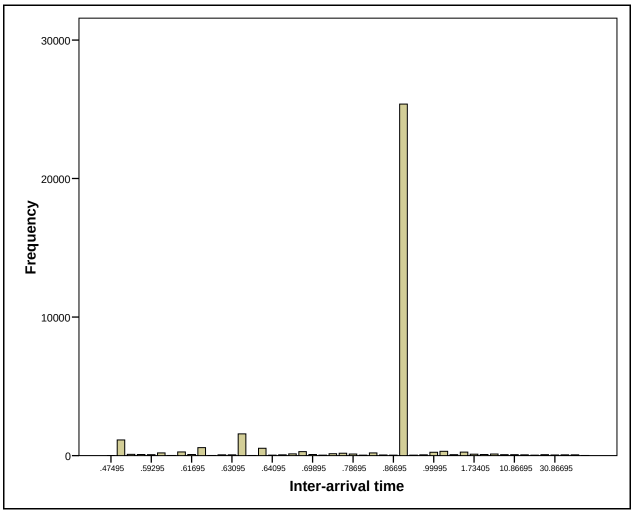


Figure 7.4: Distributions of inter-arrival times (simulated minutes) between untested engines used in the unadjusted model parameters

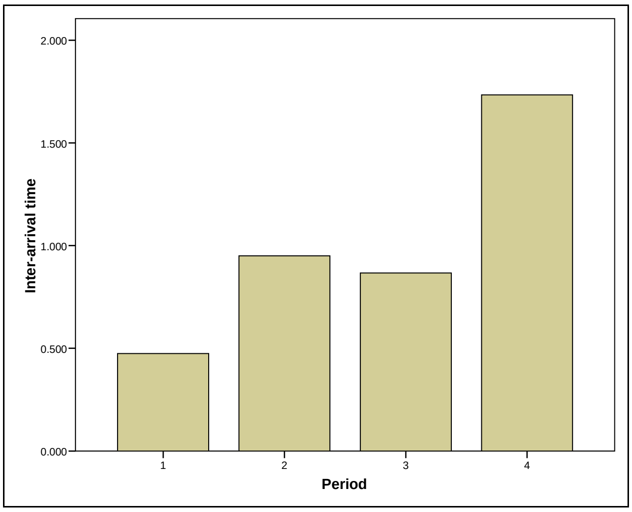


Figure 7.5: Distributions of inter-arrival times (simulated minutes) between untested engines used in the adjusted model parameters

Furthermore, the game models with the adjusted parameters will also produce scenes that are more extreme and perhaps difficult than those that actually occur in the real world. For instance, there will be a greater interspersion of untested 2*l* and 2.4*l* engines entering from Assembly Line B into the hot-test operations than in the real world. The distributions of engines that were used in the unadjusted parameters and adjusted parameters are tabulated in Table 7.1. These distributions were actually applied in a repetitive order. As an illustration, the first batch of untested engines that enters into the hot-test operations in a game model with unadjusted parameters will be made up of 381 2*l* engines. This will then be followed by a second batch of 30 untested 2.4*l* engines, and so on. When the final batch of 52 untested 2.*l* engines enters into the hot-test operations, the distribution will be repeated with the next batch of untested engines being made up of 381 2*l* engines again.

Engine capacity (litres)	Quantity of engines	
	Unadjusted	Adjusted
· 2 *l*	381	10
· 2.4 *l*	30	20
· 2 *l*	120	40
· 2.4 *l*	30	10
· 2 *l*	100	20
· 2.4 *l*	161	40
· 2 *l*	150	10
· 2.4 *l*	56	20
· 2 *l*	52	40
· 2.4 *l*	-	10
· 2 *l*	-	60

Table 7.1: Distributions of untested engines entering into the hot-test operations used in the unadjusted and adjusted model parameters

As well, the game models with the adjusted parameters will have more engines detected as defective in the hot-test operations, and less so in the UV and ATD operations than in the real world. Since all defective engines are eventually routed to the repair station within the hot-test operations for rectification (Figure 5.1), the defective rates in the post-hot-test operations (*i.e.* the UV and ATD) need to be reduced in order not to choke the hot-test operations with defective engines and bring the simulation run to a possible standstill. Lastly, the hot-test cells in the game models with the adjusted parameters will also break down more often than they actually do in the real world. These differences are summarised in Table 7.2.

Model parameter	Unadjusted	Adjusted
· Percentage of engines that fail in the hot-test operations, and require major (minor) repair	5% (5%)	12% (15%)
· Percentage of engines that fail in the UV operations	2%	1.5%
· Percentage of engines that fail in the ATD operations	2%	1.5%
· Percentage of times a hot-test cell breaks down	Value varies for each hot-test cell	Larger of triple 'Unadjusted' value, and 15%

Table 7.2: A summary of engine defective rate and hot-test cell breakdown rate used in the unadjusted and adjusted model parameters

As such, the overall effect from making these adjustments is to create more chaotic scenes of the hot-test operations being swarmed with an erratic mixture of untested 2*l* and 2.4*l* engines, tested 2*l* and 2.4*l* engines as well as an unusually high proportion of defective engines, and coupled with more frequent hot-test cell breakdowns.

7.2 GAME MODEL ASSESSMENT

Finally, the fully adapted game models for condition I (2D representation with unadjusted parameters) and V (3D representation with unadjusted parameters) were assessed for their face validity from the experts' point of view (Pidd, 2005). In addition, they were also assessed for their usability from the experiment's point of view. These assessments were carried out with the help of two hot-test personnel. They are the staff who are identified earlier in Section 5.5 as being qualified to perform switch operations, but whose commitment to participate in the experiment could not be secured. The game model for condition III (2½D representation with unadjusted parameters) was not assessed, as it is identical to the game model for condition I except for the three-dimensionality of its icons. Likewise, the game models for condition II (2D representation with adjusted parameters), IV (2½D representation with adjusted parameters) and VI (3D representation with adjusted parameters) were not assessed, as they are all visual duplicates of the game models for condition I, III and V respectively.

7.2.1 ASSESSMENT FOR MODEL FACE VALIDITY

There are mainly two means of assessing a VIS model for its validity: black-box validation, and white-box validation (Robinson, 2004; Pidd, 2005). On the one hand, black-box validation is about performing macro checks to determine whether the overall VIS model represents the real world with sufficient accuracy for the purposes at hand. However, as the experiment does not rely on the game models' overall ability to mimic and/or predict the real world for its purposes, it is not material to establish black-box validity. On the other hand, white-box validation is about performing micro checks to determine whether the constituent parts of the VIS model represent the corresponding real-world elements with sufficient accuracy for the purposes at hand. In effect, it is establishing the model's face validity. As such, white-box validation is used as the means of assessment here.

It is crucial to recruit the assistance of the people who are knowledgeable about the real-world system and tap their detailed knowledge for testing for white-box validity (Robinson, 2004; Pidd, 2005). These people are the hot-test personnel described above, and they were enlisted to do so via verbal description and visual check. Firstly, Pidd (2005) suggests that the static logic of the VIS model is checked by describing it verbally to the hot-test personnel using a natural-language (in this case, the English language). The static logic in a VIS model comes in the form of decision rules that govern the behaviour of model entities during a simulation run. Secondly, Robinson (2004) and Pidd (2005) also suggest that the dynamic logic of the VIS model is checked by running the model and asking the hot-test personnel to watch its behaviour. In both cases, the hot-test personnel were asked to provide their feedback on the game models' logic.

7.2.2 ASSESSMENT FOR MODEL USABILITY

The game models were assessed for their usability through observation interviews, a knowledge elicitation technique described in Section 2.4.2. These were conducted in tandem with a series of trial runs. Initially, the two hot-test personnel were asked to try out the game models and their activities with the game models were observed and noted. In the meantime, the hot-test personnel were also asked to comment on the representativeness of the game models' layouts, as well as the suitability and completeness of the information that were presented whilst the game models were running. Lastly, they were also asked to provide their feedback on the game models' run speed and ease of use.

7.2.3 OUTCOME OF ASSESSMENTS

Following the assessments above, the game models were fine-tuned with the feedback that was received from the two hot-test personnel. Overall, the hot-test personnel were satisfied with both the static and dynamic logic of the game models. This established the game models' white-box validity and hence, face validity. Moreover, the hot-test personnel found it easy to use the control bar for executing their switching decisions in the game models. However, they commented that the game models would be easier to manage, if they were running at a visibly slower pace. Eventually, a slower run speed was agreed with the hot-test personnel after running the game models a few more times at various run speeds. Although the game models are running slower now, they only take approximately 10 minutes more than the initial 30 minutes or so to complete their runs.

In addition, they had been observed to display some degree of awkwardness when they tried to locate the corresponding cell/stand control switch for switching on or off a hot-test cell and its respective adjacent waiting stand. This was because the cell/stand control switches in the original control bar were not arranged in the same way as the hot-test cells in the game model (which were arranged in the same way as those in the real world). A screenshot to contrast the arrangement of the hot-test cells in the real world against the arrangement of the original control bar is displayed in Figure 7.6. In response, the cell/stand control switches were rearranged to give the improved control bar displayed earlier in Figure 7.2. Now, the cell/stand control switches share the same arrangement as the hot-test cells in the game model now. As such, locating a corresponding cell/stand control switch had become more intuitive and takes much less effort than before.

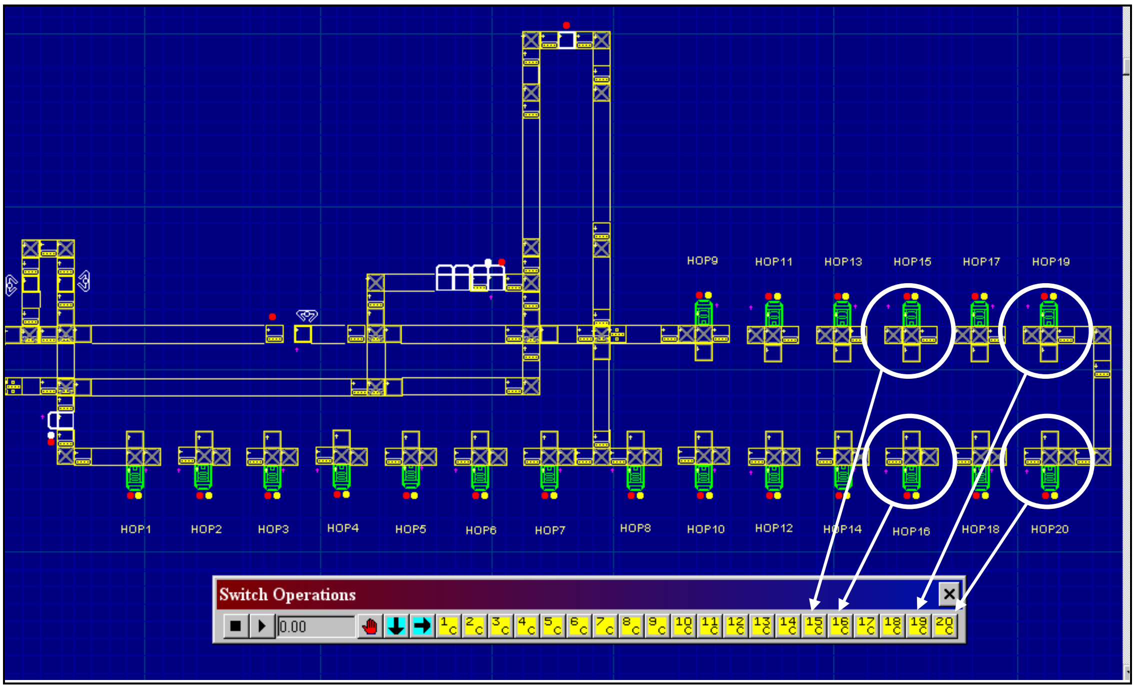


Figure 7.6: The original control bar set against a 2D game model

7.3 A CATALOGUE OF GAME MODEL SCREENSHOTS

A series of screenshots were captured from the finished 2D, 2½D and 3D game models after running them for 671.91 minutes part way through a shift, and displayed below. These are shown for comparing the different views that an expert can have under different visual representation dimensions, when the game model is paused at a certain point in simulated time. Moreover, these screenshots are also shown for contrasting the overall visual effect of a full set of model icons under different visual representation dimensions.

7.3.1 Screenshots from a 2D game model

The appearances of the 2D game model in the screenshots had been restricted in keeping with the information consistency principle described in Section 7.1.4. The left-half, middle-half and right-half of the game model are shown in Figure 7.7, Figure 7.8 and Figure 7.9 respectively. Also, the hot-test cell icons in the 2D game model had been coloured differently, according to their respective operational states. These colours were expected to change when the game model resumed running and the operational states changed.

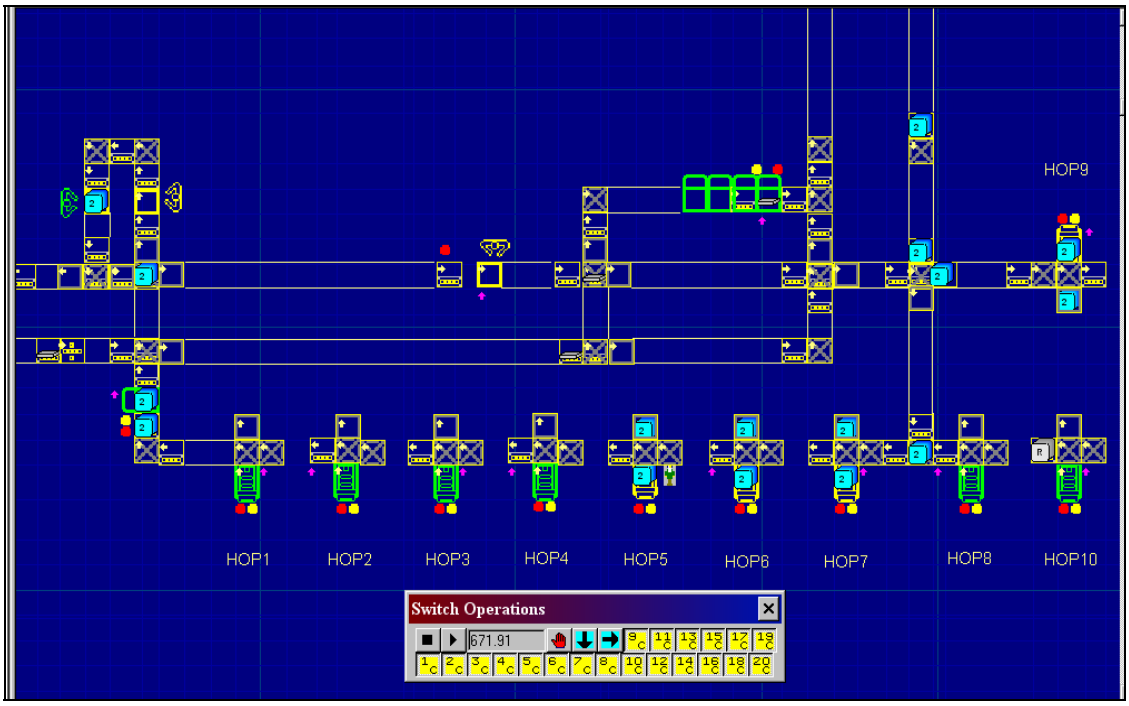


Figure 7.7: Screenshot 1 from the 2D model

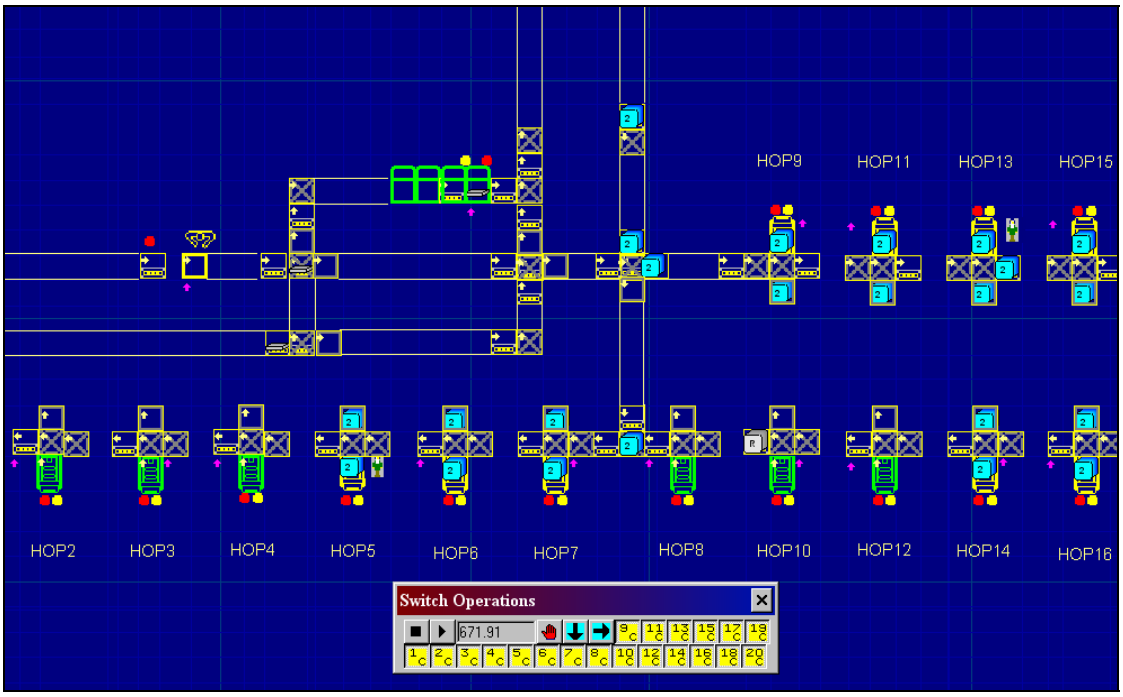


Figure 7.8: Screenshot 2 from the 2D model

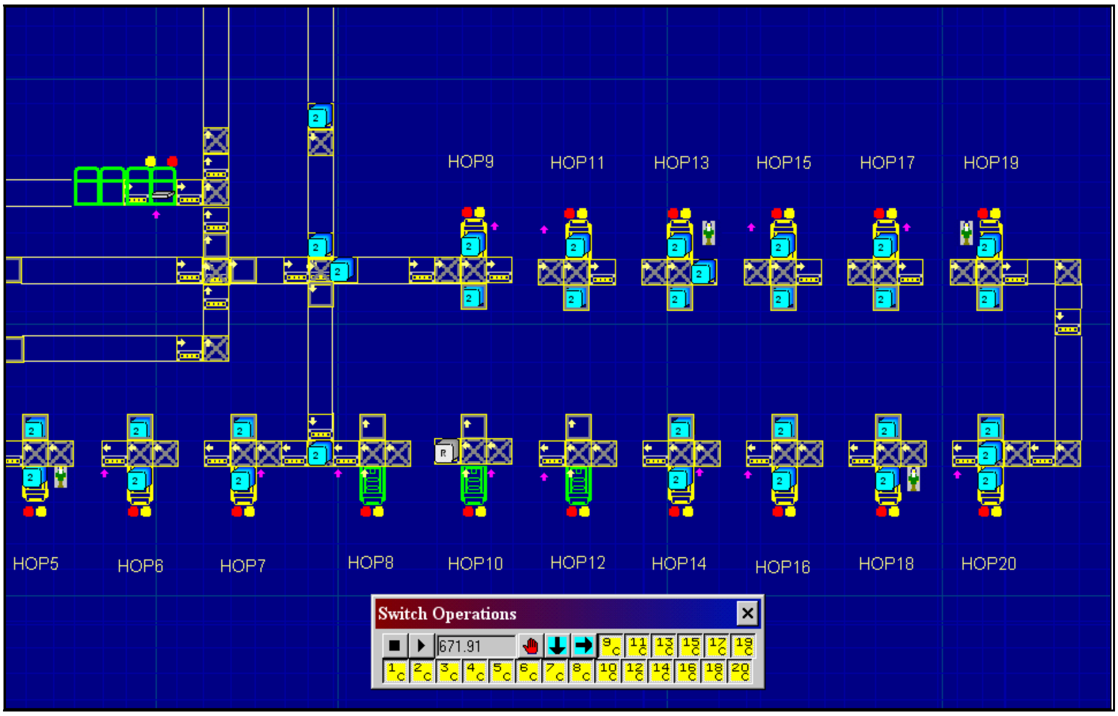


Figure 7.9: Screenshot 3 from the 2D model

7.3.2 SCREENSHOTS FROM A 2½D GAME MODEL

The appearances of the 2½D game model in the screenshots had been restricted in keeping with the information consistency principle described in Section 7.1.4. The left-half, middle-half and right-half of the game model are shown in Figure 7.10, Figure 7.11 and Figure 7.12 respectively. Also, the hot-test cell icons in the 2½D game model had been coloured differently, according to their respective operational states. These colours were expected to change when the game model resumed running and the operational states changed.

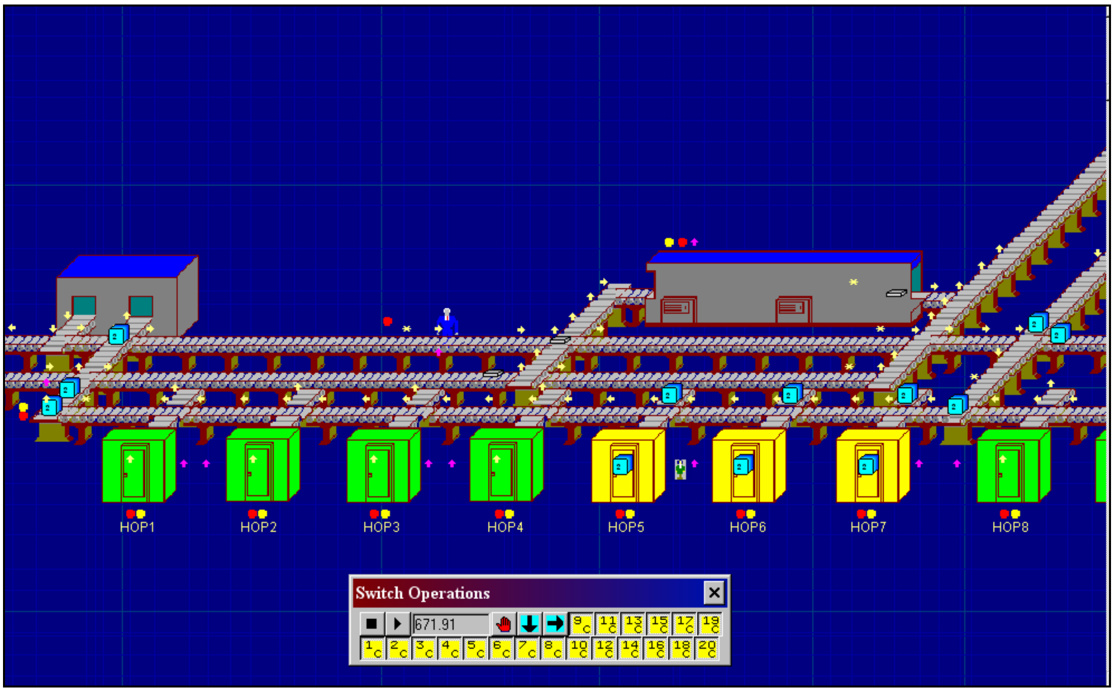


Figure 7.10: Screenshot 1 from the 2½D model

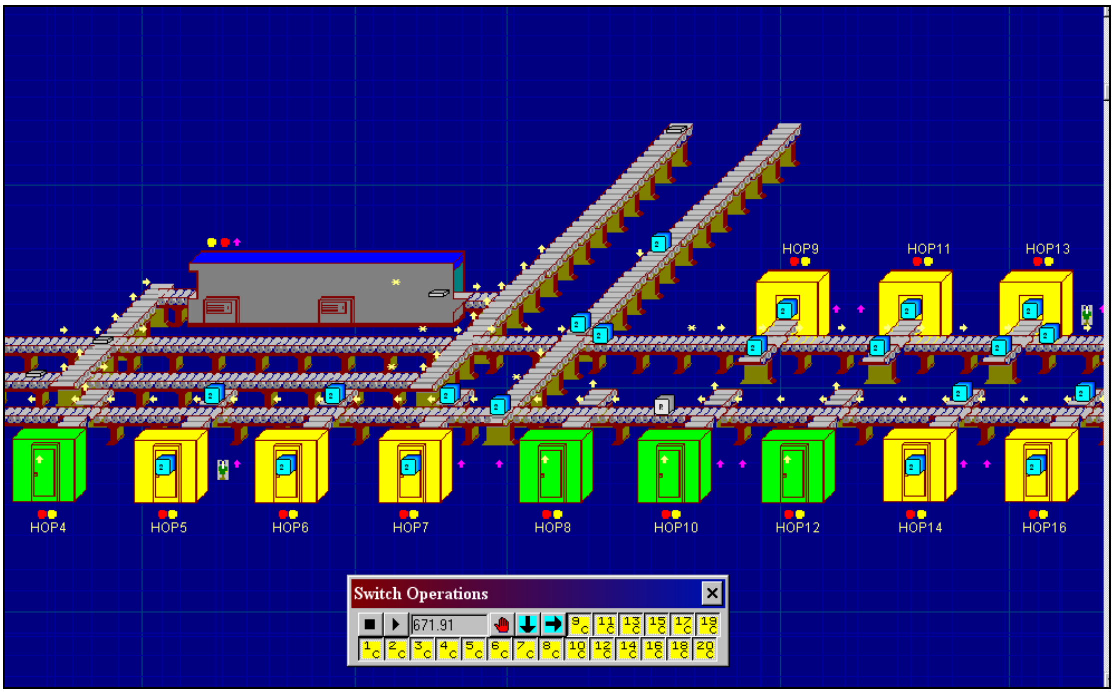


Figure 7.11: Screenshot 2 from the 2½D model

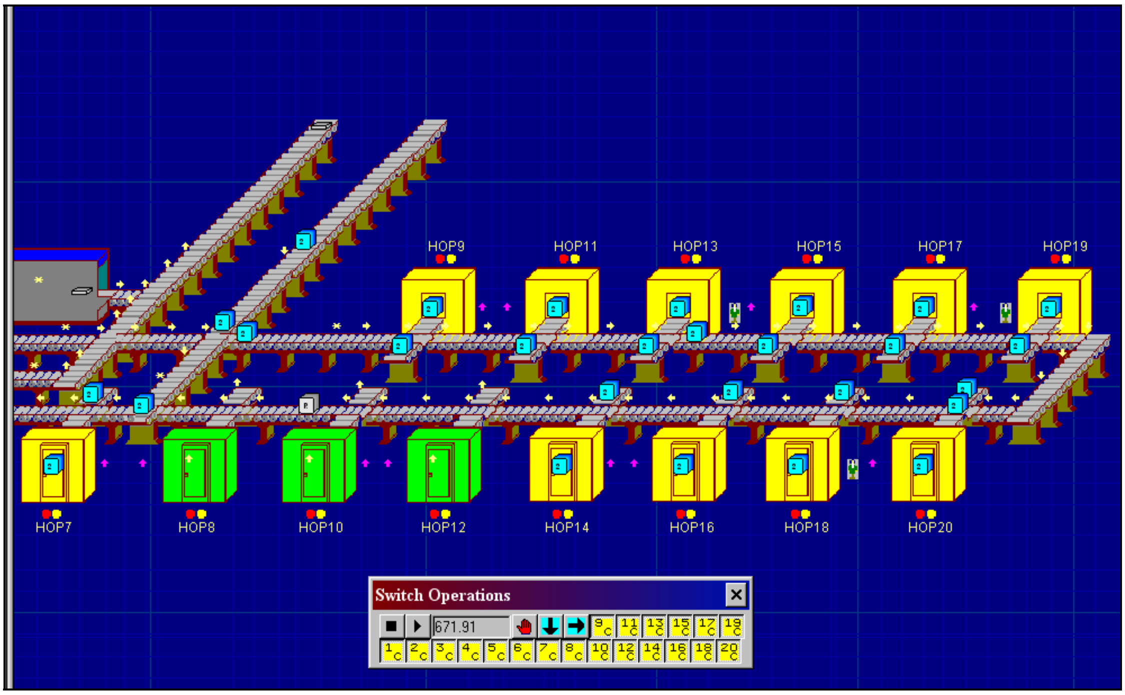


Figure 7.12: Screenshot 3 from the 2½D model

7.3.3 SCREENSHOTS FROM A 3D GAME MODEL

The visual effect of perspective projection is illustrated fully by the series of screenshots from the 3D game model. Figure 7.14 to Figure 7.18 illustrate the view provided by the 3D game model as one traverses from its left end to its middle. Also, in keeping with the information consistency principle described in Section 7.1.4, the small cylinders attached to the hot-test cell icons (above their labels) in the 3D game model had been coloured differently, according to the respective hot-test cells' operational states. These colours were expected to change when the game model resumed running and the operational states changed.

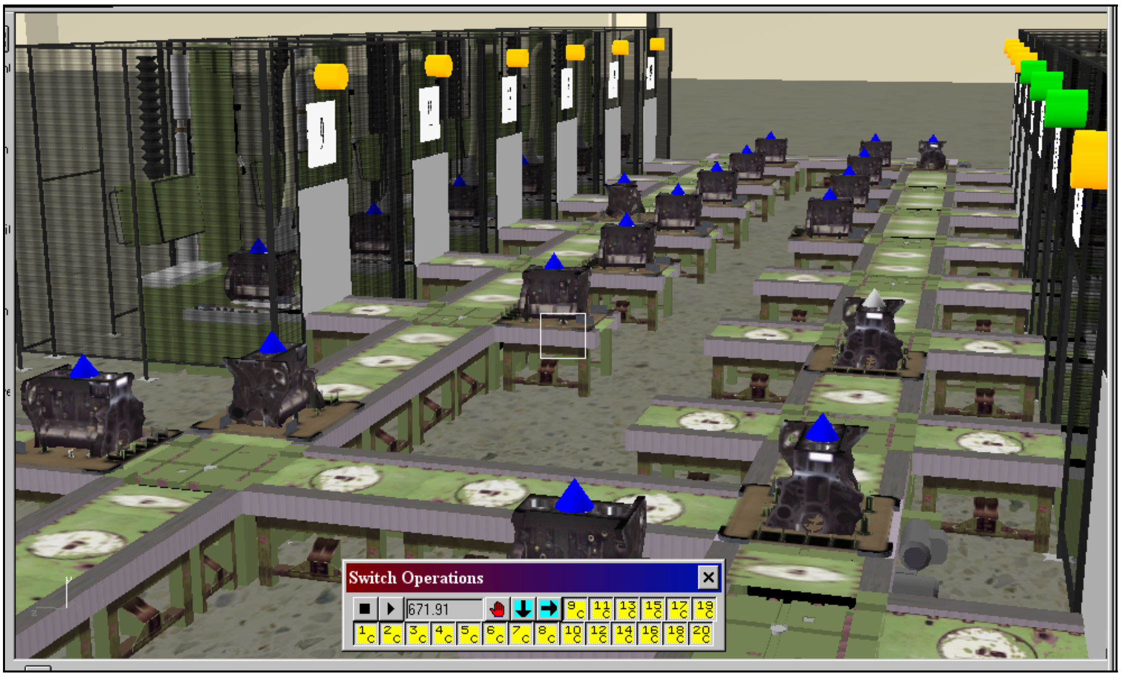


Figure 7.13: Screenshot 1 from the 3D model

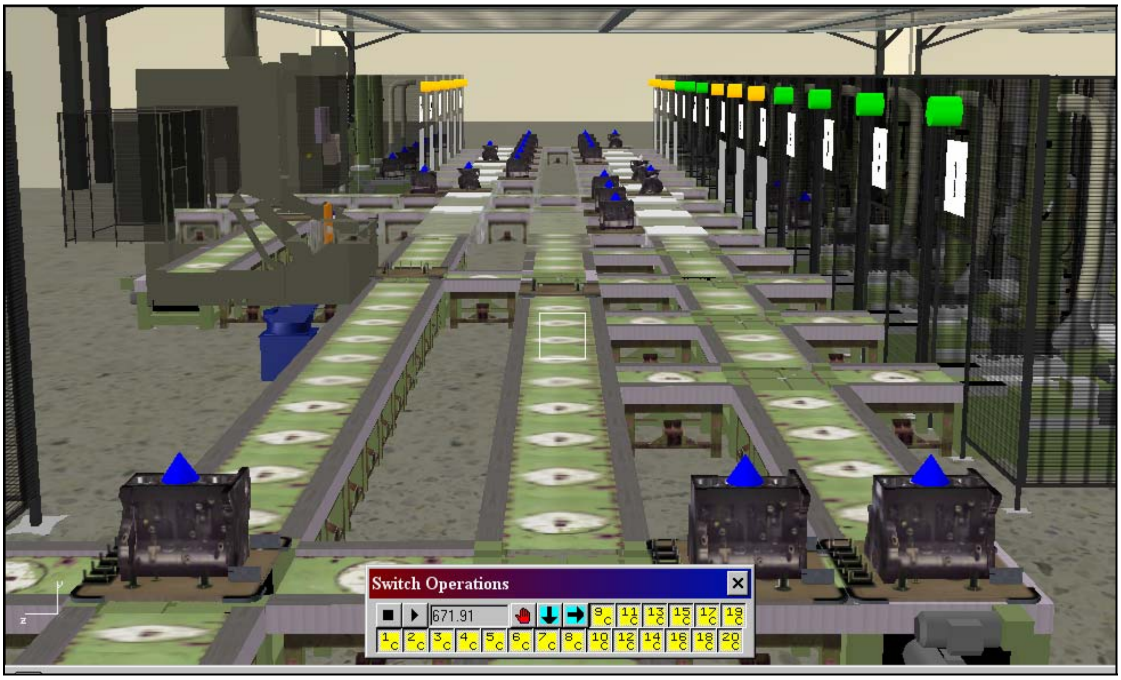


Figure 7.14: Screenshot 2 from the 3D model

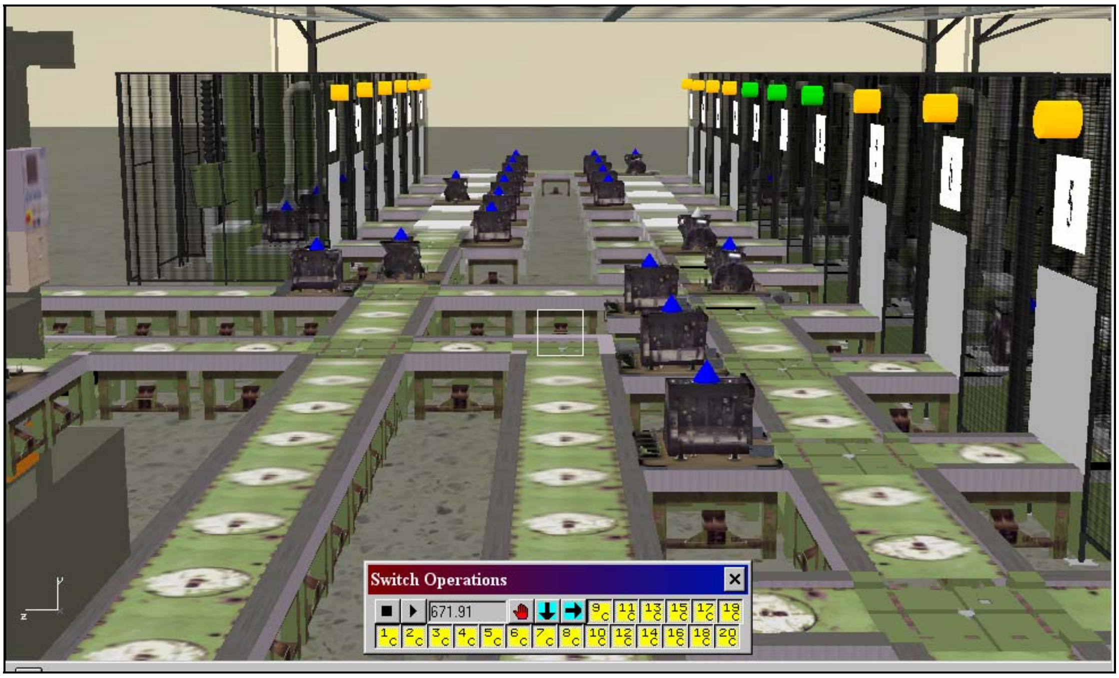


Figure 7.15: Screenshot 3 from the 3D model

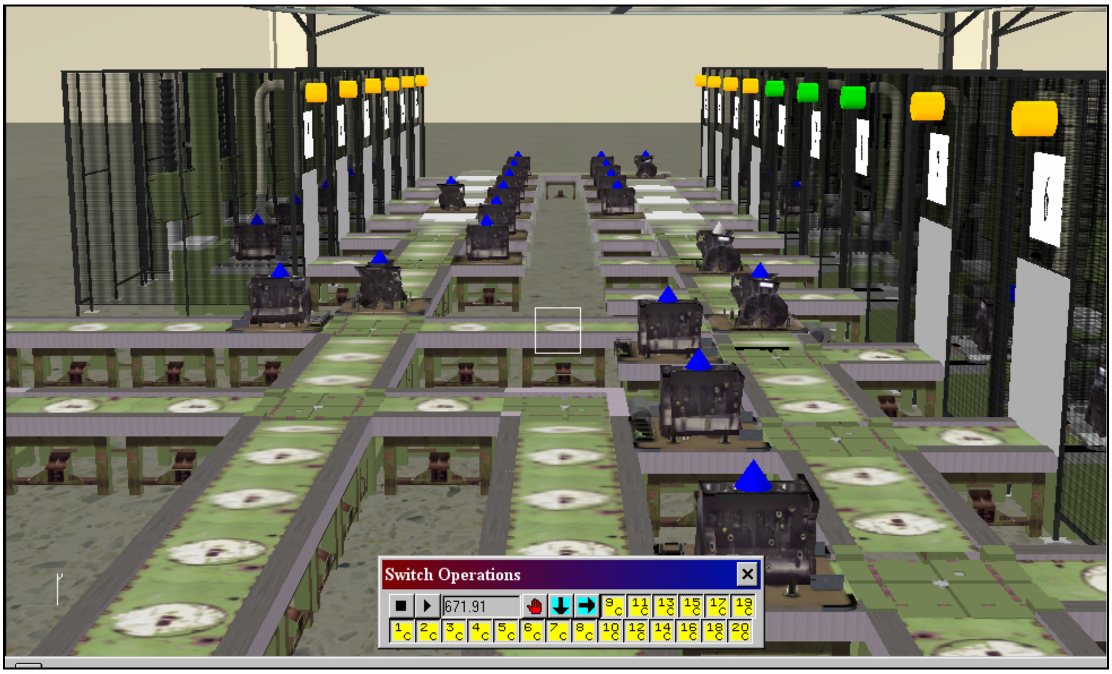


Figure 7.16: Screenshot 4 from the 3D model

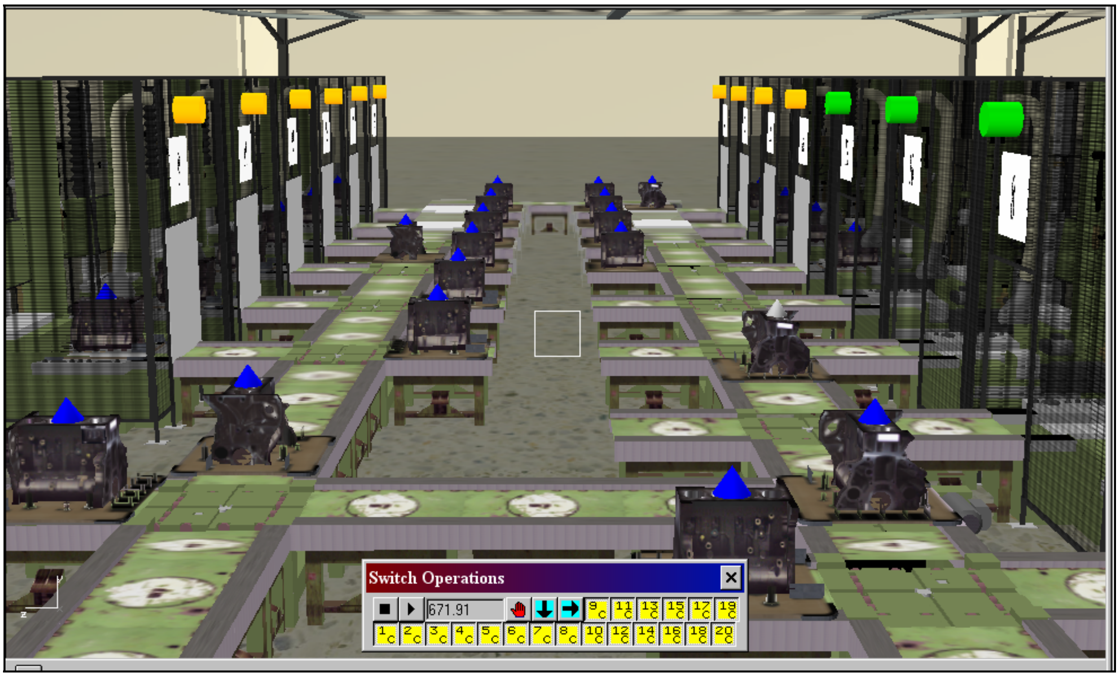


Figure 7.17: Screenshot 5 from the 3D model

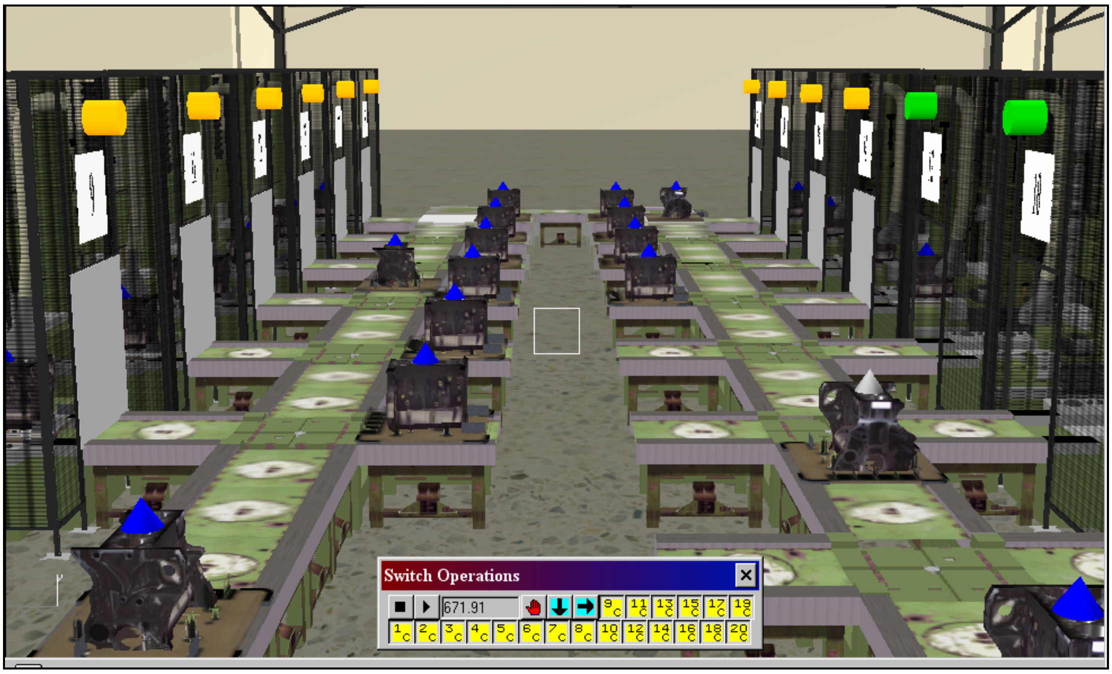


Figure 7.18: Screenshot 6 from the 3D model

7.4 CONCLUSION

This chapter begins by adopting Ford's current 2D-VIS model of the Puma diesel engine assembly line as a base model. It continues by describing the various adaptations made on the base model to produce the game models for the experiment. Firstly, work was carried out to improve the base model's utility and logic. Secondly, the base model was transformed into a game model by equipping the former with a control bar for gaming purpose. With this control bar, the experts will be able to access the gaming functions that were built earlier into the base model. Lastly, the game model was reproduced and then duplicated to produce six different versions for investigating the experimental conditions defined in Section 6.1. Whilst doing so, efforts were also made to ensure that the game models' fidelity is in line with the information consistency principle. Finally, the game models were assessed for their face validity and usability, followed by some finishing adjustments.

In the next chapter, four measures were conceived for evaluating the constructs that were used for assessing elicitation effectiveness and efficiency in the experiment. These constructs have been defined previously in Section 4.2. The next chapter belongs to the process of devising the measures (process v in Figure 4.1) and is the last of three concurrent processes described in the research methodology in Section 4.5, which follow the process of understanding the case study (process i) in Chapter 5.

8

Measures for Evaluating Elicitation Effectiveness and Efficiency

Three constructs are identified in Section 4.2 for assessing improvement in knowledge elicitation effectiveness. They are decision fidelity, state space and case quantity. In addition, a fourth construct – collection rate – is also identified for assessing improvement in knowledge elicitation efficiency. These are then used as the cornerstones for laying down the research propositions and specifying the hypotheses for testing in this thesis. In this chapter, they will be used as the basis for devising measures that are then computed to test the hypotheses subsequently.

To recall from Section 4.5, there is a series of activation, deactivation and reactivation between the process of devising measures and the next process of collecting data; the latter is described in the next chapter. This implies that whilst some measures are determined directly from the constructs' definitions before data collection (case quantity and collection rate), others are either improvised iteratively during data collection (decision fidelity), or derived after data collection (state space). In the following sections, the constructs will be revisited briefly, and the measures devised for evaluating them are described in more detail.

8.1 MEASURE FOR EVALUATING DECISION FIDELITY (CONSTRUCT ONE)

Decision fidelity is the first of three constructs identified for assessing knowledge elicitation effectiveness. It relates to the proximity to reality of the decisions elicited from experts. Hence, keeping in mind that an example case is made up of a decision element and an attribute element (Section 4.1), an example case is considered to have a high degree of decision fidelity if its decision element bears close resemblance to the decision that the expert would have made in a reality described by the corresponding attribute element.

The attempt to measure decision fidelity met two inherent difficulties. Firstly, the hot-test operations had no facility to record the actual decisions made by the experts in the real world, nor the attributes that described the hot-test operations when these decisions are made. As such, there was no basis that could be used to gauge the degree of decision fidelity in the example cases collected. Secondly, due to the experts' limited availability to participate in the experiment, the game models were only run for a simulated shift and an average of 56 example cases were collected from each knowledge elicitation session. Since the decision and attribute elements in each example case are very large (with an initial 21 and 551 variates respectively – these are described in more detail in Appendix B), the decision model that might be developed from the example cases collected in a session would not be robust. This implies that any evidence gained from using such a decision model to test the concept of decision fidelity will be weak. Consequently, it was concluded highly unlikely that a method for measuring decision fidelity definitively could be devised. Instead, it was deemed more sensible to devise a method that could indicate the presence of a certain degree of decision fidelity.

In the time spent to understand the hot-test operations and decision-making process, it was observed that each expert had a tendency to turn more switches (decisions) controlling one zone of the hot-test operations than other zones. This observation is not surprising since it is in line with a basic assumption shared by many theories of learning: experience shapes behaviour (Kowalski and Westen, 2005). Therefore, as the experts accrue their experiences on assigning untested engines to vacant hot-test cells (Section 5.2), they are also expected to form and consolidate certain switching behaviours. In light of this motion, it is thought if the proportion of switches turned in each zone in a knowledge elicitation session is close to the corresponding proportion observed in the real world, then this indicates that the real-world switching behaviour has been replicated in the elicitation session and suggests a certain degree of decision fidelity in the example cases collected.

Subsequently, four different zones of hot-test operations were determined. They, together with the panels of switches for controlling them, are identified in Figure 8.1 using a four-colour coding scheme. The zones are:

i. Zone 1 (Green) – Junction J;

ii. Zone 2 (Red) – Section of hot-test operations along Conveyor B;

iii. Zone 3 (Yellow) – Section of hot-test operations along Conveyor F; and

iv. Zone 4 (Blue) – Section of hot-test operations along Conveyor E.

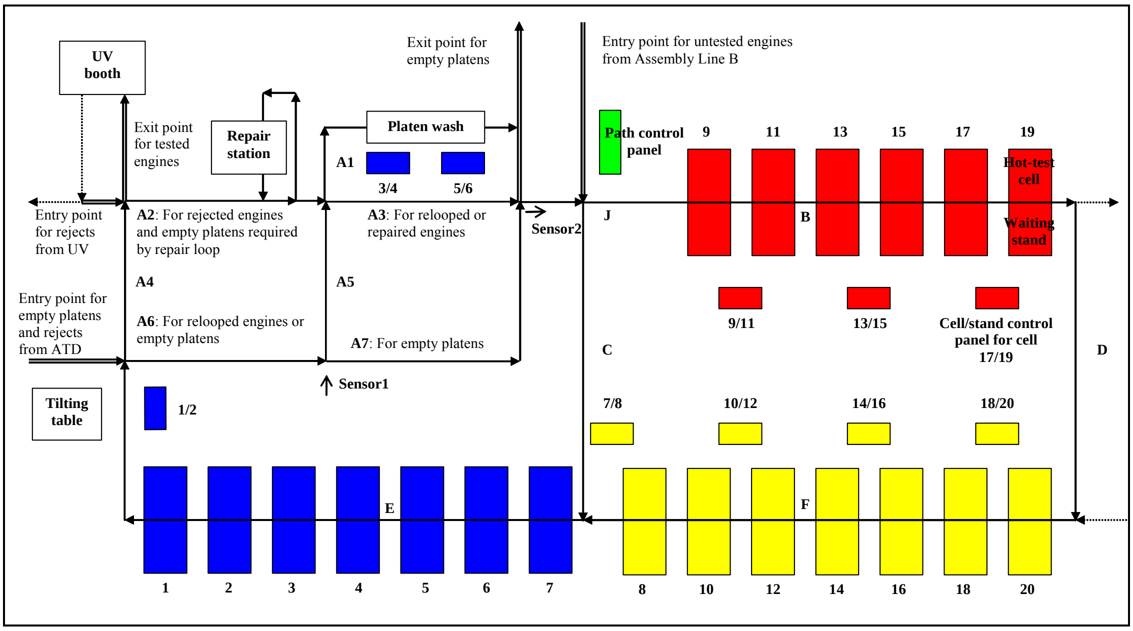

UV booth
Exit point for empty platens
Entry point for untested engines from Assembly Line B
Exit point for tested engines
Repair station
Platen wash
A1
3/4
5/6
Path control panel
9
11
13
15
17
19
Hot-test cell
Waiting stand
Entry point for rejects from UV
A2: For rejected engines and empty platens required by repair loop
A3: For relooped or repaired engines
Sensor2
J
B
A4
A5
Entry point for empty platens and rejects from ATD
A6: For relooped engines or empty platens
A7: For empty platens
9/11
13/15
Cell/stand control panel for cell 17/19
C
D
Sensor1
Tilting table
1/2
7/8
10/12
14/16
18/20
E
F
1
2
3
4
5
6
7
8
10
12
14
16
18
20

Figure 8.1: The four zones in the hot-test operations where decisions are made

As such, the measures for evaluating decision fidelity are two differently-sourced sets of quantities of switches turned in the four zones determined above. A set is determined from the decision element data collected in the knowledge elicitation sessions, whilst the other set is determined from the actual data collected in the real world. These two sets of quantities are then compared in the data analysis process that follows (Chapter 9).

8.2 MEASURE FOR EVALUATING STATE SPACE (CONSTRUCT TWO)

State space is the second of three constructs identified for assessing knowledge elicitation effectiveness. It relates to the adequacy of the range of situations from which the example cases are collected for training a knowledge base. At this juncture, it is helpful to reiterate that a situation is actually the state of the game model, which is defined by and recorded as a set of attribute values when an example case is being collected. Hence, if the example cases collected have k -number of values in their respective attribute elements, then from a spatial point of view, each of them will have an exact location in a k -dimensional space given by its set of attribute values. Thus, it can be followed that a set of example cases is considered to occupy a large state space if their attribute elements collectively cover a wide range of values for all attributes, so that the example cases are well-scattered in the k -dimensional space.

Therefore, the search for a measure to evaluate the state space occupied by a set of example cases will entail finding a way to evaluate their scatter in a k -dimensional space. A convenient place to begin the search is in the field of Descriptive Statistics. However, as prevalent descriptive statistics are formulated to describe quantitative univariate data only, there is a need to find reasonable alternative equivalents that can address a mixture of quantitative and qualitative multivariate data. This quest eventually led to ideas being borrowed from the fields of Geostatistics and Cluster Analysis. These are described in more detail in the following sections.

8.2.1 DESCRIPTIVE STATISTICS

Descriptive Statistics is a branch of statistics that includes any of the many techniques used to organise, summarise and present a set of data. Two major characteristics that data are normally summarised with are their central tendency and dispersion. The central tendency of a distribution is an estimate of the ‘centre’ of a distribution of values. There are three main measures of central tendency – mean, median and mode, with (arithmetic) mean being the most commonly used measure. However, central tendency does not necessarily provide enough information to describe data adequately, as two set of data with the same mean (a measure of central tendency) may have very dissimilar dispersions.

Dispersion refers to the scatter of data points around the central tendency. It is this concept of dispersion that is used for measuring the magnitude of a set of example cases' state space. There are four main measures of dispersion – range, inter-quartile range, standard deviation and its square, the variance. Out of these, the most commonly used measures of dispersion are standard deviation and variance, and this thesis will be using the former for measuring the scatter (or rather, the state space) of a set of example cases.

8.2.2 MEASURE OF DISPERSION FOR UNIVARIATE DATA: STANDARD DEVIATION

The sample standard deviation s for a sample of n univariate observations, comprising $x_1, x_2, \ldots, x_n$, is defined as the square root of the average squared deviation between the observations $i(i = 1,2,\ldots,n)$ and the sample mean $\bar{x}$:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Equation 1

A quick interpretation given by Field (2006) states a small standard deviation (relative to the value of the mean itself) indicates that many data points are close to the mean.

Conversely, a large standard deviation (relative to the mean) indicates that many data points are distant from the mean. A standard deviation of zero would mean that all the data points have the same value.

Unfortunately, the decisions and attributes in the example cases are represented by a mixture of counts, binary and nominal values (Section 7.1.2). This immediately poses a major challenge for using Equation 1 on the attribute element data collected, as it should be used only on univariate data that can be ranked meaningfully. Hence, there is a need to derive an alternative measure of dispersion that is a reasonable equivalent of the univariate standard deviation, which can be used on any (ranked or otherwise) multivariate data.

8.2.3 MEASURE OF DISPERSION FOR BIVARIATE DATA: STANDARD DISTANCE – A BACKGROUND

The next port of call was to look for other measures of dispersion that are used on data with a higher dimension. This attempt brought the search to the field of Geostatistics eventually, which has a measure that is commonly used to evaluate the dispersion of bivariate coordinates. In essence, this measure known as Bachi's standard distance s_d is used to summarise the spatial dispersion of locations (cases) around a fixed central location (mean centre). Rogerson (2006) interprets it as the square root of the average squared distance between the cases $i(i = 1,2,\ldots,n)$ and their mean centre c:

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} d_{ic}^2}{n}}$$

Equation 2

where d_{ic} represents the distance between case i and the mean centre c.

When the standard distance is first put forward by Bachi (1962), he assumes that the geographical distribution of an entire population over a territory is known. Each case of the population is indicated by an integer $i(i = 1,2,\ldots,n)$, and its location is given by a pair of values (x_i, y_i). Here, x_i and y_i are the coordinates of case i's place of residence, and are measured with regard to a system of orthogonal axes[7] such as the longitude and latitude.

Bachi (1962) proposes that a simple and intuitive measure of dispersion of the population over the territory can be obtained by averaging the distances d_{ij} between all possible pairs of cases i and j. In this way, if the population is widely dispersed, then the average distance D will be high, and *vice versa*. Whilst various formulae can be used to compute D, Bachi thinks it is most appropriate to do so by square-rooting the average squared distance between any pairs of cases for practical and theoretical reasons. If the calculation is extended to all n^2 distances, including those between each case and itself, the following expression is obtained:

[7] The orthogonal axes are assumed to be employed on a comparatively small territory that may be considered as a plane. This assumption ignores the curvature of the Earth's surface.

$$D = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^{2}}{n^{2}}}$$

Equation 3

which can be shown to lead to:

$$D = \sqrt{2\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

Equation 4

Since the coordinates of the mean centre c of the geographical distribution are $(\bar{x}, \bar{y})$, and that the Euclidean distance d_{ij} between two cases i and j is given by:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Therefore, Equation 4 can be re-expressed as:

$$D = \sqrt{2\frac{\sum_{i=1}^{n} d_{ic}^{2}}{n}}$$

Equation 5

where d_{ic} represents the Euclidean distance between case i and the mean centre c .

Thus, the square root of the average squared distance between all pairs of cases is simply the product of $\sqrt{2}$ and the square root of the average squared distance between the cases $i(i = 1,2,\ldots,n)$ and their mean centre c . Bachi (1962) then recognises the latter as the standard distance s_d (*i.e.* Equation 2), which he later promotes as a measure of dispersion for bivariate geographical coordinates. The standard distance appears to enjoy some important properties as a measure of dispersion, as well as some other distinct advantages like ease of calculation and algebraic manipulation. More importantly, it bears a close resemblance to a spatial version of the standard deviation (compare Equation 4 against Equation 1), which makes it conceptually appealing.

8.2.4 MEASURE OF DISPERSION FOR MULTIVARIATE DATA: STANDARD DISTANCE* – A PROOF

Whilst the measures for evaluating the dispersion of univariate and bivariate data can be found, the same cannot be said for multivariate data. As a result, Bachi's standard distance is extended below to find a reasonable alternative measure. To begin, consider a population of cases with k variates. Let each case be denoted by an integer $i(i = 1,2,\ldots,n)$, and its location in a k-dimensional Euclidean space (also known as a hyperspace for k > 3) be given by $(x_{i1}, x_{i2},\ldots,x_{ik})$. Using Bachi's argument in the last section, Equation 3 is reproduced below:

$$D^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2}{n^2}}$$

Equation 6

where * is added to denote that D is computed in a k-dimensional Euclidean space.

As the Euclidean distance d_{ij} between two cases i and j in a k-dimensional Euclidean space is given by:

$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ik} - x_{jk}\right)^2}$$

Equation 7

Then, substituting Equation 7 into Equation 6,

$$D^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ik} - x_{jk}\right)^2\right]}{n^2}}$$

$$D^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_{i1} - x_{j1}\right)^2}{n^2} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_{i2} - x_{j2}\right)^2}{n^2} + \ldots + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_{ik} - x_{jk}\right)^2}{n^2}}$$

Equation 8

By observation, the components under the square root of Equation 8 above have a generic form:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2}{n^2}$$

Expanding it becomes:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^2 - 2x_i x_j + x_j^2\right)}{n^2}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} 2x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n} x_j^2}{n^2}$$

$$= \frac{n\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j + n\sum_{j=1}^{n} x_j^2}{n^2}$$

$$= \frac{2n\sum_{i=1}^{n} x_i^2 - 2\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}$$

, as $\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$ and $\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{n} x_j^2$

$$= 2\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right]$$

$$= 2\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2\right]$$

$$= 2\sigma_{x_i}^2$$

, as the expression in the square brackets is an alternative form of the variance of x_i $\left(\sigma_{x_i}^2\right)$

$$= 2\left[\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}\right]$$

, where alternatively, $\sigma_{x_i}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$

Hence, continuing from Equation 8 above,

$$D^* = \sqrt{2\left[\frac{\sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2}{n} + \frac{\sum_{i=1}^{n} (x_{i2} - \overline{x}_2)^2}{n} + \ldots + \frac{\sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2}{n}\right]}$$

$$D^* = \sqrt{2\frac{\sum_{i=1}^{n} \left[(x_{i1} - \overline{x}_1)^2 + (x_{i2} - \overline{x}_2)^2 + \ldots + (x_{ik} - \overline{x}_k)^2\right]}{n}}$$

Equation 9

Since the location of the mean centre c in a k -dimensional Euclidean space is $(\overline{x}_1, \overline{x}_2,\ldots, \overline{x}_k)$, the Euclidean distance between case i and the mean centre c in a k -dimensional Euclidean space is given by:

$$d_{ic} = \sqrt{(x_{i1} - \overline{x}_1)^2 + (x_{i2} - \overline{x}_2)^2 + \ldots + (x_{ik} - \overline{x}_k)^2}$$

Hence, Equation 9 becomes:

$$D^* = \sqrt{2\frac{\sum_{i=1}^{n} d_{ic}^2}{n}}$$

Equation 10

As such, the square root of the average squared distance between all pair of cases (in a k -dimensional Euclidean space) can be expressed similarly as Equation 5, which is still simply $\sqrt{2}$ multiplied by the square root of the average squared distance between the cases $i(i = 1,2,\ldots,n)$ and their mean centre c . Like in Bachi (1962), the latter shall also be defined as the standard distance s_d^* (* is added to distinguish it from Bachi's standard distance for bivariate data), and advocated as a measure of dispersion for multivariate data. That is,

$$s_d^* = \sqrt{\frac{\sum_{i=1}^{n} d_{ic}^2}{n}}$$

Equation 11

As a further extension, the standard distance s_d^* can also be expressed in term of the Sum of Squared Distances (SSD) between all pairs of cases. It is formed by first substituting Equation 11 into Equation 10, and then equating it with Equation 6,

$$\sqrt{2} \cdot s_d^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2}{n^2}}$$

$$s_d^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2}{2n^2}}$$

Equation 12

That is,

$$s_d^* = \sqrt{\frac{SSD}{2n^2}}$$

Equation 13

However, in spite of finding a reasonable alternative measure of dispersion for multivariate data, another problem still exists: the Euclidean distance d_{ij} between two cases i and j (as defined in Equation 7) that is used to compute the numerator in Equation 13 can be used only on quantitative variates made up of discrete or continuous values (Krzanowski, 2005). It cannot be used on qualitative variates made up of binary or nominal values. More importantly, it also cannot be used on variates that have a mixture of data types, as in the case of the attribute element data collected in the experiment. Thus, there is a need to look for a more robust distance measure that can handle a mixture of quantitative and qualitative variates to substitute the Euclidean distance d_{ij} used in Equation 13.

8.2.5 DISTANCE MEASURE FOR MIXED DATA: GENERAL DISTANCE COEFFICIENT

Cluster Analysis (CA) is a branch of multivariate data analysis, which is predicated on computing the distances between observations. Briefly, Hair *et al.* (2006) explain that CA groups observations into clusters so that observations in the same cluster are more similar to one another than they are to observations in other clusters. This attempt is to maximise the homogeneity of observations within the clusters whilst also maximising the heterogeneity between the clusters. If the classification is successful, the observations within clusters will be close together when plotted geometrically, and different clusters will be far apart. This gives rise to the notion that dissimilarity is closely linked to the idea of distance: greater (smaller) distance indicates greater dissimilarity (similarity).

Krzanowski (2005) points out that various distance measures have been proposed in CA over the years for measuring dissimilarity between cases where the variates have the same type. A sample of these measures includes the Euclidean distance and the Minkowski metric for quantitative (discrete or continuous) data, and the Jaccard coefficient and the Czekanowski coefficient for binary data. On occasions where only similarity measures are available, then corresponding dissimilarity measures are obtained by using a monotonically decreasing transformation. As a similarity measure usually ranges from zero to one, with a value of one being most similar, a simple transformation is to subtract the similarity measure from a value of one. As an illustration, if x is a similarity measure, then '1 - x' will be the corresponding dissimilarity measure.

Like this research, CA also faces the problem of measuring dissimilarity between cases where different types of data are measured for each case. Typically, a case's measurements may consist of a mixture of numeric values, counts, rankings, binary attributes and categorical variates. Gower (1971) suggests a possible approach is to compute a separate dissimilarity value for each variate between any pair of cases, and then average these individual dissimilarity values to derive a final dissimilarity value for the two cases. He then uses this approach to define a general coefficient of similarity that can be applied to cases with different types of data. This is known as Gower's general similarity coefficient.

Later, Wishart (2001, 2006) adapts Gower's general similarity coefficient to set forth a general distance coefficient. It is essentially the converse of Gower's general similarity coefficient and is able to compute the distances between cases with mixed data types. Wishart defines the squared general Euclidean distance between any pair of cases i and j as follows (* is added to distinguish it from the Euclidean distance in Equation 7):

$$d_{ij}^{2*} = \frac{\sum_k w_{ijk} d_{ijk}^{2*}}{\sum_k w_{ijk}}$$

Equation 14

where d_{ijk}^{*} is a distance component for the k^{th} variate. Its determination is dependent on the scale that is used to measure the variate:

i. Nominal scale – $d_{ijk}^{2*} = 0$ if cases i and j have the same value for the variate (*i.e.* if $x_{ik} = x_{jk}$), and $d_{ijk}^{2*} = 1$ if they have different values (*i.e.* if $x_{ik} \neq x_{jk}$);

ii. Binary scale – $d_{ijk}^{2*} = 0$ if attribute k is present or absent in both cases i and j, and $d_{ijk}^{2*} = 1$ if attribute k is present in one case and absent in the other; and

iii. Ordinal, interval or ratio scale – d_{ijk}^{*} takes the value of $x_{ik} - x_{jk}$.

w_{ijk} is a binary variable that has a value of 1, if the k^{th} variate has a valid value in the context of the measurement scale used, in both cases i and j. Otherwise, w_{ijk} has a value of 0. For instance, if the k^{th} variate is measured on a binary scale and has a valid value of either 0 or 1 in cases i and j, then w_{ijk} = 1. However, if the k^{th} variate has an invalid value that is neither 0 nor 1 in at least one of the cases, then w_{ijk} = 0. At this point, it should be noted that the inclusion of w_{ijk} has the effect of averaging the squared general Euclidean distance over the number of valid variates, so that comparison of distances between different pairs of cases is fair.

Lastly, the Euclidean distance d_{ij}^{*} between cases i and j can be obtained by taking the square root of d_{ij}^{2*} in Equation 14.

8.2.6 STANDARD DISTANCE* FOR MIXED MULTIVARIATE DATA

A measure for evaluating the dispersion of mixed multivariate data is obtained by combining the distance measure for mixed data in Equation 14 with the dispersion measure for multivariate data in Equation 12. The standard distance per (square-rooted) valid variate for mixed, multivariate data thus formed is as expressed below:

$$s_d^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\sum_k w_{ijk} d_{ijk}^{2*}}{\sum_k w_{ijk}}\right)}{2n^2}}$$

Equation 15

However, as all variates in the attribute element data are valid, the inclusion of w_{ijk} is actually irrelevant. It is then removed from Equation 15 to give a simpler and more meaningful standard distance for mixed, multivariate data in Equation 16:

$$s_d^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k} d_{ijk}^{2*}}{2n^2}}$$

Equation 16

Finally, the expression in Equation 16 gives the measure for evaluating the state space occupied by a set of example cases. The numerator, which is actually a protracted form of SSD, can be computed easily using a CA software known as Clustan (Clustan, 2007).

8.3 MEASURE FOR EVALUATING CASE QUANTITY (CONSTRUCT THREE)

Case quantity is the last of three constructs identified for assessing knowledge elicitation effectiveness. It relates to the size of the set of example cases collected for training a knowledge base. Hence, case quantity refers to the total number of example cases recorded in a knowledge elicitation session. As such, a self-evident measure for evaluating case quantity is the number of example cases that are collected in a knowledge elicitation session.

8.4 MEASURE FOR EVALUATING COLLECTION RATE (CONSTRUCT FOUR)

Collection rate is the only construct identified for assessing knowledge elicitation efficiency. It relates to the expediency of the elicitation process in real time. In this thesis, a unit of real time is taken to be one minute. Hence, an obvious measure for evaluating collection rate is the average number of example cases recorded over an elicitation session.

8.5 CONCLUSION

This chapter describes the measures for evaluating the four constructs identified in Section 4.2. They are decision fidelity, state space, case quantity and collection rate. Whilst the measures for case quantity and collection rate were determined before the data collection process, the measures for decision fidelity and state space were determined during and after the data collection process respectively. In short, the process of devising measures began before the experiment, and ended after it.

To summarise the work carried out so far, an understanding of the hot-test operations and its environment was first gained by using an array of complementary techniques (process i in Figure 4.1 – Chapter 5). The information gathered from this was then used to design the experiment to be carried out later (process ii – Chapter 6), build and assess the game models to be used in the experiment (processes iii and iv – Chapter 7), and lastly devise the measures in this chapter (process v). In the next chapter, the outcomes from these earlier chapters will be pieced together to carry out the experiment (process vi). Then, the data that are collected in the experiment will be used to compute the measures for subsequent analysis (process vii).

9 Data Collection and Analysis (Hypothesis One)

The foundation for the data collection and analysis processes is laid with the completion of the activities described in earlier chapters. Next, the outcomes from these chapters will be put together to carry out the experiment, collect data and test the hypotheses in Section 4.4. In a nutshell, the experimental design and knowledge elicitation schedule that are established in Chapter 6 will be implemented with the help of eight experts identified in Chapter 5 and six game models built in Chapter 7. The data collected in the ensuing knowledge elicitation sessions are then used to compute the measures devised in Chapter 8. Following this, the measures determined for testing Hypothesis 1 are analysed, whilst the rest will be used to analyse Hypothesis 2 to 6 in the next chapter. The work performed and its findings are described in more detail below.

9.1 THE EXPERIMENT

To recap, eight experts are identified in Chapter 5, a repeated measures design is established in Chapter 6 and six game models are built in Chapter 7 for the experiment. Under the repeated measures experimental design, each expert was exposed to all conditions of the experiment, where an experimental condition is defined as a unique combination of the factors that are being investigated. These experimental conditions are reproduced from Section 6.1 as below:

I. 2D representation with unadjusted parameters;

II. 2D representation with adjusted parameters;

III. 2½D representation with unadjusted parameters;

IV. 2½D representation with adjusted parameters;

V. 3D representation with unadjusted parameters; and

VI. 3D representation with adjusted parameters.

Since there are six experimental conditions, a complete experiment trial with each expert comprises six knowledge elicitation sessions; an elicitation session was carried out for each experimental condition, with the game model that was adapted for that condition. In order to standardise the way that the experiment was conducted, a set of procedures was developed and followed to carry out every knowledge elicitation session. Notwithstanding all the planning that was undertaken, the elicitation schedule still took a significant time to complete. These issues are discussed further below.

9.1.1 STANDARD PROCEDURES

Kowalski and Westen (2005) suggest that a set of standard procedures should be followed to carry out an experiment, so that the only things that vary from expert to expert are the experimental conditions and the experts' behaviour in response to them. In so doing, these procedures will help to maximise the likelihood that any differences observed in the experts' behaviour can be attributed to the experimental manipulation, allowing the experiment to draw inferences about cause and effect more conclusively. Subsequently, a set of standard procedures was developed and followed in each elicitation session, which addressed issues on the random number streams used, the game model's warm-up period and run speed, and the instructions for playing a game model.

Firstly, an identical set of random number streams was used in each game model. This is to ensure that the same sequence of events was recreated for every model entity that used random numbers. However, as each game model's run was subject to the player's *ad hoc* interventions, these events would occur at different times in the simulation. For instance, the event that a particular hot-test cell breaks down after testing a certain number of engines will be replicated in all elicitation sessions. However, the breakdown will occur at a different simulated time in each session.

Secondly, the game model was warmed up for a period of a shift (*i.e.* 480 minutes) before the knowledge elicitation session began. This is to emulate the real-world conditions in the hot-test operations at the start of a shift.

Thirdly, the run speed of the game model was also set to a pre-determined level (Section 7.2.2), by using a fixed time scale factor. The latter is a WITNESS function that adjusts the relationship between real time units and simulation time units. In this case, the time scale factor used actually slowed down the game model such that the expert would have sufficient time to appreciate the information presented in the game model and respond appropriately without extending the model run-time excessively.

Lastly, the expert participating in the knowledge elicitation session was given a brief refresher on how to play the game model before it began. This is necessary because a long period of time might have elapsed between the expert's current and last elicitation sessions, and he might need to re-familiarise himself with the model's functions.

9.1.2 LET THE GAMES BEGIN!

The game model was handed over to the expert after the standard procedures described above were performed. This would officially mark the beginning of a knowledge elicitation session. Then he was asked to play with the game model for a simulated shift, in which his interactions with the game model would be recorded as example cases (Section 7.1.2). Also, the actual time taken by him to complete the elicitation session was recorded manually. The actual times taken by the experts to complete all the sessions are summarised in Table 9.1.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	30	37	27	24	73	93
B	50	37	35	33	96	97
C	45	30	34	33	64	87
D	40	40	29	27	76	82
E	20	25	22	28	65	80
F	45	35	36	34	78	88
G	50	45	30	32	79	90
H	34	44	36	29	67	95

Table 9.1: A summary of collection times

It is known that management generally resists the notion of losing productivity as a result of the experts absenting themselves to participate in an experiment (Section 6.3). This consequently restricted the plan and initial schedule for carrying out the 48 knowledge elicitation sessions (eight experiment trials of six knowledge elicitation sessions each). Unfortunately, the immense difficulty in getting the management to release the experts that were required to participate in the elicitation sessions for each week was not fully anticipated. This is in spite of the fact that an elicitation session that involved either the 2D or the 2½D game model required only 34 minutes on average. Hence, the initial (static) schedule was turned into a rolling one, and was revised on a weekly basis.

Subsequently, it seemed very unlikely that the management would release the experts to participate in elicitation sessions that involved the 3D game models, which would require 82 minutes on average on a current state of the art computer notebook. Thus, after much deliberation, plans were made to request the experts to participate in these elicitation sessions outside their shifts. Fortunately, as a strong rapport was already established with the experts from the beginning of the investigation (Section 5.1) and maintained throughout, they were happy to oblige. Some even offered to do so *pro bono*. Nevertheless, they were all compensated at a rate of £10 per hour for their help in the end.

All 48 knowledge elicitation sessions were carried out over an extended period of 19 weeks. The actual timeline for the experiment is reproduced from Section 6.3 in Table 9.2, where I to VI denote the six experimental conditions that the elicitation sessions were carried out under. It can be seen from Table 9.2 that there were several periods of lull time between successive elicitation sessions. These were used meaningfully to shadow the experts whilst they were working, and record any decisions made by them in the meantime. The real-world data thus collected were used later for testing Hypothesis 1.

Subject	Week																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A					I					II					III	IV			V VI
B		I II		III	IV			V	VI										
C	I		II	III						IV			V			VI			
D	I		II	III				IV					V	VI					
E	I			II				III		IV				V			VI		
F			I II		III				IV							V		VI	
G									I			II			III IV		V	VI	
H										I			II			III IV		V	VI

I to VI denote the six conditions (Section 9.1) in the experiment

Table 9.2: The actual knowledge elicitation timeline for the experiment

9.2 ANALYSIS FOR HYPOTHESIS ONE: DECISION FIDELITY & VISUAL REPRESENTATION DIMENSION

The overarching pair of null and alternative hypotheses of interest are replicated below:

$H_{1(0)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{1(a)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session improves as a higher dimension of visual representation is used.

Decision fidelity has been defined, in Section 4.2.1, as the resemblance that a decision element of an example case bears to the decision that the expert would have made in a reality described by the corresponding attribute element. It is also explained in Section 8.1 that two sets of quantities of switches turned in the four zones identified in Figure

8.1 will be used to evaluate decision fidelity. The latter is reproduced in Figure 9.1 and the zones are:

i. Zone 1 (Green) – Junction J;

ii. Zone 2 (Red) – Section of hot-test operations along Conveyor B;

iii. Zone 3 (Yellow) – Section of hot-test operations along Conveyor F; and

iv. Zone 4 (Blue) – Section of hot-test operations along Conveyor E.

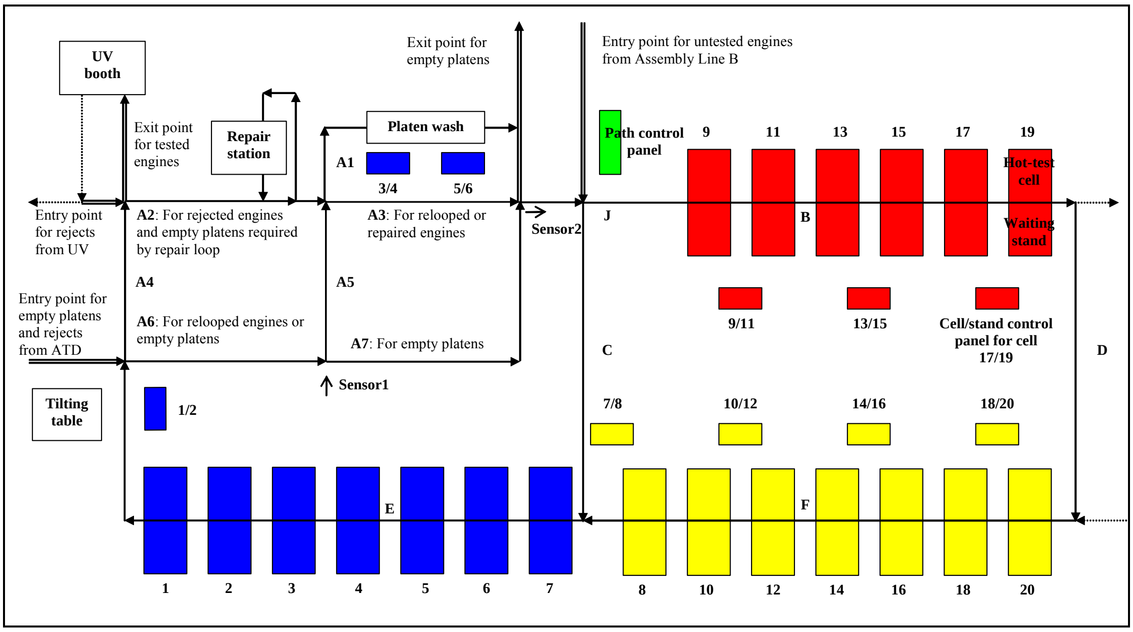

UV booth
Exit point for tested engines
Repair station
Exit point for empty platens
Platen wash
Entry point for untested engines from Assembly Line B
Path control panel
A1
3/4
5/6
9
11
13
15
17
19
Hot-test cell
Waiting stand
Entry point for rejects from UV
A2: For rejected engines and empty platens required by repair loop
A3: For relooped or repaired engines
Sensor2
J
B
A4
A5
Entry point for empty platens and rejects from ATD
A6: For relooped engines or empty platens
A7: For empty platens
9/11
13/15
Cell/stand control panel for cell 17/19
C
D
Sensor1
Tilting table
1/2
7/8
10/12
14/16
18/20
E
F
1
2
3
4
5
6
7
8
10
12
14
16
18
20

Figure 9.1: The four zones in the hot-test operations where decisions are made

The first set of quantities was determined from the decision element of the example cases collected in the knowledge elicitation sessions, whilst the second set was determined from the real-world data collected by shadowing the experts (Section 9.1.2). It was thought that if the proportions of switches turned in the four zones in an elicitation session are close to the corresponding proportions that are observed in the real world, then this might indicate that a certain degree of decision fidelity is present in the example cases collected. It was further thought that if this phenomenon was observed repeatedly in elicitation sessions supported by game models that used a specific visual representation dimension, then these observations would constitute some evidence towards testing Hypothesis 1.

The data that were used for determining the first set of quantities are from knowledge elicitation sessions supported by game models preset with the unadjusted set of model parameters only. This includes elicitation sessions carried out under experimental conditions I (2D representation with unadjusted parameters), III (2½D representation with unadjusted parameters) and V (3D representation with unadjusted parameters). The data from elicitation sessions supported by game models preset with the adjusted set of model parameters are not used here, as the game models would produce more unlikely scenarios that are most probably not similar to those observed in reality, thus eroding the basis for comparison.

The data that were used for determining the second set of quantities were collected manually by shadowing the experts whilst they were working. Although there are eight experts who are qualified to participate in the experiment, only five of them were working as switch operators at the time of data collection. They are Subject A, B, C, G and H. This limited any real-world data collection efforts to revolve around these five experts. The real-world data collected took the form of a sequence of switches that were turned whilst the experts were being observed. Each of them was shadowed for different periods of time over several work-shifts in order to avoid bias in the data. They were eventually shadowed for 842 minutes on average (approximately 1.98 work-shifts). The actual times spent on shadowing the experts are summarised in Table 9.3.

Subject	Total shadow time (minutes)
A	811
B	260
C	1,309
G	1,341
H	490

Table 9.3: A summary of total time spent to shadow Subject A, B, C, G and H

Altogether, there are 15 sets of quantities that were determined from the knowledge elicitation sessions carried out under experimental condition I, III and V for Subject A, B, C, G and H. Also, five sets of quantities were determined from shadowing the experts. These are summarised in Table 9.4, Table 9.5, Table 9.6, Table 9.7 and Table 9.8. Next, these quantities are explored and used to test Hypothesis 1.

Zone	Game model with Unadjusted model parameters			Real-world
	2D	2½D	3D	
1	17	13	17	103
2	54	36	110	301
3	36	21	34	204
4	1	1	18	153

Table 9.4: A summary of quantities of switches turned by Subject A

Zone	Game model with Unadjusted model parameters			Real-world
	2D	2½D	3D	
1	24	20	27	54
2	154	122	204	78
3	148	126	188	86
4	66	28	110	80

Table 9.5: A summary of quantities of switches turned by Subject B

Zone	Game model with Unadjusted model parameters			Real-world
	2D	2½D	3D	
1	6	10	16	154
2	88	50	120	390
3	46	6	24	254
4	20	20	36	204

Table 9.6: A summary of quantities of switches turned by Subject C

Zone	Game model with Unadjusted model parameters			Real-world
	2D	2½D	3D	
1	22	17	24	162
2	60	58	114	617
3	52	4	20	546
4	25	4	12	335

Table 9.7: A summary of quantities of switches turned by Subject G

Zone	Game model with Unadjusted model parameters			Real-world
	2D	2½D	3D	
1	11	14	16	70
2	28	25	40	195
3	12	1	10	159
4	1	2	2	102

Table 9.8: A summary of quantities of switches turned by Subject H

9.2.1 DATA EXPLORATION

The quantities determined from data that were collected whilst shadowing Subject A, B, C, G and H are first converted into proportions and then compared with those determined from the knowledge elicitation sessions in Figure 9.2. For ease of reference, the colours used in the graphs follow those used to differentiate the zones in Figure 9.1. As an illustration, the proportion of switches turned in blue-coloured Zone 4 in Figure 9.1 is also coded in blue in the graphs.

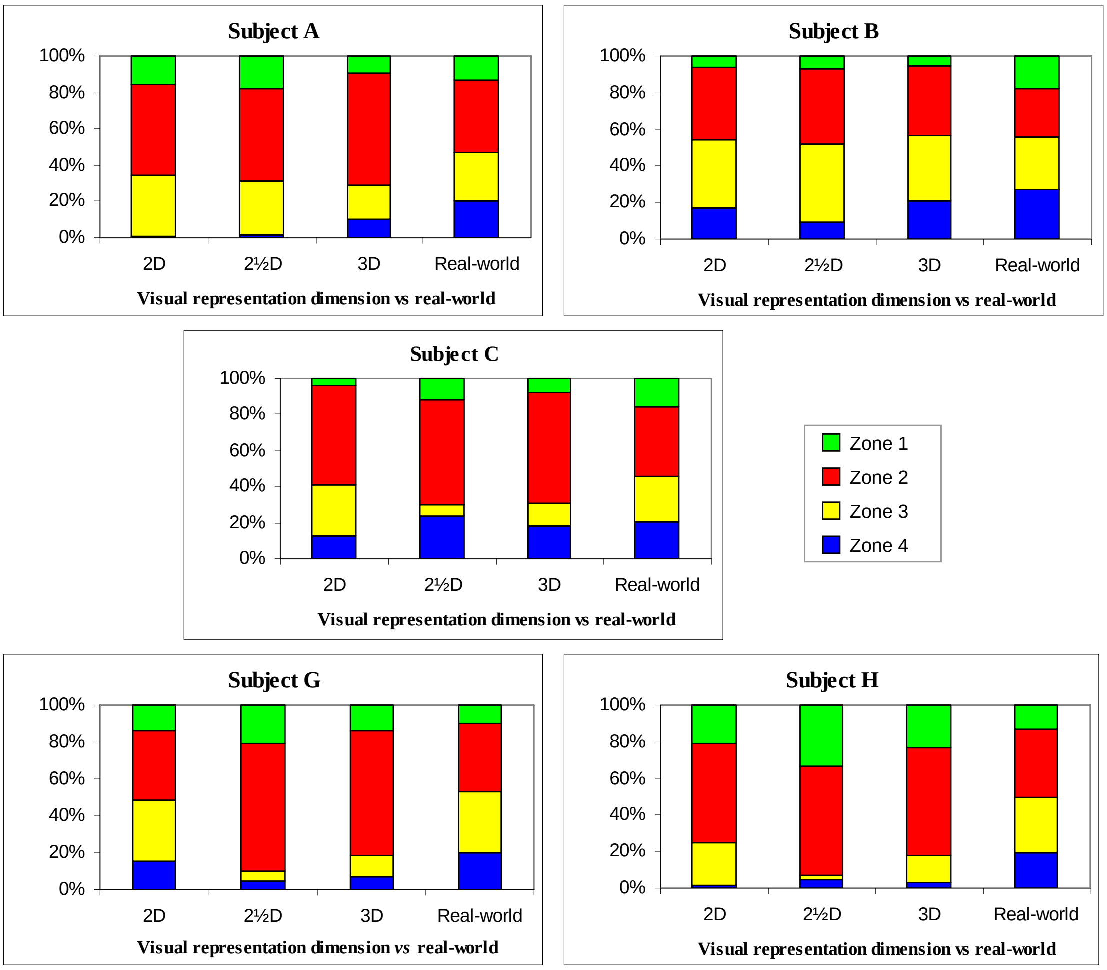


Figure 9.2: A comparison of proportions of switches turned by Subject A, B, C, G and H

It can be observed that the proportions of switches turned in the four zones by all five experts in reality are quite similar. It is also immediately apparent for Subject G that his elicitation session supported by a 2D game model seemed to produce proportions that were very similar to those from the real world. However, the same cannot be said for Subject A, B, C and H, as it is not obvious whether there was an elicitation session that produced proportions similar to those from the real world.

9.2.2 HYPOTHESIS TESTING

The measures need to be tested statistically to support the observations made above. Here, a non-parametric chi-squared $\left(\chi^2\right)$ goodness-of-fit test is used to compare the quantities of switches turned in the four zones during a knowledge elicitation session, against the expected quantities computed using the real-world proportions. As this test was executed for each elicitation session carried out under either experimental condition I, III or V for the five experts, hence a total of 15 χ^2 tests were executed subsequently. An appropriate pair of null and alternative hypotheses for each χ^2 test executed are given as follows:

$H_{S,G(0)}$: The proportions of switches turned in the four zones in the elicitation session are similar to those of the real world;

$H_{S,G(a)}$: The proportions of switches turned in the four zones in the elicitation session are not similar to those of the real world.

where S denotes Subject A, B, C, G or H, and G denotes 2D, 2½D or 3D game model with unadjusted model parameters.

On the one hand, if the χ^2 tests consistently show that the quantities of switches turned in elicitation sessions supported by game models using a higher visual representation dimension are more similar to the expected quantities, then their results would construe sufficient evidence in support of the alternative hypothesis ($H_{1(a)}$). On the other hand, if the χ^2 tests show otherwise, then their results would construe sufficient evidence in support of the null hypothesis ($H_{1(0)}$). The test statistics computed for all 15 χ^2 tests are summarised in Table 9.9.

Subject	Game model with Unadjusted model parameters		
	2D	2½D	3D
A	24.82	16.01	36.93
B	82.33	99.28	95.76
C	30.28	20.71	45.30
G	4.31*	63.09	81.69
H	15.43	33.64	29.09

* Statistically insignificant at 5%

Table 9.9: A summary of χ^2 test statistics

It is noted that all knowledge elicitation sessions, except for the one participated by Subject G using a 2D game model, have test statistics whose p-values are effectively zero. Thus, in relation to each of these elicitation sessions, there is sufficient evidence to reject the null hypothesis ($H_{S,G(0)}$) at a 5% level of significance, and it is concluded that the proportions of switches turned in the four zones in the elicitation session are not similar to those of the real world. However, as the test statistic's p-value for the elicitation session participated by Subject G using a 2D game model is greater than 0.05, there is insufficient evidence to reject the null hypothesis ($H_{Subject\,G,2D(0)}$) at a 5% level of significance, and it is concluded that the proportions of switches turned in the four zones in this elicitation session are similar to those of the real world.

Hence, whilst it is clear that a 2D representation is most effective in eliciting example cases with a high degree of decision fidelity from Subject G, such strong conclusions cannot be drawn for the other four experts. Nonetheless, even if all test statistics are statistically significant, their magnitude might still provide some clues with regard to which visual representation dimension is relatively more effective in eliciting example cases with a certain degree of decision fidelity. Since a smaller test statistic suggests that the proportions of switches turned in the four zones in the knowledge elicitation session are closer to those of the real world, a 2D representation appears to be relatively more effective for Subject B and H, and a 2½D representation is relatively more effective for Subject A and C. Also, even though it cannot be concluded whether a 2D or 2½D representation is more effective in eliciting example cases with a certain degree of decision fidelity, there is some corroborating evidence that signals the 3D representation to be the least effective.

9.2.3 Summary

The overarching pair of null and alternative hypotheses of interest ($H_{1(0)}$ and $H_{1(a)}$) were assessed through testing 15 pairs of null and alternative hypotheses ($H_{S,G(0)}$ and $H_{S,G(a)}$) on the knowledge elicitation sessions carried out under either experimental condition I, III or V for Subject A, B, C, G and H. Each of these hypothesis tests was executed by comparing the quantities of switches turned in the four zones of the hot-test operations during a knowledge elicitation session, against the expected quantities computed using the real-world proportions.

Subsequently, 14 out of the 15 hypothesis tests had their null hypotheses ($H_{S,G(0)}$) rejected at a 5% level of significance. Hence, it is concluded that the proportions of switches turned in the four zones in these 14 elicitation sessions are not similar to those of the real world. Since these results indicate that the degree of decision fidelity is low in the example cases collected across 14 knowledge elicitation sessions, and remains low regardless of the visual representation dimension used in the game models, they appear to construe strong, albeit negative evidence to support the overarching null hypothesis ($H_{1(0)}$).

Moreover, the results go further to present some weak evidence suggesting that the degree of decision fidelity in the example cases collected in a knowledge elicitation session improves as a lower dimension of visual representation is used. This evidence essentially conspires to contradict Proposition 1 (Section 4.3.1), which is used to form Hypothesis 1. In light of this, Proposition 1 might consider to be revised as follows:

> *A lower dimension of iconic representation would demonstrably improve, if not maintain, the degree of decision fidelity in the example cases collected in a knowledge elicitation session.*

9.3 CONCLUSION

This chapter describes the set of standard procedures that was executed before each knowledge elicitation session was carried out. They pertain to the random number streams used, the game model's warm-up period and run speed, and the instructions for playing a game model. In addition, the difficulty encountered during the data collection process and its resolution are also elaborated.

Following the end of the data collection process, the example cases collected were analysed. In particular, Hypothesis 1, which postulates a causal link between visual representation dimension (cause) and decision fidelity (effect), was tested. A series of

χ^2 tests were executed subsequently, whose results support the null hypothesis ($H_{1(0)}$). Furthermore, the results present some weak evidence that suggests an inverse relationship between visual representation dimension and decision fidelity.

In the next chapter, the data analysis process (process vii in Figure 4.1) continues with the testing of the remaining hypotheses (Hypothesis 2 to 6).

10

Data Analysis (Hypothesis Two to Six)

This chapter continues with the data analysis process started in Chapter 9. As with Hypothesis 1 in the last chapter, the measures determined for Hypothesis 2 to 6 are analysed here. The analysis that was carried out subsequently followed an established parametric framework for analysing data collected from repeated measures experiments. The framework is outlined briefly below, and described in more detail in Appendix B. In addition, the work performed to test the hypotheses and its findings are also discussed. In order to minimise any unnecessary duplication, these are organised according to the constructs (Section 4.2) that the hypotheses are founded on. They are state space (Hypothesis 4), case quantity (Hypothesis 2 and 5) and finally, collection rate (Hypothesis 3 and 6).

10.1 AN OVERVIEW OF THE ANALYTICAL FRAMEWORK

In general, Kowalski and Westen (2005) suggest that an analysis should use both descriptive and inferential statistics. On the one hand, descriptive statistics are used to summarise the data's essential features, which can also be depicted using appropriate line graphs. Together, these constitute a preliminary exploration of the data to gain a quick appreciation of any trend or pattern underlying them. On the other hand, inferential statistics are used to yield tests of statistical significance, which are then used to ascertain whether the observations made above are meaningful. Hence, in so doing, the hypotheses set out in this thesis are tested.

As a repeated measures experimental design was used to collect the data for investigating two factors, the most apt inferential statistical test for assessing their effects is the two-way repeated measures ANalysis Of VAriance (ANOVA). However, since ANOVA is a parametric test, the data need to show that they meet the criterion for normality before ANOVA can be applied on them. The tests for this criterion include the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D) and skewness tests. If the data initially failed to show that they meet the criterion for normality, then Field (2006) suggests executing a logarithmic transformation on them before putting them through another cycle of the K-S, A-D and skewness tests. Moreover, the data also need to be assessed for the criterion of sphericity. In this case, the test for this criterion is Mauchly's test. If the data failed to show that they meet the criterion for sphericity, then Field (2006) remarks that it is necessary to revise the critical values that are used for assessing the test statistics from the following ANOVA with a Greenhouse-Geisser correction.

In the event that the ANOVA's results are able to conclude a significant effect in a factor, a series of planned or post-hoc/pairwise comparisons are carried out to determine the specifics of the factor's effect. On the one hand, if the alternative hypothesis is specified to test a *a priori prediction* about the data and the preliminary data exploration that is completed earlier supports the trend/pattern described in the prediction, then planned comparisons are executed. Otherwise, it is more meaningful to perform post-hoc/pairwise comparisons to investigate the factor's effect. On the other hand, if the alternative hypothesis is specified to explore the data for *any differences* due to treatment levels in a factor, then post-hoc comparisons are executed. Finally, after the statistically significant effects are identified, their materiality and importance are determined by computing their sizes.

10.2 ANALYSIS FOR HYPOTHESIS FOUR: STATE SPACE & MODEL PARAMETERS

The overarching pair of null and alternative hypotheses of interest are replicated below:

$H_{4(0)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session is not affected by the model parameters used;

$H_{4(a)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

State space has been defined, in Section 4.2.2, as the coverage collectively made by the attribute elements of a set of example cases in a hyperspace defined by the ranges of values possible for all attributes. It is first mentioned in Section 8.2 that each example case's attribute element describes the state of the game model when an expert interacts with it. It is initially made up of 551 variates (Appendix B), which are measured using either binary, nominal or ratio scales. However, many of these variates are actually redundant and should be removed. In addition, many variates are actually components of various attributes or cover large ranges of values. Therefore, there is also a need to recode or rescale these variates respectively. Finally, the attribute element data are screened for outliers, before they are used to compute the measures for evaluating state space.

10.2.1 DATA PREPARATION

Remove rogue and redundant data

To begin, the values from all 48 sets of attribute element data collected (from eight trials of six knowledge elicitation sessions each) were compiled and profiled. The profiles, which informed on the highest and lowest values in each variate, can then be used for the following two purposes.

Firstly, the profiles can be used for identifying rogue attribute elements, which may arise from recording errors. As the binary and nominal variates have known upper and lower limits, the highest and lowest values in their profiles are expected to fall within these limits. Therefore, if a variate's profile overlaps its known limits, then the attribute elements that contained the offending values are classified as rogue observations and will not be used in any subsequent analysis. For instance, a binary variate is expected to have either a '0' or '1' value. If a value other than '0' or '1' is discovered in the binary variate's profile, then the attribute element that has the offending value is removed from further analysis. Nonetheless, since the collected example cases were not recorded manually, rogue attribute elements are not expected. Subsequently, no rogue data were detected.

Secondly, the profiles can also be and were used for identifying redundant variates, which were created mistakenly without realising that the information they were supposed to collect did not exist in the first place. If a variate's highest and lowest values are found to be zero, then it can be safely assumed to be redundant and will not be used in any subsequent analysis. 110 redundant variates were found and discarded subsequently, leaving behind 441 useful variates.

Recode data

It is mentioned earlier that many variates are actually components of various attributes, where related binary component-variates are organised and interpreted together to provide complete information on the attributes. Hence, the number of variates in each attribute element can be reduced to a more manageable size by combining these binary component-variates appropriately to form nominal composite-variates without losing any vital information.

As an illustration, assume that a pair of binary component-variates (c_1, c_2) are originally used to indicate the presence (yes or no) and type of engine (2*l* or 2.4*l*) on a conveyor section. If there is a 2*l* or 2.4*l* engine on the conveyor section, then the binary variates will appear as $(1,0)$ or $(0,1)$ respectively. However, if there is no engine on the conveyor section, then the variates will appear as $(0,0)$. Notwithstanding, these variates can be combined to form a nominal composite-variate (c) where the latter takes the value of 0,1 or 2 to represent no engine, a 2*l* engine, or a 2.4*l* engine respectively.

In this way, the remaining 441 cleaned variates were reduced further to 184 variates.

Rescale data

It is determined in Section 8.2 that a standard distance measure will be used to evaluate state space. In essence, the standard distance measure is predicated on computing the distances between pairs of example cases, which in turn is based on computing the distances between corresponding variates (distance components) from the pairs of example cases. Krzanowski (2005) warns that if some variates exhibit a much greater range of values than the others, then they are likely to create larger distance components. These will then go on to dominate and bias the distances computed between example cases, and *ergo* the standard distance. The converse is also true.

Fortunately, Wishart (2001, 2006) has defined the distance components based on binary and nominal variates in such a manner that they have a unit range (Section 8.2.5). That is, they range from zero to one. However, the same cannot be said for the distance component based on ordinal, interval or ratio variates; its value is simply the difference between the pair of ordinal, interval or ratio variates. Hence, there is only a need to rescale ordinal, interval or ratio variates.

In short, whilst the cleaned and recoded data include binary, nominal and ratio variates with varying ranges of values, only the latter need to be rescaled. In order not to bias the distances that are computed subsequently, the ratio variates should be scaled such that the distance components based on them will share the same unit range as those based on the binary or nominal variates.

Consider an observation x with a ratio k^{th} variate. Therefore, a variate, x_k, can be scaled down by performing the following operation:

$$\frac{x_k - x_{k,\min}}{x_{k,\max} - x_{k,\min}}$$

where $x_{k,\max}$ and $x_{k,\min}$ are the respective empirical maximum and minimum values of x_k across all 48 sets of attribute element data. These values are available from the profiling carried out earlier.

Accordingly, a scaled down x_k with a value of 0 or 1 implies that its original value is the empirical minimum or maximum value respectively. In this way, a distance component based on a pair of rescaled ratio variates is made to have a unit range.

Remove outlying data

Finally, the cleaned, recoded and rescaled data were inspected for outliers using a series of Andrews plots. Likewise, consider an observation x with k variates. The Andrews curve that corresponds to a typical observation, $(x_1, x_2, \ldots, x_k)$, can be obtained by computing the following function and plotting it over the range $-\pi < t < \pi$:

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots$$

Therefore, a set of observations will appear as a set of curves drawn across the plot. Subsequently, an Andrews plot was drawn for each set of attribute element data, and inspected for outliers. An example of the Andrew plots drawn for the data collected from Subject B is displayed in Figure 10.1; each line in the plot represents the attribute element of an example case. The data used for the plot are collected from the knowledge elicitation session using the 2D representation with adjusted parameters.

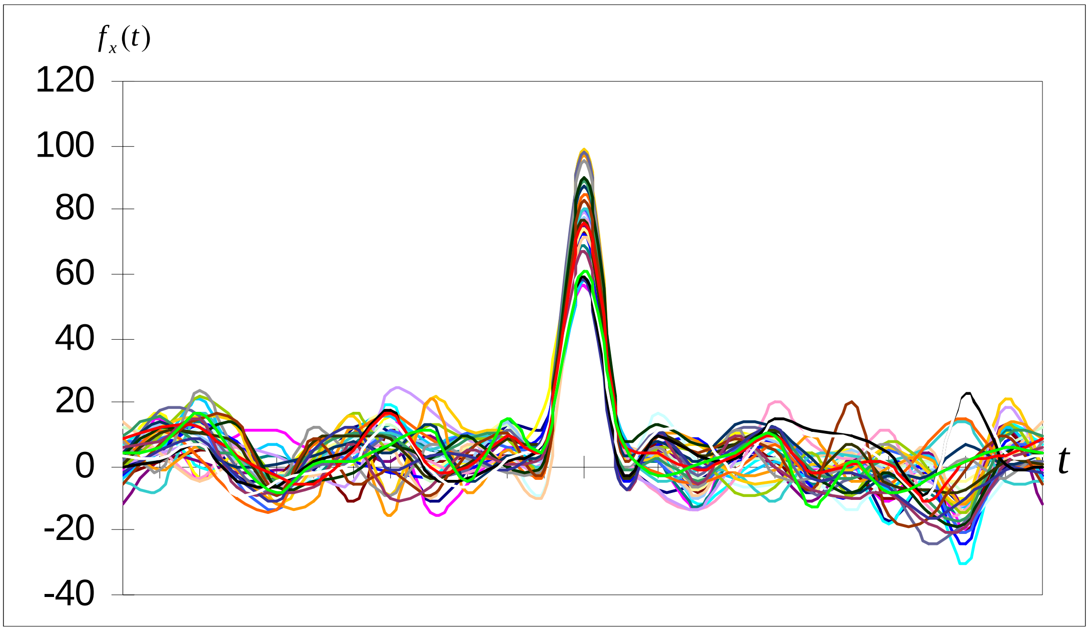


Figure 10.1: An example of the Andrews plots drawn for Subject B (2D representation with adjusted parameters)

Krzanowski (2005) advises that these Andrew plots have properties which make them suitable for detecting outliers in the data. These properties will cause two observations with similar sets of variate values to be represented by curves that are close together. Conversely, two observations with different sets of variate values will be represented by curves that differ markedly in at least some parts of the curves. Hence, if there is a curve in an Andrews plot that behaves very differently from the rest, then the observation represented by the curve will be regarded as an outlier.

Using Figure 10.1 for illustration, except for the spike in the middle, the Andrews curves do not appear to show a congruent pattern. Nonetheless, since there is no curve that appears to deviate significantly from the others, it is concluded that there is no obvious outlier in the data used to draw the Andrews plot.

Eventually, only a few apparent outliers were detected among the 48 sets of attribute element data, and these were deleted from subsequent analysis.

10.2.2 STATE SPACE MEASURE COMPUTATION

A suitable measure for evaluating the state space of a set of example cases is found in the standard distance for mixed, multivariate data (Equation 16, Section 8.2.6):

$$s_d^* = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k} d_{ijk}^{2*}}{2n^2}}$$

Equation 16

The numerator, also known as the SSD, can be computed easily using Clustan (a Cluster Analysis software) on the prepared data. The denominator, n , is simply the case quantity, which is defined as the total number of example cases recorded in a knowledge elicitation session (Section 4.2.3). The computed SSD values and case quantities for the entire experiment are summarised in Table 10.1 and Table 10.2 respectively.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	84,996	203,537	38,813	32,433	241,920	664,667
B	1,167,679	80,932	682,295	147,147	2,584,074	1,137,512
C	167,742	79,958	81,852	56,231	242,875	525,234
D	184,507	390,663	78,699	46,316	466,446	196,604
E	13,775	19,629	5,836	13,556	17,414	38,551
F	581,905	72,448	57,841	45,206	98,853	431,166
G	294,330	332,116	68,378	150,762	159,426	274,732
H	30,915	27,342	21,377	17,618	85,929	369,860

Table 10.1: A summary of computed SSD values $\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k} d_{ijk}^{2*}\right)$

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	42	55	29	22	70	97
B	161	35	122	47	240	132
C	59	34	42	29	71	89
D	63	76	40	26	98	53
E	17	17	11	14	19	24
F	110	32	35	25	44	79
G	79	70	37	48	55	63
H	25	20	21	16	41	73

Table 10.2: A summary of case quantities (n)

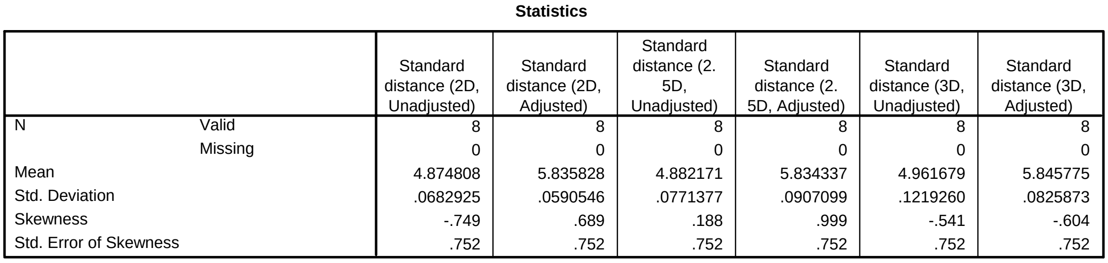
Putting the values from Table 10.1 and Table 10.2 into Equation 16, the standard distances for the 48 sets of attribute element data are computed and summarised in Table 10.3. Next, these values are explored and used to test Hypothesis 4.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	4.908	5.800	4.804	5.788	4.968	5.943
B	4.746	5.747	4.788	5.771	4.736	5.713
C	4.909	5.881	4.817	5.782	4.908	5.758
D	4.821	5.815	4.959	5.853	4.928	5.916
E	4.882	5.828	4.911	5.881	4.911	5.785
F	4.904	5.948	4.859	6.014	5.053	5.877
G	4.856	5.821	4.997	5.720	5.133	5.883
H	4.973	5.846	4.923	5.866	5.056	5.891

Table 10.3: A summary of standard distances $\left(s_d^*\right)$

10.2.3 DATA EXPLORATION

The values from Table 10.3 can be described in Table 10.4.



Statistics

		Standard distance (2D, Unadjusted)	Standard distance (2D, Adjusted)	Standard distance (2. 5D, Unadjusted)	Standard distance (2. 5D, Adjusted)	Standard distance (3D, Unadjusted)	Standard distance (3D, Adjusted)
N	Valid	8	8	8	8	8	8
	Missing	0	0	0	0	0	0
Mean		4.874808	5.835828	4.882171	5.834337	4.961679	5.845775
Std. Deviation		.0682925	.0590546	.0771377	.0907099	.1219260	.0825873
Skewness		-.749	.689	.188	.999	-.541	-.604
Std. Error of Skewness		.752	.752	.752	.752	.752	.752

Table 10.4: Some descriptive statistics for the standard distances

At first glance, it is obvious that the average standard distances can be split into two groups; the values from knowledge elicitation sessions using adjusted parameters are clearly larger than those from using unadjusted parameters.

The main effects of the visual representation dimension and model parameters factors, and their interaction effect on standard distance were further assessed through a visual inspection of the graphs in Figure 10.2 and Figure 10.3. Figure 10.2 plots standard distance against visual representation dimension for both treatment levels of the model parameters factor, and Figure 10.3 plots standard distance against model parameters for all treatment levels of the visual representation dimension factor. As eight experts had participated fully in the experiment carried out using three different visual representation dimensions and under two different sets of model parameters, there are 16 groups and 24 pairs of data available for plotting Figure 10.2 and Figure 10.3 respectively.

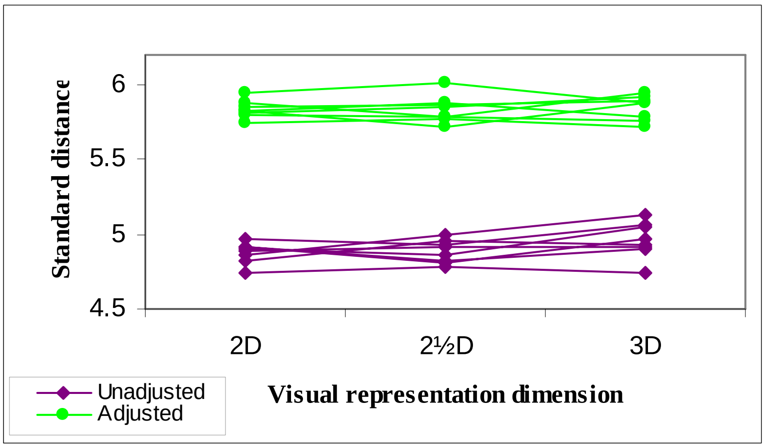


Figure 10.2: Hypothesis 4 – A comparison of standard distances under different visual representation dimensions

It can be observed in Figure 10.2 that the lines for unadjusted (purple) and adjusted (green) model parameters are roughly parallel; this indicates that the visual representation dimension and model parameters factors do not interact. Also, it is noted that the purple and green lines are distinctly segregated from each other; this implies that the model parameters factor's main effect is likely to be significant. Moreover, as the green lines are well above the purple lines, this seems to provide some support for the alternative hypothesis ($H_{4(a)}$). Further to this, regardless of model parameters, the lines are generally flat; this suggests that the visual representation dimension factor's main effect is probably not significant.

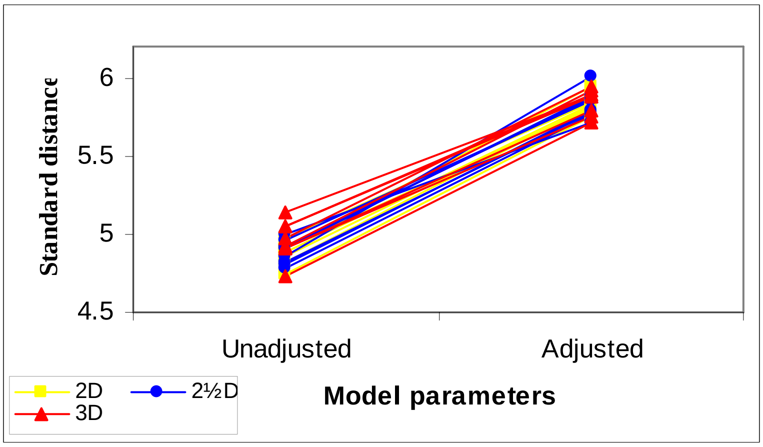


Figure 10.3: Hypothesis 4 – A comparison of standard distances under different model parameters

It can be observed in Figure 10.3 that the lines for the 2D (yellow), 2½D (blue) and 3D (red) representations appear to have a strong positive gradient; this reinforces the earlier observation that the model parameters factor's main effect is likely to be significant, and that adjusted parameters might lead to a larger state space (measured by the standard distance). Also, it is noted that all lines are parallel and close to each other; this implies that there is probably no interaction effect between the visual representation dimension and model parameters factors, and no main effect from the visual representation dimension factor.

10.2.4 HYPOTHESIS TESTING

Test of normality

Following the analytical framework, a histogram is generated for each experimental condition in Table 10.3 to provide an initial indication of whether the standard distances are normally distributed. The six histograms generated are shown in Figure 10.4. Due to the limited number of values plotted in each graph, none of them is seen to have a convincing bell shape.

Given the graphs' inability to show clearly if the distributions are close enough to normality to be useful, goodness-of-fit tests (K-S and A-D tests) are performed on each set of standard distances to ascertain their distributions. The results from the series of K-S and A-D tests executed are summarised in Table 10.5. The p-values for all six sets of standard distances in both tests are greater than 0.05. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of standard distances are normally distributed at a 5% level of significance.

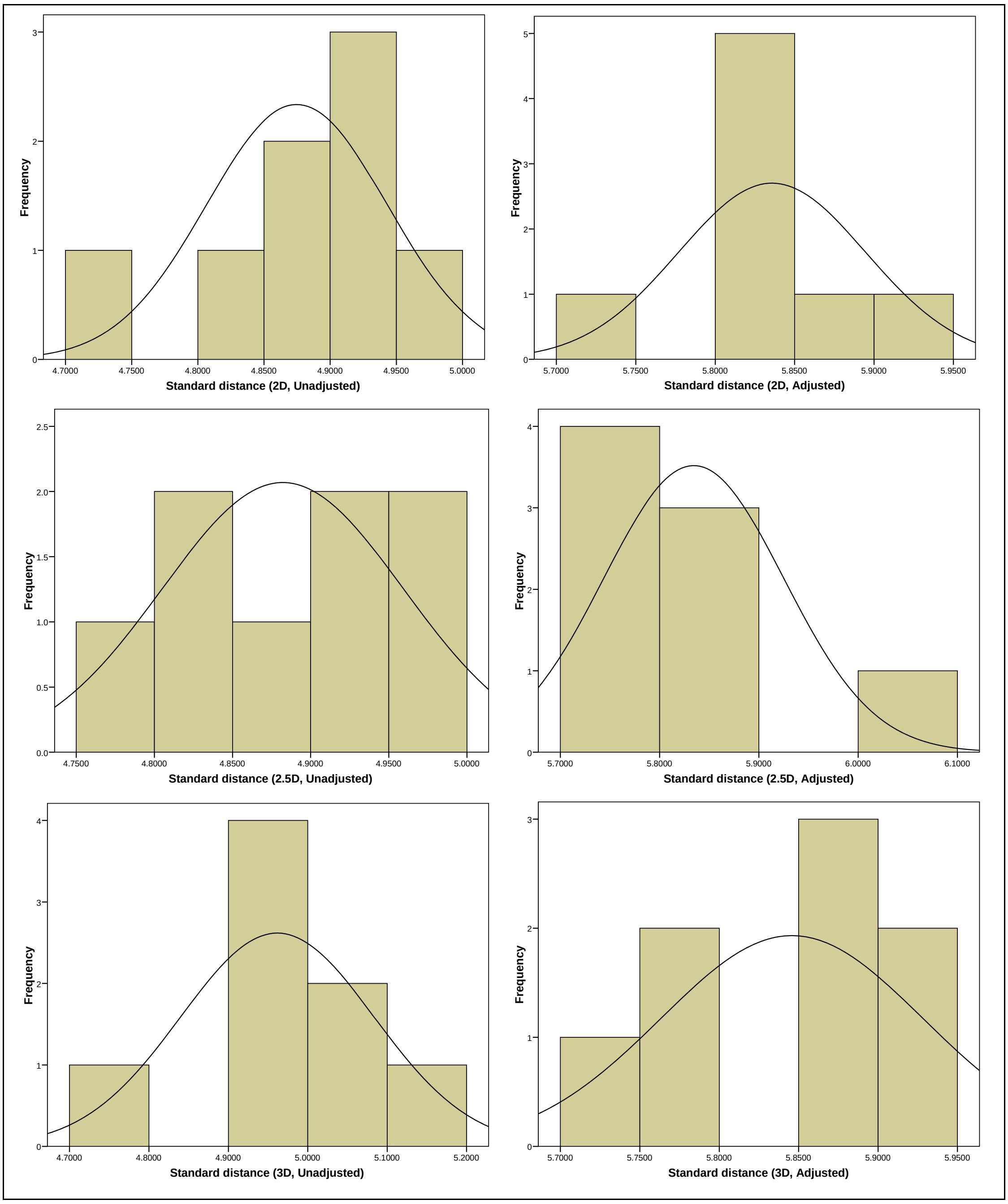


Figure 10.4: Hypothesis 4 – A summary of histograms generated for the standard distances

	Kolmogorov-Smirnov(a)			Anderson-Darling	
	Statistic	df	Sig.	Statistic	Sig.
Standard distance (2D, Unadjusted)	.186	8	.200(*)	.318	.450
Standard distance (2D, Adjusted)	.181	8	.200(*)	.300	.502
Standard distance (2½D, Unadjusted)	.177	8	.200(*)	.248	.643
Standard distance (2½D, Adjusted)	.194	8	.200(*)	.339	.397
Standard distance (3D, Unadjusted)	.205	8	.200(*)	.284	.532
Standard distance (3D, Adjusted)	.274	8	.079	.420	.241

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Table 10.5: Hypothesis 4 – A summary of results from the Kolmogorov-Smirnov and Anderson-Darling tests performed on the standard distances

Further to this, a series of skewness tests are also performed on the six sets of standard distances to determine if their distributions are symmetrical. Using the relevant descriptive statistics from Table 10.4, the test statistics for the skewness tests are computed and summarised in Table 10.6.

	Skewness	Standard Error	Test Statistic
Standard distance (2D, Unadjusted)	-0.749	0.752	-0.996
Standard distance (2D, Adjusted)	0.689	0.752	0.916
Standard distance (2½D, Unadjusted)	0.188	0.752	0.250
Standard distance (2½D, Adjusted)	0.999	0.752	1.328
Standard distance (3D, Unadjusted)	-0.541	0.752	-0.719
Standard distance (3D, Adjusted)	-0.604	0.752	-0.803

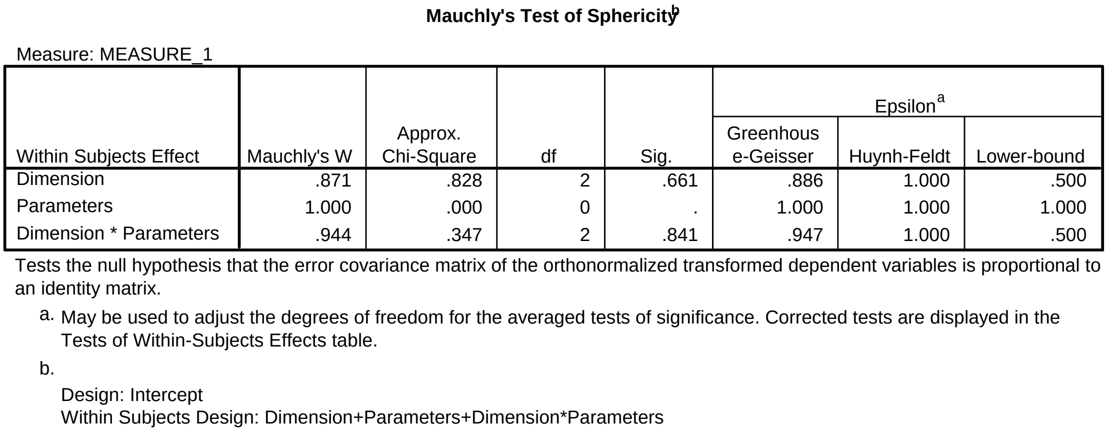
Table 10.6: Hypothesis 4 – A summary of results from the skewness tests performed on the standard distances

The absolute values of the test statistics for all six sets of standard distances are less than 1.96. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of standard distances are symmetrically distributed at a 5% level of significance.

In short, it can be shown that the values in Table 10.3 meet the criterion for normality, and hence parametric tests can be used on them.

Test of sphericity

Whilst the results above sanction the use of parametric ANOVAs on the standard distance data, the latter still needs to be tested for sphericity. This test is required for deciding whether it is necessary to revise the critical values that are used for assessing the test statistics from the ANOVAs that follow. As outlined in the analytical framework, Mauchly's test will be performed on the differences between the treatment levels of each possible main and interaction effect to assess the severity of departure from sphericity. The results from the series of Mauchly's tests performed are summarised in Table 10.7.



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Dimension	.871	.828	2	.661	.886	1.000	.500
Parameters	1.000	.000	0	.	1.000	1.000	1.000
Dimension * Parameters	.944	.347	2	.841	.947	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: Dimension+Parameters+Dimension*Parameters

Table 10.7: Hypothesis 4 – A summary of results from the Mauchly's tests performed on the relevant differences

The p-values for the visual representation dimension factor's main effect and the interaction effect between the latter and the model parameters factor are 0.661 and 0.841 respectively. Since these are more than 0.05, there is insufficient evidence to reject the null hypotheses that the variances of differences are not different at a 5% level of significance. This means that the criterion for sphericity is met, and the corresponding critical values for the following ANOVAs need not be revised. On the other hand, as the model parameters factor has only two treatment levels, the sphericity criterion is not relevant. As such, the Mauchly's test of sphericity and its results for the model parameters factor in Table 10.7 are not used.

Test of main and interaction effects

The results from the two-way repeated measures ANOVA performed are summarised in Table 10.8. The output is split into sections that refer to the different effects and their associated error terms.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Sphericity Assumed	.024	2	.012	2.259	.141
	Greenhouse-Geisser	.024	1.772	.013	2.259	.149
	Huynh-Feldt	.024	2.000	.012	2.259	.141
	Lower-bound	.024	1.000	.024	2.259	.177
Error(Dimension)	Sphericity Assumed	.073	14	.005		
	Greenhouse-Geisser	.073	12.401	.006		
	Huynh-Feldt	.073	14.000	.005		
	Lower-bound	.073	7.000	.010		
Parameters	Sphericity Assumed	10.433	1	10.433	1839.127	.000
	Greenhouse-Geisser	10.433	1.000	10.433	1839.127	.000
	Huynh-Feldt	10.433	1.000	10.433	1839.127	.000
	Lower-bound	10.433	1.000	10.433	1839.127	.000
Error(Parameters)	Sphericity Assumed	.040	7	.006		
	Greenhouse-Geisser	.040	7.000	.006		
	Huynh-Feldt	.040	7.000	.006		
	Lower-bound	.040	7.000	.006		
Dimension * Parameters	Sphericity Assumed	.014	2	.007	2.070	.163
	Greenhouse-Geisser	.014	1.894	.007	2.070	.166
	Huynh-Feldt	.014	2.000	.007	2.070	.163
	Lower-bound	.014	1.000	.014	2.070	.193
Error(Dimension* Parameters)	Sphericity Assumed	.048	14	.003		
	Greenhouse-Geisser	.048	13.256	.004		
	Huynh-Feldt	.048	14.000	.003		
	Lower-bound	.048	7.000	.007		

Table 10.8: Hypothesis 2 – A summary of results from the two-way repeated measures ANOVA performed on the standard distances

On the one hand, as the criterion for sphericity is met with respect to the visual representation dimension factor’s main effect and the interaction effect, the p-values that correspond to ‘Sphericity Assumed’ in Table 10.8 are used for each effect. They are 0.141 and 0.163 respectively. Since these are more than 0.05, there is insufficient evidence to reject the null hypothesis that if type of model parameters used is ignored, using different types of visual representation dimension does not affect state space at a 5% level of significance. Also, there is insufficient evidence to reject the null hypothesis that the effect which visual representation dimension (model parameters) has on state space is independent of the model parameters (visual representation dimension) associated with it at a 5% level of significance. That is, there is no interaction between the two factors in relation to state space.

On the other hand, as the criterion for sphericity is not relevant for the model parameters factor, there is only one p-value for its main effect in Table 10.8: zero. Since it is less than 0.05, there is sufficient evidence to reject the null hypothesis that if the type of visual representation dimension used is ignored, using different types of model parameters does not affect state space at a 5% level of significance.

Planned comparison

The preliminary data exploration in Section 10.2.3 seems to support the alternative hypothesis ($H_{4(a)}$) that state space (measured by the standard distance) increases in size as model parameters are adjusted to develop more uncommon and extreme scenes. Therefore, a planned comparison should be used instead of a post-hoc test to establish this observation.

Nonetheless, in view of the fact that there are only two treatment levels in the model parameters factor, this planned comparison is not necessary. It is because the results from the planned comparison will not differ from that of the ANOVA in Table 10.8. That is, there is sufficient evidence to reject the null hypothesis that these treatment levels are not different in their effects on standard distance at a 5% level of significance.

In addition, the earlier observation on the mean standard distances from knowledge elicitation sessions using adjusted parameters being clearly larger than those from using unadjusted parameters suggests that the adjusted parameters lead to a larger state space than the unadjusted parameters.

Effect size

Substituting the relevant values from the SPSS output[8] in Appendix E.1 into the expression for computing effect size,

$$r_{Unadjusted\ vs\ Adjusted} = \sqrt{\frac{1{,}839.127}{1{,}839.127 + 7}} = 0.998$$

Since $r_{Unadjusted\ vs\ Adjusted}$ is more than 0.50, it can be concluded that the effect between unadjusted and adjusted parameters is large.

10.2.5 Summary

The overarching pair of null and alternative hypotheses of interest are assessed by performing various tests prescribed by the analytical framework. Subsequently, there is sufficient evidence to reject the null hypothesis ($H_{4(0)}$) at a 5% level of significance. Hence, it can be concluded that the size of state space occupied by the example cases collected in a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes. Also, it can be concluded that using the adjusted set of parameters over the unadjusted set has a large effect on the size of state space.

[8] The SPSS output is generated from performing appropriate planned contrasts. In this case, the values are also available in Table 10.8.

10.3 ANALYSIS FOR HYPOTHESIS TWO: CASE QUANTITY & VISUAL REPRESENTATION DIMENSION, AND HYPOTHESIS FIVE: CASE QUANTITY & MODEL PARAMETERS

The overarching pairs of null and alternative hypotheses of interest are replicated below:

$H_{2(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{2(a)}$: The size of case quantity of a knowledge elicitation session increases as a higher dimension of visual representation is used.

$H_{5(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the model parameters used;

$H_{5(a)}$: The size of case quantity of a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

Case quantity has been defined, in Section 4.2.3, as the total number of example cases recorded in a knowledge elicitation session. They are as summarised in Table 10.2.

10.3.1 DATA EXPLORATION

The values from Table 10.2 can be described in Table 10.9.

Statistics

		Case quantity (2D, Unadjusted)	Case quantity (2D, Adjusted)	Case quantity (2.5D, Unadjusted)	Case quantity (2.5D, Adjusted)	Case quantity (3D, Unadjusted)	Case quantity (3D, Adjusted)
N	Valid	8	8	8	8	8	8
	Missing	0	0	0	0	0	0
Mean		69.50	42.38	42.13	28.38	79.75	76.25
Std. Deviation		47.431	22.136	33.909	12.817	68.938	31.994
Skewness		1.032	.553	2.283	.800	2.196	.155
Std. Error of Skewness		.752	.752	.752	.752	.752	.752

Table 10.9: Some descriptive statistics for the case quantities

A quick review reveals that that the average case quantities from knowledge elicitation sessions using the 3D representation are larger than those from using the 2D representation, which in turn are larger than those from using the 2½D representation. Also, it is noticeable that the average case quantities from knowledge elicitation sessions using unadjusted parameters are larger than those from using adjusted parameters.

The visual representation dimension and model parameters factors' main effects, and their interaction effect on case quantity were further assessed through a visual inspection of the graphs in Figure 10.5 and Figure 10.6. Figure 10.5 plots case quantity against visual representation dimension for both treatment levels of the model parameters factor, and Figure 10.6 plots case quantity against model parameters for all treatment levels of the visual representation dimension factor. As eight experts had participated fully in the experiment carried out using three different visual representation dimensions and under two different sets of model parameters, there are

16 groups and 24 pairs of data available for plotting Figure 10.5 and Figure 10.6 respectively.

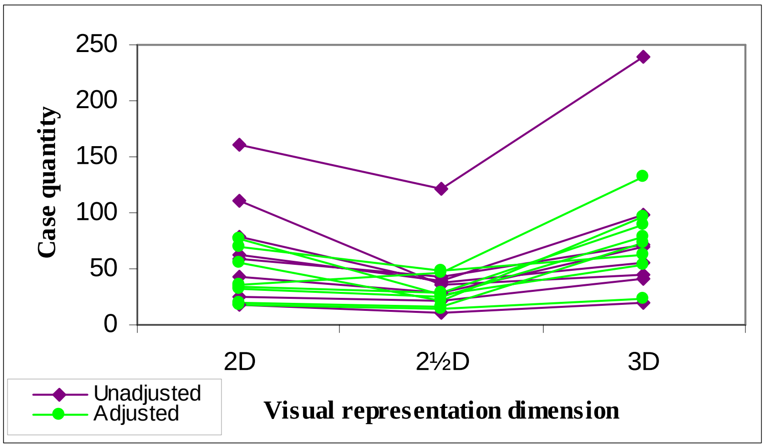


Figure 10.5: Hypothesis 2 and 5 – A comparison of case quantities under different visual representation dimensions

It can be observed in Figure 10.5 that the lines for unadjusted (purple) and adjusted (green) model parameters are roughly similar in shape: the lines appear to create a valley, being higher at both ends at the 2D and 3D representations. This suggests that the visual representation dimension and model parameters factors do not interact. Also, this suggests that the visual representation dimension factor's main effect might be significant, and that a 2½D representation seems to be the least effective in encouraging the experts to interact with the game models. In addition, it is noted that the purple and green lines are not distinctly segregated from each other; this implies that the model parameters factor's main effect might not be significant.

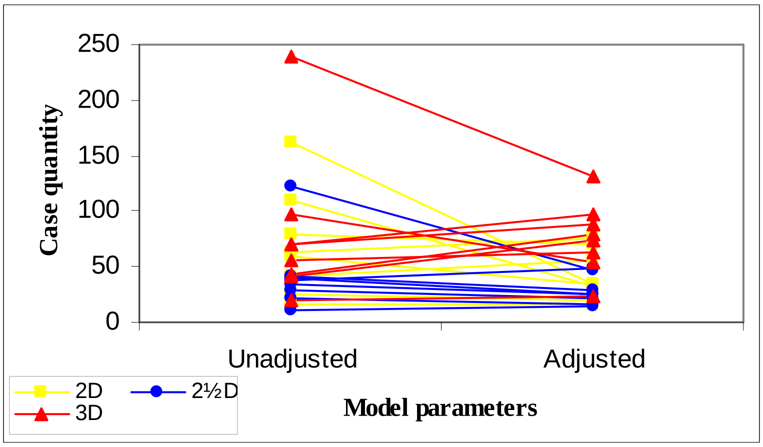


Figure 10.6: Hypothesis 2 and 5 – A comparison of case quantities under different model parameters

It can be observed in Figure 10.6 that the lines for the 2D (yellow), 2½D (blue) and 3D (red) representations do not display a strong pattern; this reinforces the earlier observation that the model parameters factor's main effect might not be significant. Also, it is noted that the blue lines appear to congregate at the bottom of the graph, away from the red and yellow lines; this hints that the visual representation dimension factor might have a main effect.

10.3.2 HYPOTHESIS TESTING

Test of normality

Following the analytical framework, a histogram is generated for each experimental condition in Table 10.2 to provide an initial indication of whether the case quantities are normally distributed. The six histograms generated are shown in Figure 10.7. Except for the histogram generated for the 3D representation with adjusted parameters, the remaining graphs do not appear to have a convincing bell shape. This is again due to the limited number of values plotted in each graph.

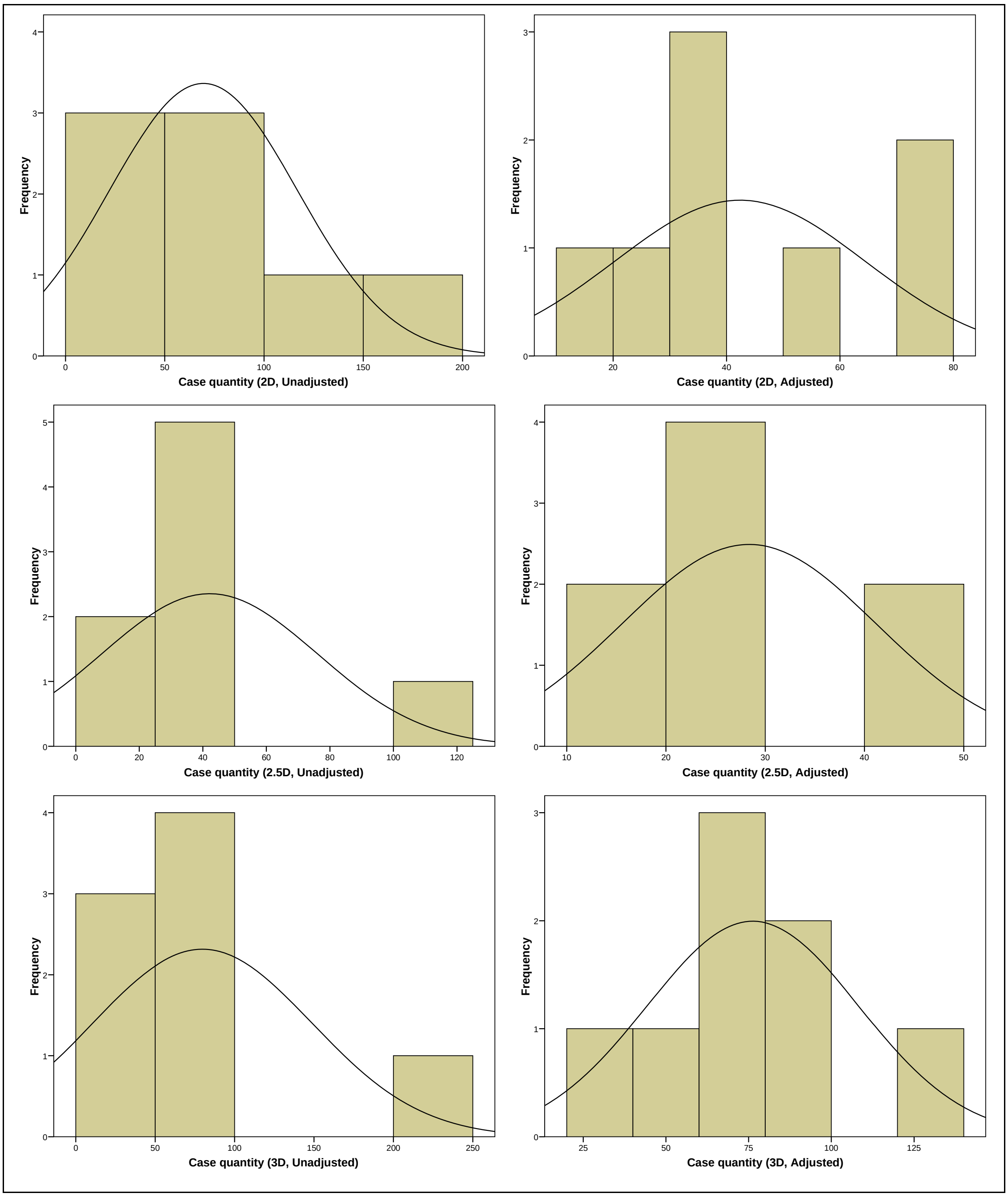
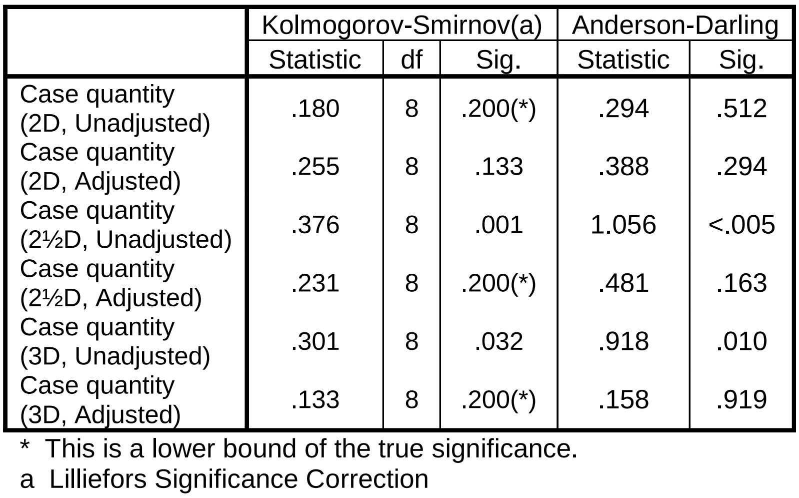


Figure 10.7: Hypothesis 2 and 5 – A summary of histograms generated for the case quantities

Given the graphs' inability to show clearly if the distributions are close enough to normality to be useful, goodness-of-fit tests (K-S and A-D tests) are performed on each set of case quantities to ascertain their distributions. The results from the series of K-S and A-D tests performed are summarised in Table 10.10. On the one hand, the p-values for four sets of case quantities (2D representation with unadjusted parameters, and adjusted parameters with any visual representation dimension) are greater than 0.05 in both tests. Therefore, there is insufficient evidence to reject the null hypotheses that these four sets of case quantities are normally distributed at a 5% level of significance. On the other hand, the p-values for the other two sets of case quantities (unadjusted parameters with 2½D or 3D representation) are less than 0.05 in both tests. Therefore, there is sufficient evidence to reject the null hypotheses that these two sets of case quantities are normally distributed at a 5% level of significance. That is, these two sets of case quantities are not normally distributed.



	Kolmogorov-Smirnov(a)			Anderson-Darling	
	Statistic	df	Sig.	Statistic	Sig.
Case quantity (2D, Unadjusted)	.180	8	.200(*)	.294	.512
Case quantity (2D, Adjusted)	.255	8	.133	.388	.294
Case quantity (2½D, Unadjusted)	.376	8	.001	1.056	<.005
Case quantity (2½D, Adjusted)	.231	8	.200(*)	.481	.163
Case quantity (3D, Unadjusted)	.301	8	.032	.918	.010
Case quantity (3D, Adjusted)	.133	8	.200(*)	.158	.919

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Table 10.10: Hypothesis 2 and 5 – A summary of results from the Kolmogorov-Smirnov and Anderson-Darling tests performed on the case quantities

Further to this, a series of skewness tests are also performed on the six sets of case quantities to determine if their distributions are symmetrical. Using the relevant descriptive statistics from Table 10.9, the test statistics for the skewness tests are computed and summarised in Table 10.11.

	Skewness	Standard Error	Test Statistic
Case quantity (2D, Unadjusted)	1.032	0.752	1.372
Case quantity (2D, Adjusted)	0.553	0.752	0.735
Case quantity (2½D, Unadjusted)	2.283	0.752	3.036
Case quantity (2½D, Adjusted)	0.800	0.752	1.064
Case quantity (3D, Unadjusted)	2.196	0.752	2.920
Case quantity (3D, Adjusted)	0.155	0.752	0.206

Table 10.11: Hypothesis 2 and 5 – A summary of results from the skewness tests performed on the case quantities

Likewise, the test statistics for the same four sets of case quantities (2D representation with unadjusted parameters, and adjusted parameters with any visual representation dimension) have absolute values that are less than 1.96. Therefore, there is insufficient evidence to reject the null hypotheses that these four sets of case quantities are symmetrically distributed at a 5% level of significance. As well, the test statistics for the other two sets of case quantities (unadjusted parameters with 2½D or 3D representation) have absolute values that are more than 1.96. Therefore, there is sufficient evidence to reject the null hypotheses that these two sets of case quantities are symmetrically distributed at a 5% level of significance. That is, these two sets of case quantities are not symmetrically distributed.

In short, it can be shown that the values in Table 10.2 do not meet the criterion for normality, and hence parametric tests cannot be used on them. As a remedy, Field (2006) and Hair *et al*. (2006) suggest executing a logarithmic transformation on the data, before putting them through another cycle of the K-S, A-D and skewness tests. The transformed case quantities are summarised in Table 10.12.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	1.623	1.740	1.462	1.342	1.845	1.987
B	2.207	1.544	2.086	1.672	2.380	2.121
C	1.771	1.531	1.623	1.462	1.851	1.949
D	1.799	1.881	1.602	1.415	1.991	1.724
E	1.230	1.230	1.041	1.146	1.279	1.380
F	2.041	1.505	1.544	1.398	1.643	1.898
G	1.898	1.845	1.568	1.681	1.740	1.799
H	1.398	1.301	1.322	1.204	1.613	1.863

Table 10.12: Hypothesis 2 and 5 – A summary of transformed case quantities

Likewise, a histogram is generated for each experimental condition in Table 10.12 to provide an initial indication of whether the transformed case quantities are normally distributed. The six histograms generated are shown in Figure 10.8. Unlike those in Figure 10.7, the histograms generated for the transformed data have a closer resemblance to a bell shape now. To support this observation, the K-S, A-D and skewness tests were carried out, with the results summarised in Table 10.13 and Table 10.14 respectively. The values that were used to compute the test statistics for the skewness tests are available in Appendix D.1.

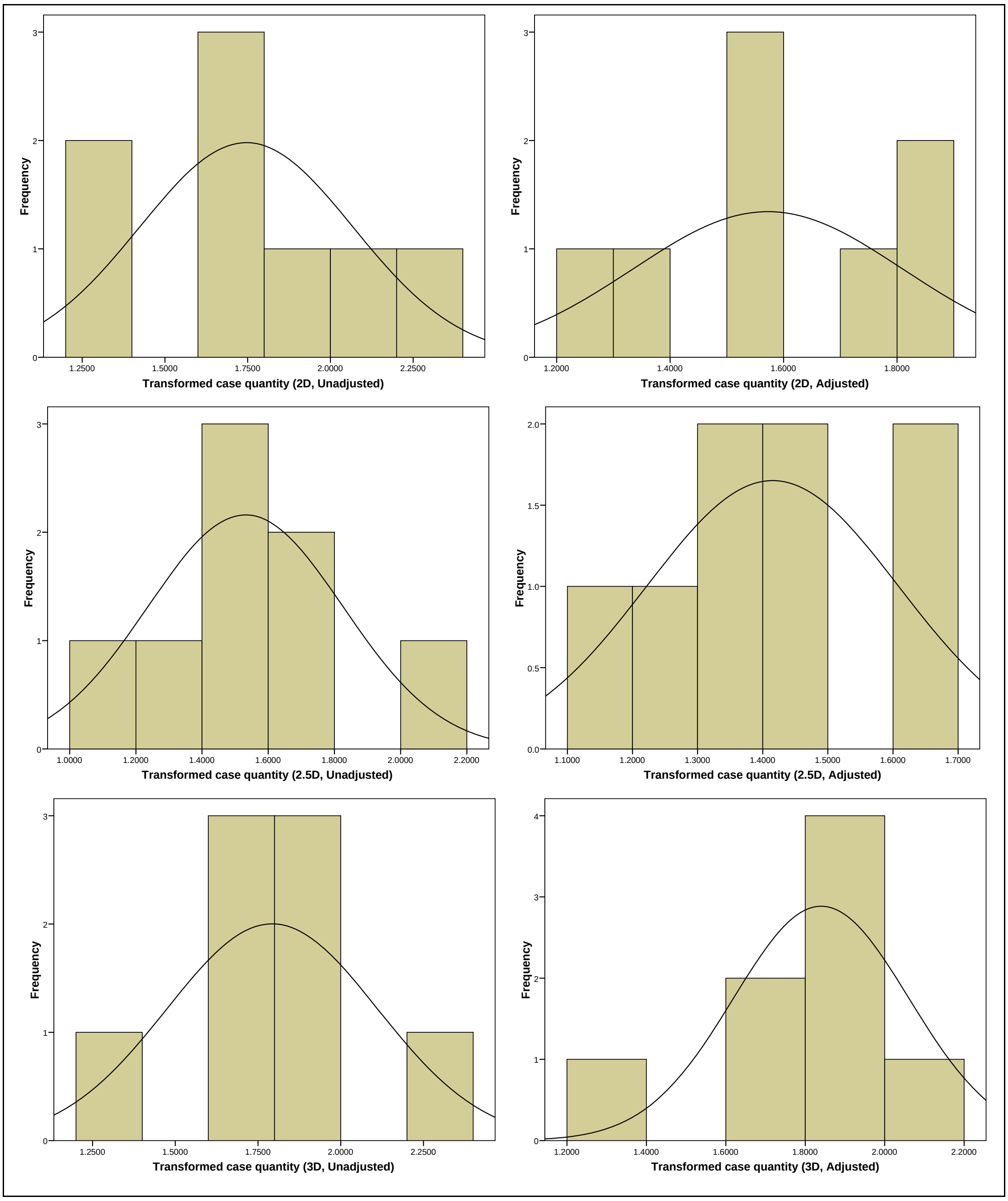


Figure 10.8: Hypothesis 2 and 5 – A summary of histograms generated for the transformed case quantities

	Kolmogorov-Smirnov(a)			Anderson-Darling	
	Statistic	df	Sig.	Statistic	Sig.
Transformed case quantity (2D, Unadjusted)	.156	8	.200(*)	.154	.926
Transformed case quantity (2D, Adjusted)	.172	8	.200(*)	.270	.570
Transformed case quantity (2½D, Unadjusted)	.253	8	.141	.411	.256
Transformed case quantity (2½D, Adjusted)	.158	8	.200(*)	.271	.566
Transformed case quantity (3D, Unadjusted)	.177	8	.200(*)	.258	.609
Transformed case quantity (3D, Adjusted)	.177	8	.200(*)	.370	.328

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Table 10.13: Hypothesis 2 and 5 – A summary of results from the Kolmogorov-Smirnov and Anderson-Darling tests performed on the transformed case quantities

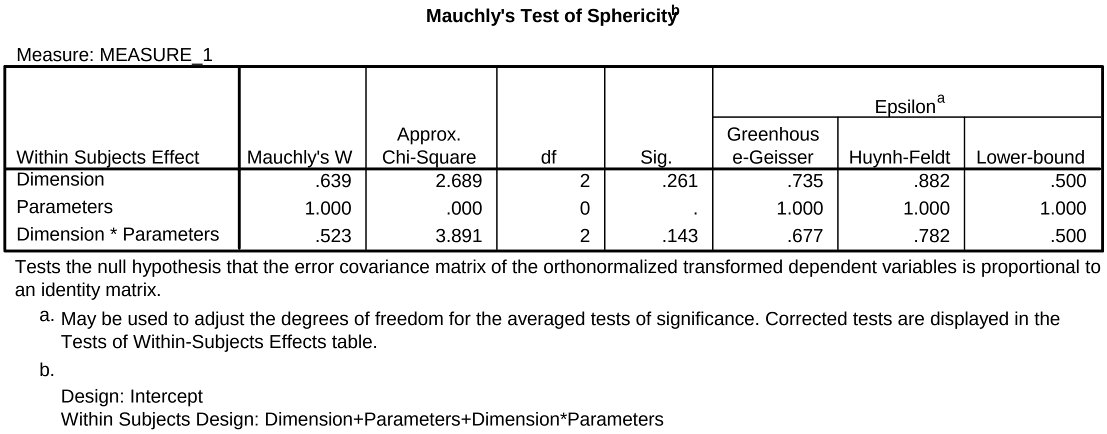
	Skewness	Standard Error	Test Statistic
Tformed case quantity (2D, Unadjusted)	-0.305	0.752	-0.406
Tformed case quantity (2D, Adjusted)	-0.084	0.752	-0.112
Tformed case quantity (2½D, Unadjusted)	0.334	0.752	0.444
Tformed case quantity (2½D, Adjusted)	0.175	0.752	0.233
Tformed case quantity (3D, Unadjusted)	0.384	0.752	0.511
Tformed case quantity (3D, Adjusted)	-1.271	0.752	-1.690

Table 10.14: Hypothesis 2 and 5 – A summary of results from the skewness tests performed on the transformed case quantities

The p-values for all six sets of transformed case quantities in both tests in Table 10.13 are greater than 0.05. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of transformed case quantities are normally distributed at a 5% level of significance. In addition, the absolute values of the test statistics computed for the skewness tests in Table 10.14 are all less than 1.96. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of transformed case quantities are symmetrically distributed at a 5% level of significance. As such, it can be shown that the transformed values in Table 10.12 meet the criterion for normality, and hence parametric tests can be used on them.

Test of sphericity

Whilst the results above sanction the use of parametric ANOVAs on the transformed case quantity data, the latter still needs to be tested for sphericity. This test is required for deciding whether it is necessary to revise the critical values that are used for assessing the test statistics from the ANOVAs that follow. As outlined in the analytical framework, Mauchly's test is performed on the differences between the treatment levels of each possible main and interaction effect to assess the severity of departure from sphericity. The results from the series of Mauchly's tests performed are summarised in Table 10.15.



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Dimension	.639	2.689	2	.261	.735	.882	.500
Parameters	1.000	.000	0	.	1.000	1.000	1.000
Dimension * Parameters	.523	3.891	2	.143	.677	.782	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: Dimension+Parameters+Dimension*Parameters

Table 10.15: Hypothesis 2 and 5 – A summary of results from the Mauchly's tests performed on the relevant differences

The p-values for the visual representation dimension factor's main effect and the interaction effect between the latter and the model parameters factor are 0.261 and 0.143 respectively. Since these are more than 0.05, there is insufficient evidence to reject the null hypotheses that the variances of differences are not different at a 5% level of significance. This means that the criterion for sphericity is met, and the corresponding critical values for the following ANOVAs need not be revised. On the other hand, as the model parameters factor has only two treatment levels, the sphericity criterion is not relevant. As such, the Mauchly's test of sphericity and its results for the model parameters factor in Table 10.15 are not used.

Test of main and interaction effects

The results from the two-way repeated measures ANOVA performed are summarised in Table 10.16. The output is split into sections that refer to the different effects and their associated error terms.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Sphericity Assumed	.945	2	.473	22.960	.000
	Greenhouse-Geisser	.945	1.469	.643	22.960	.000
	Huynh-Feldt	.945	1.764	.536	22.960	.000
	Lower-bound	.945	1.000	.945	22.960	.002
Error(Dimension)	Sphericity Assumed	.288	14	.021		
	Greenhouse-Geisser	.288	10.285	.028		
	Huynh-Feldt	.288	12.346	.023		
	Lower-bound	.288	7.000	.041		
Parameters	Sphericity Assumed	.078	1	.078	1.821	.219
	Greenhouse-Geisser	.078	1.000	.078	1.821	.219
	Huynh-Feldt	.078	1.000	.078	1.821	.219
	Lower-bound	.078	1.000	.078	1.821	.219
Error(Parameters)	Sphericity Assumed	.301	7	.043		
	Greenhouse-Geisser	.301	7.000	.043		
	Huynh-Feldt	.301	7.000	.043		
	Lower-bound	.301	7.000	.043		
Dimension * Parameters	Sphericity Assumed	.105	2	.053	3.180	.073
	Greenhouse-Geisser	.105	1.354	.078	3.180	.099
	Huynh-Feldt	.105	1.564	.067	3.180	.090
	Lower-bound	.105	1.000	.105	3.180	.118
Error(Dimension* Parameters)	Sphericity Assumed	.231	14	.017		
	Greenhouse-Geisser	.231	9.477	.024		
	Huynh-Feldt	.231	10.949	.021		
	Lower-bound	.231	7.000	.033		

Table 10.16: Hypothesis 2 and 5 – A summary of results from the two-way repeated measures ANOVA performed on the transformed case quantities

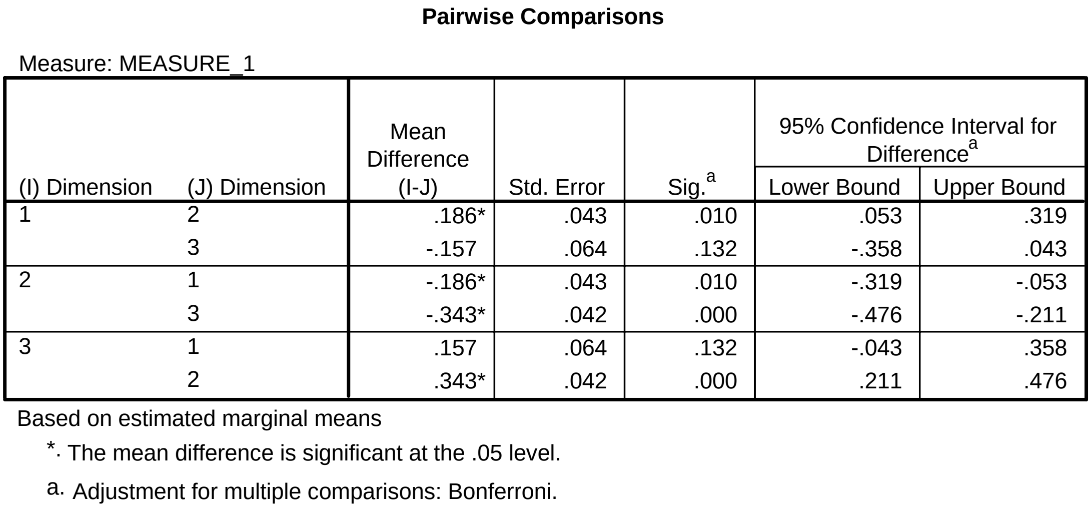
On the one hand, as the criterion for sphericity is met with respect to the visual representation dimension factor's main effect and the interaction effect, the p-values that correspond to 'Sphericity Assumed' in Table 10.16 are used for each effect. They are zero and 0.073 respectively. Since the former is less than 0.05, there is sufficient evidence to reject the null hypothesis that if type of model parameters used is ignored, using different types of visual representation dimension does not affect case quantity at a 5% level of significance. However, as the p-value for the interaction effect is more than 0.05, there is insufficient evidence to reject the null hypothesis that the effect which the visual representation dimension (model parameters) has on case quantity is independent of the model parameters (visual representation dimension) associated with it at a 5% level of significance. That is, there is no interaction between the two factors in relation to case quantity.

On the other hand, as the criterion for sphericity is not relevant for the model parameters factor, there is only one p-value for its main effect in Table 10.16: 0.219. Since it is more than 0.05, there is insufficient evidence to reject the null hypothesis that if the type of visual representation dimension used is ignored, using different types of model parameters does not affect case quantity at a 5% level of significance.

Post-hoc test

The preliminary data exploration in Section 10.3.1 generally agrees that case quantities from knowledge elicitation sessions using the 2D representation are larger than those from using the 2½D representation. As such, it does not appear to support the alternative hypothesis ($H_{2(a)}$) that case quantity increases as a higher dimension of visual representation is used. Consequently, it is more meaningful to use a post-hoc test, instead of a series of planned comparisons, to investigate the visual representation dimension factor's main effect on case quantity.

Having determined that using different types of visual representation dimension affect case quantity, the next step is to identify the visual representation dimension that leads to a larger case quantity. The results from the pairwise comparisons performed on the visual representation dimension factor are summarised in Table 10.17.



Pairwise Comparisons

Measure: MEASURE_1

(I) Dimension	(J) Dimension	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.186*	.043	.010	.053	.319
	3	-.157	.064	.132	-.358	.043
2	1	-.186*	.043	.010	-.319	-.053
	3	-.343*	.042	.000	-.476	-.211
3	1	.157	.064	.132	-.043	.358
	2	.343*	.042	.000	.211	.476

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Table 10.17: Hypothesis 2 and 5 – A summary of results from the pairwise comparisons of the visual representation dimension factor

Using a Bonferroni-adjusted critical value that maintains the overall Type I error rate at a 5% level of significance, SPSS was used to evaluate the mean differences between treatment level 1 (2D representation) and 2 (2½D representation), and between treatment level 2 and 3 (3D representation). These were found to be significant. Hence, there is sufficient evidence to reject the null hypotheses that these two pairs of treatment levels are not different in their effects on case quantity at an overall 5% level of significance. Moreover, there is insufficient evidence to reject the null hypothesis that treatment level 1 and 3 are not different in their effects on case quantity at an overall 5% level of significance.

Furthermore, the positive mean difference between treatment level 1 and 2 suggests that the 2D representation leads to a larger case quantity than the 2½D representation, whereas the negative mean difference between treatment level 2 and 3 suggests that the 3D representation leads to a larger case quantity than the 2½D representation.

Effect sizes

Substituting the relevant values from the SPSS output[9] in Appendix E.2 into the expression for computing effect size,

$$r_{2D\,vs\,2½D} = \sqrt{\frac{19.069}{19.069 + 7}} = 0.855$$

$$r_{2½D\,vs\,3D} = \sqrt{\frac{65.971}{65.971 + 7}} = 0.950$$

Since both $r_{2D\,vs\,2½D}$ and $r_{2½D\,vs\,3D}$ are more than 0.50, it can be concluded that the effects between the 2D and 2½D representations, and between the 2½D and 3D representations are large.

10.3.3 SUMMARY

The overarching pair of null and alternative hypotheses of interest are assessed by performing various tests prescribed by the analytical framework.

In relation to Hypothesis 2, there is sufficient evidence to reject the null hypothesis ($H_{2(0)}$) at a 5% level of significance. However, it cannot be concluded that the size of case quantity of a knowledge elicitation session generally increases as a higher dimension of visual representation is used. It is because whilst the 3D representation is

[9] The SPSS output is generated from performing appropriate planned contrasts.

shown to lead to a larger case quantity than the 2½D representation, the latter is not shown to lead to a larger case quantity than the 2D representation. Instead, the 2D representation can be shown to lead to a larger case quantity than the 2½D representation. Nevertheless, it can be concluded that using either the 2D or 3D representation over the 2½D representation has a large effect on the size of case quantity.

In relation to Hypothesis 5, there is insufficient evidence to reject the null hypothesis ($H_{5(0)}$) at a 5% level of significance. Hence, it is concluded that the size of case quantity of a knowledge elicitation session is not affected by the model parameters used.

10.4 ANALYSIS FOR HYPOTHESIS THREE: COLLECTION RATE & VISUAL REPRESENTATION DIMENSION, AND HYPOTHESIS SIX: COLLECTION RATE & MODEL PARAMETERS

The overarching pairs of null and alternative hypotheses of interest are replicated below:

$H_{3(0)}$: The collection rate in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{3(a)}$: The collection rate in a knowledge elicitation session is affected by the visual representation dimension used.

$H_{6(0)}$: The collection rate in a knowledge elicitation session is not affected by the model parameters used;

$H_{6(a)}$: The collection rate in a knowledge elicitation session is affected by the model parameters used.

Collection rate has been defined, in Section 4.2.4, as the number of example cases recorded per unit of real-time in a knowledge elicitation session. In this thesis, a unit of real-time is taken to be one minute. Using the data in Table 9.1 and Table 10.2, the rates for the experiments can be computed and are summarised in Table 10.18.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	1.400	1.486	1.074	0.917	0.959	1.043
B	3.220	0.946	3.486	1.424	2.500	1.361
C	1.311	1.133	1.235	0.879	1.109	1.023
D	1.575	1.900	1.379	0.963	1.289	0.659
E	0.850	0.680	0.500	0.500	0.292	0.300
F	2.444	0.914	0.972	0.735	0.564	0.898
G	1.580	1.556	1.233	1.500	0.709	0.700
H	0.735	0.455	0.583	0.552	0.612	0.768

Table 10.18: A summary of collection rates

10.4.1 DATA EXPLORATION

The values from Table 10.18 can be described in Table 10.19.

Statistics

		Rate (2D, Unadjusted)	Rate (2D, Adjusted)	Rate (2.5D, Unadjusted)	Rate (2.5D, Adjusted)	Rate (3D, Unadjusted)	Rate (3D, Adjusted)
N	Valid	8	8	8	8	8	8
	Missing	0	0	0	0	0	0
Mean		1.639375	1.133750	1.307750	.933750	1.004250	.844000
Std. Deviation		.8245868	.4839784	.9339486	.3661560	.6838466	.3158933
Skewness		1.083	.245	2.192	.598	1.690	-.114
Std. Error of Skewness		.752	.752	.752	.752	.752	.752

Table 10.19: Some descriptive statistics for the collection rates

A quick scan reveals that that the average collection rates from knowledge elicitation sessions using the 2D representation are higher than those from using the 2½D representation, which in turn are higher than those from using the 3D representation. Also, it is noticed that the average collection rates from knowledge elicitation sessions using unadjusted parameters are higher than those from using adjusted parameters.

The visual representation dimension and model parameters factors' main effects, and their interaction effect on collection rate were further assessed through a visual inspection of the graphs in Figure 10.9 and Figure 10.10. Figure 10.9 plots collection rate against visual representation dimension for both treatment levels of the model parameters factor, and Figure 10.10 plots collection rate against model parameters for all treatment levels of the visual representation dimension factor. As eight experts had participated fully in the experiment carried out using three different visual representation dimensions and under two different sets of model parameters, there are 16 groups and 24 pairs of data available for plotting Figure 10.9 and Figure 10.10 respectively.

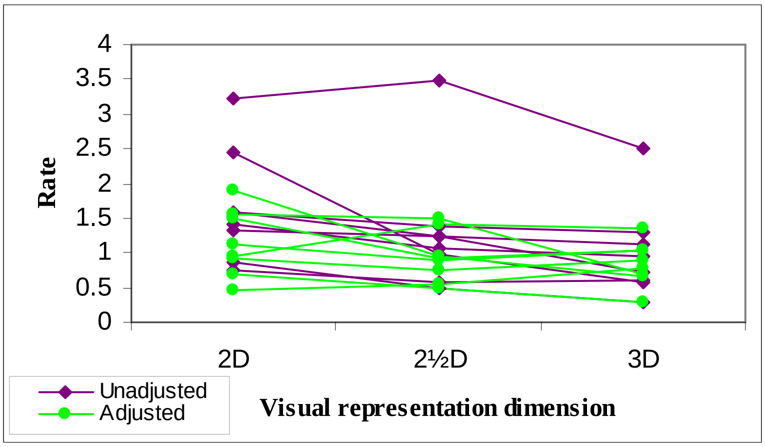


Figure 10.9: Hypothesis 3 and 6 – A comparison of collection rates under different visual representation dimensions

Except for two uncharacteristic purple lines, it can be observed in Figure 10.9 that the remaining lines for unadjusted (purple) and adjusted (green) model parameters are otherwise roughly parallel to each other; this suggests that the visual representation dimension and model parameters factors do not interact. Also, it is noted that the purple and green lines are not distinctly segregated from each other; this implies that the model parameters factor's main effect might not be significant. In addition, with the exception of the two uncharacteristic purple lines, the remaining lines are quite flat. This suggests that the visual representation dimension factor's main effect is probably not significant, though it might not be true for the experts represented by the two purple lines.

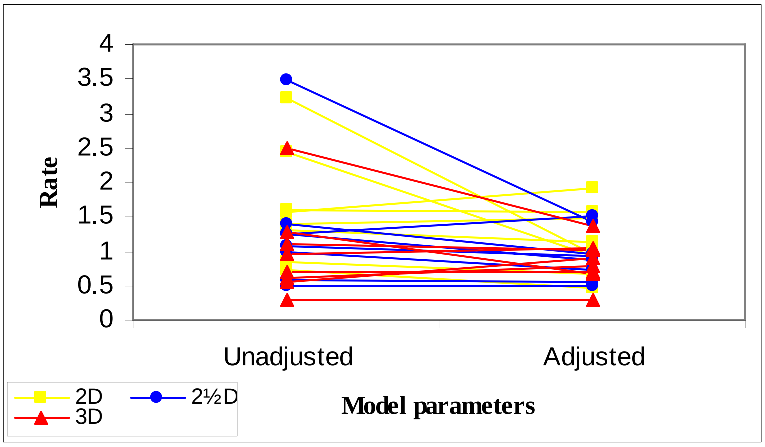


Figure 10.10: Hypothesis 3 and 6 – A comparison of collection rates under different model parameters

It can be observed in Figure 10.10 that the lines for the 2D (yellow), 2½D (blue) and 3D (red) representations do not display a strong pattern; this reinforces the earlier observation that the model parameters factor's main effect might not be significant. Also, it is noted that the yellow lines appear to congregate above the blue and red lines; this hints that the visual representation dimension factor might have a main effect.

10.4.2 Hypothesis testing

Test of normality

Following the analytical framework, a histogram is generated for each experimental condition in Table 10.18 to provide an initial indication of whether the collection rates are normally distributed. The six histograms generated are shown in Figure 10.11. Except for the histogram generated for the 3D representation with unadjusted parameters, the remaining graphs do not appear to have a convincing bell shape. This is again due to the limited number of values plotted in each graph.

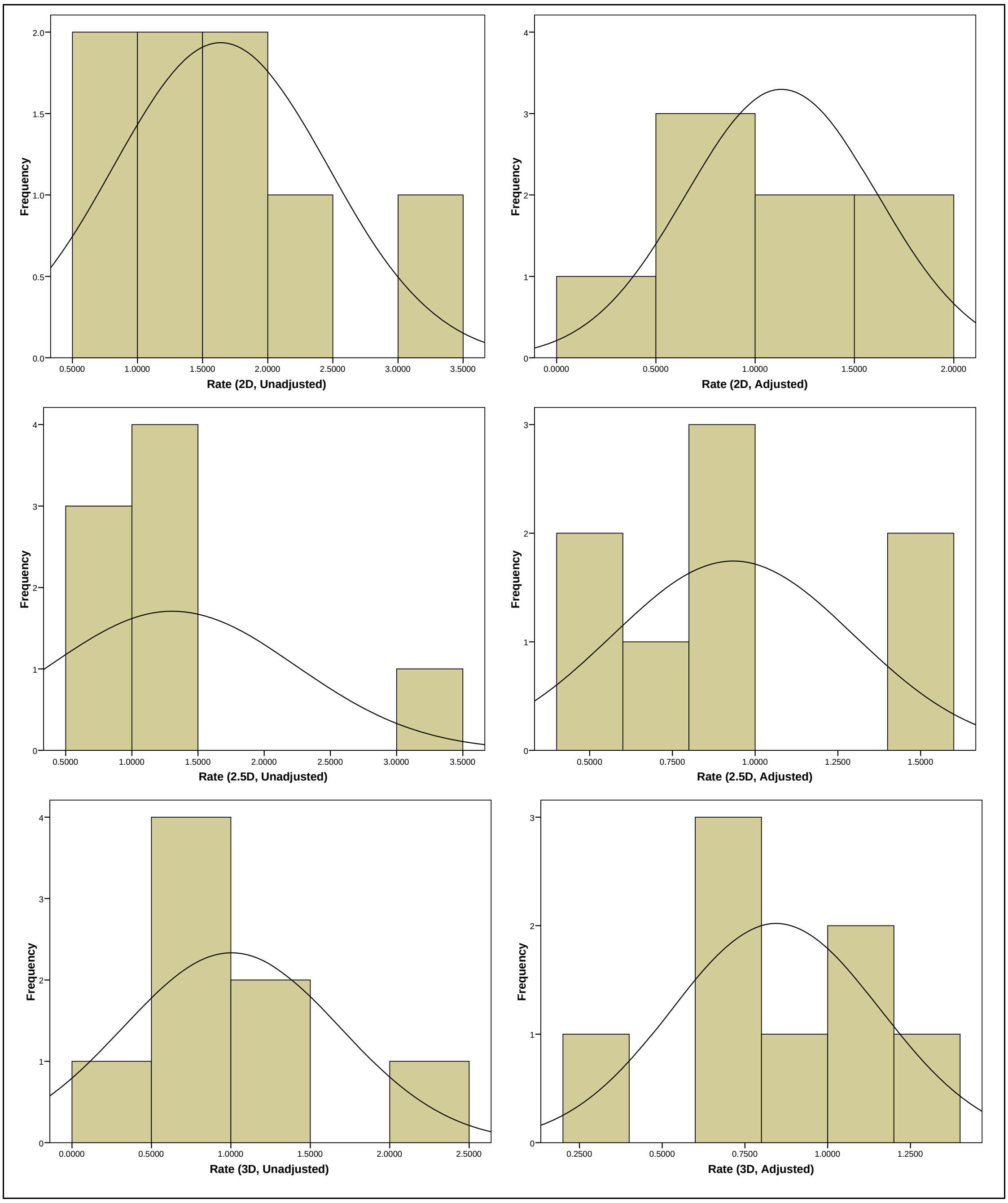


Figure 10.11: Hypothesis 3 and 6 – A summary of histograms generated for the collection rates

Given the graphs' inability to show clearly if the distributions are close enough to normality to be useful, goodness-of-fit tests (K-S and A-D tests) are performed on each set of collection rates to ascertain their distributions. The results from the series of K-S and A-D tests performed are summarised in Table 10.20. The p-values for all but one set of collection rates (2½D representation with unadjusted parameters) are greater than 0.05 in both tests. Therefore, there is insufficient evidence to reject the null hypotheses that these five sets of collection rates are normally distributed at a 5% level of significance. Conversely, there is sufficient evidence to reject the null hypothesis that the set of data from knowledge elicitation sessions using the 2½D representation with unadjusted parameters are normally distributed at a 5% level of significance.

	Kolmogorov-Smirnov(a)			Anderson-Darling	
	Statistic	df	Sig.	Statistic	Sig.
Rate (2D, Unadjusted)	.279	8	.067	.429	.229
Rate (2D, Adjusted)	.151	8	.200(*)	.180	.876
Rate (2½D, Unadjusted)	.345	8	.006	.961	.008
Rate (2½D, Adjusted)	.218	8	.200(*)	.356	.359
Rate (3D, Unadjusted)	.214	8	.200(*)	.546	.109
Rate (3D, Adjusted)	.154	8	.200(*)	.189	.853

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

Table 10.20: Hypothesis 3 and 6 – A summary of results from the Kolmogorov-Smirnov and Anderson-Darling tests performed on the collection rates

Further to this, a series of skewness tests are also performed on the six sets of collection rates to determine if their distributions are symmetrical. Using the relevant descriptive statistics from Table 10.19, the test statistics for the skewness tests are computed and summarised in Table 10.21.

	Skewness	Standard Error	Test Statistic
Rate (2D, Unadjusted)	1.083	0.752	1.440
Rate (2D, Adjusted)	0.245	0.752	0.326
Rate (2½D, Unadjusted)	2.192	0.752	2.915
Rate (2½D, Adjusted)	0.598	0.752	0.795
Rate (3D, Unadjusted)	1.690	0.752	2.247
Rate (3D, Adjusted)	-0.114	0.752	-0.152

Table 10.21: Hypothesis 3 and 6 – A summary of results from the skewness tests performed on the collection rates

In this case, the test statistics for four sets of collection rates (2D representation with unadjusted parameters, and adjusted parameters with any visual representation dimension) have absolute values that are less than 1.96. Therefore, there is insufficient evidence to reject the null hypotheses that these four sets of collection rates are symmetrically distributed at a 5% level of significance. As well, the test statistics for the other two sets of collection rates (unadjusted parameters with 2½D or 3D representation) have absolute values that are more than 1.96. Therefore, there is sufficient evidence to reject the null hypotheses that these two sets of collection rates are symmetrically distributed at a 5% level of significance.

In short, it can be shown that the values in Table 10.18 do not meet the criterion for normality, and hence parametric tests cannot be used on them. As a remedy, Field (2006) suggests executing a logarithmic transformation on the data, before putting them through another cycle of the K-S, A-D and skewness tests. The transformed collection rates are summarised in Table 10.22.

Subject	2D		2½D		3D	
	Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
A	0.146	0.172	0.031	-0.038	-0.018	0.018
B	0.508	-0.024	0.542	0.154	0.398	0.134
C	0.118	0.054	0.092	-0.056	0.045	0.010
D	0.197	0.279	0.140	-0.016	0.110	-0.181
E	-0.071	-0.167	-0.301	-0.301	-0.535	-0.523
F	0.388	-0.039	-0.012	-0.134	-0.249	-0.047
G	0.199	0.192	0.091	0.176	-0.149	-0.155
H	-0.134	-0.342	-0.234	-0.258	-0.213	-0.115

Table 10.22: Hypothesis 3 and 6 – A summary of transformed collection rates

Likewise, a histogram is generated for each experimental condition in Table 10.22 to provide an initial indication of whether the transformed collection rates are normally distributed. The six histograms generated are shown in Figure 10.12. Unlike those in Figure 10.11, the histograms generated for the transformed data mostly have a closer resemblance to a bell shape now; two of them (2D representation with adjusted parameters, and 2½D representation with unadjusted parameters) still hint of bi-modality. To support this observation, the K-S, A-D and skewness tests were carried out, with the results summarised in Table 10.23 and Table 10.24 respectively. The values that were used to compute the test statistics for the skewness tests are available in Appendix D.2.

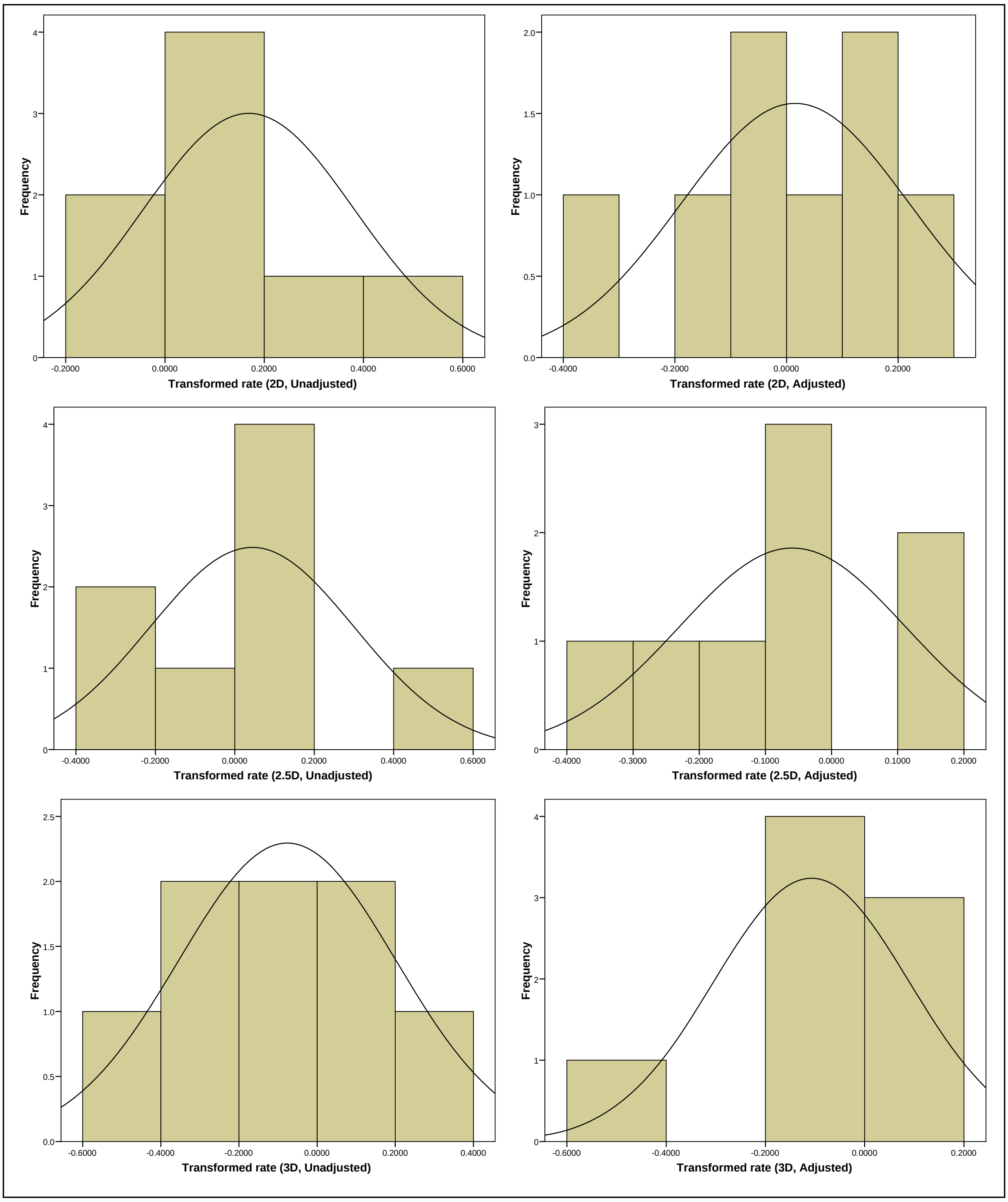


Figure 10.12: Hypothesis 3 and 6 – A summary of histograms generated for the transformed collection rates

	Kolmogorov-Smirnov(a)			Anderson-Darling	
	Statistic	df	Sig.	Statistic	Sig.
Transformed rate (2D, Unadjusted)	.194	8	.200(*)	.242	.666
Transformed rate (2D, Adjusted)	.153	8	.200(*)	.194	.839
Transformed rate (2½D, Unadjusted)	.229	8	.200(*)	.396	.281
Transformed rate (2½D, Adjusted)	.152	8	.200(*)	.248	.646
Transformed rate (3D, Unadjusted)	.143	8	.200(*)	.170	.896
Transformed rate (3D, Adjusted)	.229	8	.200(*)	.421	.240

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

Table 10.23: Hypothesis 3 and 6 – A summary of results from the Kolmogorov-Smirnov and Anderson-Darling tests performed on the transformed collection rates

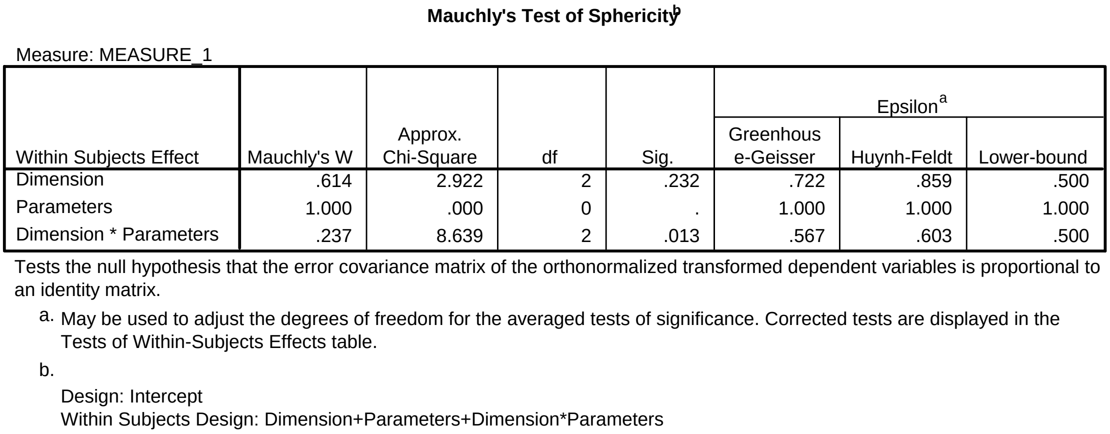
	Skewness	Standard Error	Test Statistic
Transformed rate (2D, Unadjusted)	0.157	0.752	0.209
Transformed rate (2D, Adjusted)	-0.546	0.752	-0.726
Transformed rate (2½D, Unadjusted)	0.705	0.752	0.938
Transformed rate (2½D, Adjusted)	0.011	0.752	0.015
Transformed rate (3D, Unadjusted)	0.096	0.752	0.128
Transformed rate (3D, Adjusted)	-1.356	0.752	1.803

Table 10.24: Hypothesis 3 and 6 – A summary of results from the skewness tests performed on the transformed collection rates

The p-values for all six sets of transformed collection rates in both tests in Table 10.23 are greater than 0.05. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of transformed collection rates are normally distributed at a 5% level of significance. In addition, the absolute values of the test statistics computed for the skewness tests in Table 10.24 are all less than 1.96. Therefore, there is insufficient evidence to reject the null hypotheses that these six sets of transformed collection rates are symmetrically distributed at a 5% level of significance. As such, it can be shown that the transformed values in Table 10.22 meet the criterion for normality, and hence parametric tests can be used on them.

Test of sphericity

Whilst the results above sanction the use of parametric ANOVAs on the transformed collection rate data, the latter still needs to be tested for sphericity. This test is required for deciding whether it is necessary to revise the critical values that are used for assessing the test statistics from the ANOVAs that follow. As outlined in the analytical framework, Mauchly's test is performed on the differences between the treatment levels of each possible main and interaction effect to assess the severity of departure from sphericity. The results from the series of Mauchly's tests performed are summarised in Table 10.25.



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Dimension	.614	2.922	2	.232	.722	.859	.500
Parameters	1.000	.000	0	.	1.000	1.000	1.000
Dimension * Parameters	.237	8.639	2	.013	.567	.603	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: Dimension+Parameters+Dimension*Parameters

Table 10.25: Hypothesis 3 and 6 – A summary of results from the Mauchly's tests performed on the relevant differences

The p-values for the visual representation dimension factor's main effect and the interaction effect between the latter and the model parameters factor are 0.232 and 0.013 respectively. On the one hand, as the p-value for the visual representation dimension factor's main effect is more than 0.05, there is insufficient evidence to reject the null hypothesis that the variances of the differences are not different at a 5% level of significance. This means that the criterion for sphericity is met, and the corresponding critical value for the following ANOVA needs not be revised. On the other hand, as the p-value for the interaction effect is less than 0.05, there is sufficient evidence to reject the null hypothesis that the variances of differences are not different at a 5% level of significance. This means that the criterion for sphericity is not met, and the corresponding critical value for the following ANOVA needs to be revised with a Greenhouse-Geisser correction (Field, 2006).

Nevertheless, since the model parameters factor has only two treatment levels, this implies that the sphericity criterion is not relevant. As such, the Mauchly's test of sphericity and its results for the model parameters factor in Table 10.25 are not used.

Test of main and interaction effects

The results from the two-way repeated measures ANOVA performed are summarised in Table 10.26. The output is split into sections that refer to the different effects and their associated error terms.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Sphericity Assumed	.272	2	.136	6.664	.009
	Greenhouse-Geisser	.272	1.443	.188	6.664	.020
	Huynh-Feldt	.272	1.718	.158	6.664	.014
	Lower-bound	.272	1.000	.272	6.664	.036
Error(Dimension)	Sphericity Assumed	.285	14	.020		
	Greenhouse-Geisser	.285	10.104	.028		
	Huynh-Feldt	.285	12.028	.024		
	Lower-bound	.285	7.000	.041		
Parameters	Sphericity Assumed	.110	1	.110	4.228	.079
	Greenhouse-Geisser	.110	1.000	.110	4.228	.079
	Huynh-Feldt	.110	1.000	.110	4.228	.079
	Lower-bound	.110	1.000	.110	4.228	.079
Error(Parameters)	Sphericity Assumed	.182	7	.026		
	Greenhouse-Geisser	.182	7.000	.026		
	Huynh-Feldt	.182	7.000	.026		
	Lower-bound	.182	7.000	.026		
Dimension * Parameters	Sphericity Assumed	.030	2	.015	1.331	.296
	Greenhouse-Geisser	.030	1.134	.027	1.331	.290
	Huynh-Feldt	.030	1.206	.025	1.331	.291
	Lower-bound	.030	1.000	.030	1.331	.286
Error(Dimension* Parameters)	Sphericity Assumed	.159	14	.011		
	Greenhouse-Geisser	.159	7.941	.020		
	Huynh-Feldt	.159	8.444	.019		
	Lower-bound	.159	7.000	.023		

Table 10.26: Hypothesis 3 and 6 – A summary of results from the two-way repeated measures ANOVA performed on the transformed collection rates

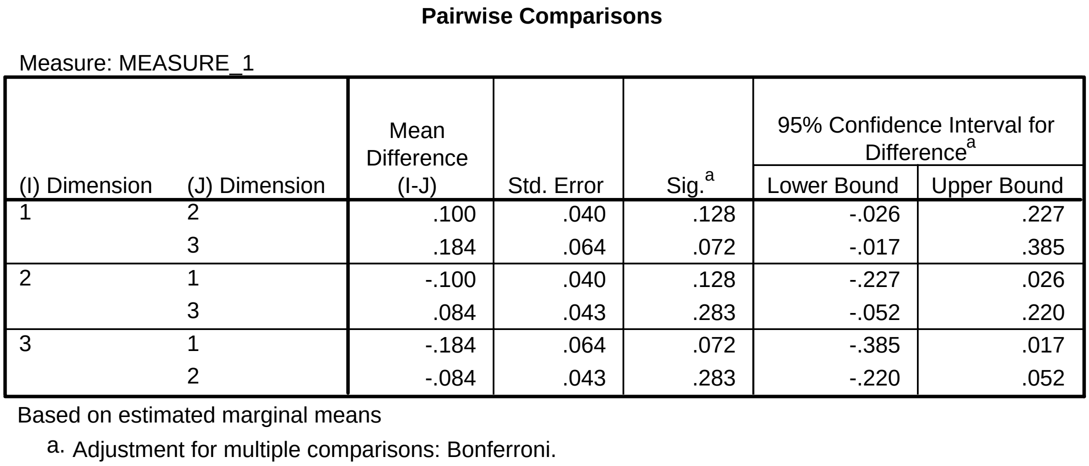
On the one hand, as the criterion for sphericity is met with respect to the visual representation dimension factor's main effect, the p-value that corresponds to 'Sphericity Assumed' in Table 10.26 is used. However, as the criterion for sphericity is not met in respect of the interaction effect, the p-value that corresponds to 'Greenhouse-Geisser' is used instead. The p-values for the main and interaction effects are 0.009 and 0.290 respectively. Since the former is less than 0.05, there is sufficient evidence to reject the null hypothesis that if type of model parameters used is ignored, using different types of visual representation dimension does not affect collection at a 5% level of significance. However, as the p-value for the interaction effect is more than 0.05, there is insufficient evidence to reject the null hypothesis that the effect which visual representation dimension (model parameters) has on collection rate is independent of the model parameters (visual representation dimension) associated with it at a 5% level of significance. That is, there is no interaction between the two factors in relation to collection rate.

On the other hand, as the criterion for sphericity is not relevant for the model parameters factor, there is only one p-value for its main effect in Table 10.26: 0.079. Since it is more than 0.05, there is insufficient evidence to reject the null hypothesis that if the type of visual representation dimension used is ignored, using different types of model parameters does not affect collection rate at a 5% level of significance.

Post-hoc test

It should be noted that Hypothesis 3 does not make any *a priori* predictions about the collection rates, and is only interested in exploring them for any differences due to the visual representation dimension used. Consequently, it is more appropriate to use a post-hoc test instead of a series of planned comparisons to investigate the visual representation dimension factor's main effect on collection rate.

Having determined that using different types of visual representation dimension affect collection rate, the next step is to identify the visual representation dimension that leads to a higher collection rate. The results from the pairwise comparisons performed on the visual representation dimension factor are summarised in Table 10.27.



Pairwise Comparisons

Measure: MEASURE_1

(I) Dimension	(J) Dimension	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.100	.040	.128	-.026	.227
	3	.184	.064	.072	-.017	.385
2	1	-.100	.040	.128	-.227	.026
	3	.084	.043	.283	-.052	.220
3	1	-.184	.064	.072	-.385	.017
	2	-.084	.043	.283	-.220	.052

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Table 10.27: Hypothesis 3 and 6 – A summary of results from the pairwise comparisons of the visual representation dimension factor

Using a Bonferroni-adjusted critical value that maintains the overall Type I error rate at a 5% level of significance, there are no significant results. Hence, there is insufficient evidence to reject the null hypotheses that all three pairs of treatment levels are not different in their effects on collection rate at an overall 5% level of significance. Unfortunately, this conclusion differs from the preceding conclusion based on the ANOVA results, which implies that there is at least a pair of treatment levels that are different in their effects on collection rate.

Nevertheless, if the overall level of significance for the two-tailed test is raised to at least 7.3%, then there is sufficient evidence to reject the null hypothesis that treatment level 1 and 3 are not different in their effects on collection rate. In such a case, the positive mean difference between treatment level 1 and 3 would suggest that the 2D representation leads to a higher collection rate than the 3D representation.

Effect size

Substituting the relevant values from the SPSS output[10] in Appendix E.3 into the expression for computing effect size,

$$r_{2D\,vs\,3D} = \sqrt{\frac{8.215}{8.215+7}} = 0.735$$

Since $r_{2D\,vs\,3D}$ is more than 0.50, it can be concluded that the effect between the 2D and 3D representations is large.

10.4.3 SUMMARY

The overarching pair of null and alternative hypotheses of interest are assessed by performing various tests prescribed by the analytical framework.

In relation to Hypothesis 3, there is sufficient evidence to reject the null hypothesis ($H_{3(0)}$) at a 5% level of significance, and conclude that the collection rate in a knowledge elicitation session is affected by the visual representation dimension used. Moreover, there is sufficient evidence to conclude that the 2D representation would lead to a higher collection rate than the 3D representation at an overall 7.3% level of significance. Further to this, it can be concluded that using the 2D representation over the 3D representation has a large effect on the collection rate. However, no significant difference in collection rate can be established between using the 2D and 2½D representations, and between the 2½D and 3D representations.

In relation to Hypothesis 6, there is insufficient evidence to reject the null hypothesis ($H_{6(0)}$) at a 5% level of significance. Hence, it can be concluded that the collection rate in a knowledge elicitation session is not affected by the model parameters used.

10.5 CONCLUSION

This chapter describes the analysis carried out to test Hypothesis 2 to 6. In a nutshell, these hypotheses postulate causal links between two factors (visual representation dimension and model parameters) and three constructs (state space, case quantity and collection rate). They are as summarised in Table 10.28.

Hypothesis	Cause	Effect
· 2 · 3	Visual representation dimension (2D, 2½D and 3D)	· Case quantity · Collection rate
· 4 · 5 · 6	Model parameters (Unadjusted and Adjusted)	· State space · Case quantity · Collection rate

Table 10.28: Postulated cause and effect relationships in Hypothesis 2 to 6

The analysis begins by exploring the measures determined for the various hypotheses. Then, following the analytical framework, a series of normality tests, sphericity tests, two-way repeated measures ANOVAs and post-hoc tests are executed on the measures to mixed results.

[10] The SPSS output is generated from performing appropriate planned contrasts.

In essence, the results reject the null hypothesis for Hypothesis 2, 3 and 4, and fail to reject the null hypothesis for Hypothesis 5 and 6. This means that state space has been found to be affected by the model parameters used. Also, both case quantity and collection rate are found to be affected by the visual representation dimension used. However, they are not affected by the model parameters used.

In the next chapter, the results from the entire data analysis process (process vii in Figure 4.1), which spans Chapter 9 and this chapter, are compiled and discussed (process viii). In addition, the limitations that were encountered throughout this research are reflected on. Last but not least, potential areas for further research are also explored.

11 Conclusion

An investigation was embarked upon to seek the answers to the research questions stated at the beginning of this thesis: Is VIS a valid tool for eliciting knowledge? If it is, how can it be adapted to make for a better knowledge elicitation tool?

In doing so, the knowledge elicitation process was formally defined (Chapter 2), and VIS was also established as a valid knowledge elicitation tool (Chapter 3). Following this, six propositions were suggested for leading the investigation on how to make a VIS-based knowledge elicitation tool better (Chapter 4). These were then used to specify the hypotheses for subsequent testing. Later, using a real-world case study set in a Ford engine assembly plant in Dagenham (East London), empirical work was planned (Chapter 5, 6, 7 and 8) and executed (Chapter 9 and 10).

The propositions and hypotheses are revisited briefly in the next section. Also, the results from the hypothesis tests are summarised and discussed. Further to this, the limitations that were encountered throughout the investigation are reflected on, before closing this thesis with some suggestions for future research.

11.1 FINDINGS FROM THE RESEARCH

Two research factors and four constructs have been determined in Chapter 4 for forming the propositions and specifying the hypotheses. The factors are the visual representation dimension and model parameters; and the constructs are decision fidelity, state space, case quantity and collection rate.

Three propositions have been suggested for investigating the effect of the visual representation dimension factor on decision fidelity, case quantity and collection rate. They are then used to frame the first three hypotheses (Hypothesis 1, 2 and 3) in this thesis. Similarly, three propositions have been suggested for investigating the effect of the model parameters factor on state space, case quantity and collection rate. They are also used to frame the next three hypotheses (Hypothesis 4, 5 and 6) in this thesis.

As it was identified that the visual representation dimension factor does not interact with the model parameters factor at all (Chapter 10), the findings for these factors can be reported separately from each other. Thus, the propositions, hypotheses and findings that pertain to the visual representation dimension factor will be addressed first, and then followed by those that pertain to the model parameters factor.

11.1.1 RESEARCH PROPOSITIONS, HYPOTHESES AND FINDINGS FOR THE VISUAL REPRESENTATION DIMENSION FACTOR

Proposition and Hypothesis One

Proposition 1 is put forward as below:

> *A higher dimension of iconic representation would demonstrably improve the degree of decision fidelity in the example cases collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{1(0)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{1(a)}$: The degree of decision fidelity in the example cases collected in a knowledge elicitation session improves as a higher dimension of visual representation is used.

Proposition and Hypothesis Two

Proposition 2 is put forward as below:

> *A higher dimension of iconic representation would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{2(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{2(a)}$: The size of case quantity of a knowledge elicitation session increases as a higher dimension of visual representation is used.

Proposition and Hypothesis Three

Proposition 3 is put forward as below:

> *Different dimension of iconic representation would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{3(0)}$: The collection rate in a knowledge elicitation session is not affected by the visual representation dimension used;

$H_{3(a)}$: The collection rate in a knowledge elicitation session is affected by the visual representation dimension used.

Findings for Hypothesis One, Two and Three – A summary and discussion

First and foremost, there is strong, albeit negative evidence to support the null hypothesis for Proposition 1 ($H_{1(0)}$). Moreover, the results also appear to contradict Proposition 1 by suggesting that decision fidelity improves as a lower dimension of visual representation is used. The latter is demonstrated most clearly by Subject G, when his decision-making demeanour in the knowledge elicitation session using the 2D game bore a very close resemblance to that in the real world (Section 9.2.2). Hence, *prima facie*, this finding goes against the belief that if the objects perceived in a real-world environment are simulated faithfully with similar visual and behavioural characteristics on a visual display, then an expert can apply the same mental processes which he uses for engaging the real-world objects onto his interactions with the simulation model (Section 4.3.1). As importantly, the apparent absence of high fidelity decision-making in the knowledge elicitation sessions implies that VIS is not a be-all and end-all tool for eliciting episodic knowledge. Subsequently, four factors that are thought to influence the extent of VIS' usefulness as a knowledge elicitation tool can be inferred from observing this phenomenon:

i. Replicability of real-world information

   VIS is more likely to be useful for eliciting knowledge if the contextual information that is vital for decision-making can be represented entirely and meaningfully by dynamic visual objects such as iconic animation and dynamically changing graphic (Section 4.3.1). In this research, however, there are pieces of information relayed through the sense of touch that cannot be represented meaningfully in the game models. This limitation might account for the discrepancy in decision fidelity and is mentioned again in Section 11.3.

ii. Alignment of the experts' motivation

   VIS is also more likely to be useful for eliciting knowledge when the motivation of the experts participating in the elicitation sessions is similar to their real-world motivation. This notion is not new, as Robinson *et al.* (2005) have discover that human decision makers are likely to act differently when there are no real consequences from their decisions made in a simulated environment (Section 1.1). In this research, there are a few (dis)incentives influencing the experts' decision-making behaviour in the real world that cannot be reproduced meaningfully in the game models. Hence, the experts' motivation during the elicitation sessions may differ from those they have in their real-world operations, which might explain the discrepancy in decision fidelity. This limitation is also mentioned again in Section 11.3.

iii. Tacitness of the experts' knowledge

VIS is probably more useful for eliciting knowledge that has not become too tacit in the experts' mind. This view is supported by comparing Subject G's experience level in the hot-test operations with the rest of the experts, and is discussed further in Section 11.4; and last but not least

iv. Consistency of the experts' decision-making

It is probable that the experts are not consistent in making their decisions, which may result in many noisy example cases being collected in the elicitation sessions and cause the phenomenon observed above. This drawback exposes a potentially serious weakness in VIS-based knowledge elicitation: the efficacy demonstrated by the latter as a computer-aided means to collect more example cases than other elicitation techniques does not necessarily imply there are as many valid example cases available for learning purpose. This strongly suggests a need to use other complementary elicitation techniques to filter out the noisy example cases, and facilitate a more robust approach to knowledge elicitation. These complementary techniques are likely to be manual methods (Section 2.4.2), which include document analysis, interview, on-site observation, questionnaire and rating scale, teach-back interview, protocol analysis, walkthrough, card-sort, and solution-characteristic matrix.

Secondly, there is sufficient evidence to reject the null hypothesis for Proposition 2 ($H_{2(0)}$) at a 5% level of significance. However, although it can be shown that case quantity is affected by the visual representation dimension used [$F(2,14) = 22.96$], it cannot be concluded that its size generally increases as a higher dimensional representation is used. Instead, it can only be concluded that using either the 2D or 3D representation leads to a significantly larger case quantity than the 2½D representation [$p_{2D\,vs\,2½D,\,one-tailed} < 0.05, r_{2D\,vs\,2½D} = 0.86$; $p_{2½D\,vs\,3D,\,one-tailed} < 0.05, r_{2½D\,vs\,3D} = 0.95$]. This finding shows that the properties that make a 3D-VIS model a better communication tool than an equivalent 2D-VIS model (Section 4.3.1) do not make the former a better knowledge elicitation tool. In addition, the finding above also forms an oblique contrast against Akpan and Brooks' (2005a) conclusion that it is easier to uncover inaccuracies in a 2½D-VIS or 3D-VIS model than in a 2D-VIS model (Section 4.3.1), by showing that it is easier for the latter to elicit responses from the experts than the former.

Last but not least, there is also sufficient evidence to reject the null hypothesis for Proposition 3 ($H_{3(0)}$) at a 5% level of significance. As such, it can be shown that collection rate is affected by the visual representation dimension used [$F(2,14) = 6.66$]. Furthermore, there is also sufficient evidence to conclude that the 2D representation leads to a significantly higher collection rate than the 3D representation [$p_{2D\,vs\,3D,\,one-tailed} < 0.05, r_{2D\,vs\,3D} = 0.74$].

In retrospect, it seems that the responses elicited through a 2D-VIS model are probably more realistic than those elicited through an equivalent 2½D-VIS or 3D-VIS model. Moreover, it also emerges that a 2D-VIS model is able to elicit significantly more responses from the experts than a 2½D-VIS model, and as many responses as a 3D-VIS model over a shorter time frame. Hence, notwithstanding the earlier concern over the responses' integrity and suitability for learning knowledge, it can be inferred from these findings that a 2D-VIS model generally makes for a better knowledge elicitation tool than a 2½D-VIS or 3D-VIS model does in the context studied.

11.1.2 RESEARCH PROPOSITIONS, HYPOTHESES AND FINDINGS FOR THE MODEL PARAMETERS FACTOR

Proposition and Hypothesis Four

Proposition 4 is put forward as below:

> *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the size of state space occupied by the example cases collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{4(0)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session is not affected by the model parameters used;

$H_{4(a)}$: The size of state space occupied by the example cases collected in a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

Proposition and Hypothesis Five

Proposition 5 is put forward as below:

> *Model parameters that are adjusted to develop more uncommon and extreme scenes would demonstrably increase the quantity of example cases collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{5(0)}$: The size of case quantity of a knowledge elicitation session is not affected by the model parameters used;

$H_{5(a)}$: The size of case quantity of a knowledge elicitation session increases as model parameters are adjusted to develop more uncommon and extreme scenes.

Proposition and Hypothesis Six

Proposition 6 is put forward as below:

> *Different sets of model parameters would have different impact on the efficiency with which the example cases are collected in a knowledge elicitation session.*

Therefore, the corresponding pair of null and alternative hypotheses are:

$H_{6(0)}$: The collection rate in a knowledge elicitation session is not affected by the model parameters used;

$H_{6(a)}$: The collection rate in a knowledge elicitation session is affected by the model parameters used.

Findings for Hypothesis Four, Five and Six – A summary and discussion

On the one hand, there is sufficient evidence to reject the null hypothesis for Proposition 4 ($H_{4(0)}$) at a 5% level of significance. Thus, it can be concluded that adjusting the model parameters to develop more uncommon and extreme scenes leads to a significantly larger state space than the original model parameters [$F(1,7) = 1{,}839.13$, $r_{Unadjusted\,vs\,Adjusted} = 1.00$] preset in the VIS model.

On the other hand, there is insufficient evidence to reject the null hypotheses for Proposition 5 and 6 ($H_{5(0)}$ and $H_{6(0)}$) at a 5% level of significance. Therefore, it can be concluded that case quantity and collection rate are not affected by the type of model settings used. These findings show that although the uncommon and extreme scenes provided in the simulation might offer the experts with more interesting situations that would retain their attention throughout the knowledge elicitation sessions, they did not materialise in more responses being elicited from the experts. Also, the experts did not find the unconventional decision-making required by these uncommon and extreme scenes more difficult, as they did not take more time to make the decisions.

In retrospect, it seems that the attribute elements of the example cases collected through a VIS model, which has been adjusted to develop more uncommon and extreme scenes, will collectively cover a wider range of values for all attributes than those from an unadjusted VIS model. Also, the latter can be achieved without any adverse impact on the quantity or rate of responses elicited from the experts. Hence, it can be inferred that an adjusted VIS model generally makes for a better knowledge elicitation tool than an unadjusted VIS model in the context studied.

All in all, a 2D-VIS model should always be chosen over a 2½D-VIS or 3D-VIS model to collect the example cases for machine learning in the context studied. Also, the chosen VIS model should always be adjusted to develop more uncommon and extreme scenes. Whilst a 2D representation does not interact with an adjusted set of model parameters to bring additional benefits into the knowledge elicitation efforts, the former will encourage a higher degree of decision fidelity in a larger set of example cases collected over a comparable period. In addition, the latter will also push for the decisions to be elicited over a wider range of situations.

11.2 CONTRIBUTIONS OF THE RESEARCH

The research initially seeks to answer two research questions: Is VIS a valid tool for eliciting knowledge? If it is, then how can VIS be adapted to make for a better knowledge elicitation tool?

The thesis commences by defining a context for the research, before attempting to establish VIS' validity as a knowledge elicitation tool. In doing so, it is realised that there is limited research on collecting an informative set of example cases for training a knowledge base. Liang *et al.* (1992) attribute this phenomenon to the flawed presumption that training example cases are either normally available, or easily collected. Unfortunately, this realisation inevitably implies that there is even less evidence for supporting VIS as a computer-aided means of collecting example cases.

Hence, the first contribution in this thesis comes from extracting and organising the circumstantial evidence in existing AI/KBS-Simulation/VIS collaboration literature that demonstrates VIS' suitability for such a purpose, and providing an affirmative answer to the first research question.

Following this, the thesis proceeds to set the scene for carrying out an experiment in order to answer the second research question. The experiment essentially investigates the effects (decision fidelity, state space, case quantity and collection rate) of using various VIS models with different levels of visual fidelity (2D, 2½D or 3D representations) and settings (unadjusted or adjusted model parameters) on the knowledge elicitation process. As Akpan and Brooks (2005a and b) recognise there are little or no empirical studies committed to comparing simulation models with 2D, 2½D or 3D representations, it is believed that this experiment constitutes the first empirical comparative study carried out on all three visual representations.

The experiment was carried out in Ford's engine assembly plant in Dagenham, East London. It involved eight real experts playing with VIS game models that had been adapted for six different experimental conditions (Section 6.1), and concluded with 48 sets of very rich data. Since the experts are actual decision makers who work in the real-world operations mimicked in the game models, their participation lend the data collected and any conclusions drawn from their analysis a rare quality of authenticity and legitimacy.

Further to this, the collected data are used to test six hypotheses (Section 4.4) based on the two factors of interest mentioned earlier: visual representation and model parameters. In the midst of doing so, a novel concept that brings together ideas from the fields of Descriptive Statistics, Geostatistics and Cluster Analysis is developed for evaluating decision fidelity in the data. Basically, it measures the dispersion of mixed multivariate data.

Last but not least, the most significant contribution from this thesis must lie with the conclusions drawn from the research, which provide an answer to the second research question. These conclusions, as detailed in Section 11.1 above, show the effects of different visual representation levels and model parameters on the effectiveness and efficiency of VIS for knowledge elicitation. Although the conclusions are specific to using VIS for collecting example cases to build a KBS via rule induction or pattern matching, they can also provide valuable insight into using VIS for aiding other techniques that elicit knowledge from example cases. These techniques include other machine learning methods such as neural network computing and genetic algorithms. Finally, the conclusions may offer an alternative perspective on using VIS as an aid for actual (as opposed to optimal) decision-making.

11.3 LIMITATIONS OF THE RESEARCH

Like all research, the investigation detailed in this thesis is not without its fair share of limitations. It is first mentioned that the order of experimental conditions, which each expert was exposed to, was not randomised sufficiently (Chapter 6). This subsequently led to a quasi-experiment being carried out in this research. Fortunately, the practice and fatigue effects that might have been developed from this lack of randomisation were mitigated by the randomising mechanism of the running game models, and the time lags built in between the knowledge elicitation sessions.

Next, as Field and Hole (2006) have pointed out, the small number of experts participating in this research might have posed a threat to its findings' validity. It is because using too few participants in an experiment will diminish its power to detect an effect in the sample that may actually exist in the population. In addition, O'Keefe and Pitt's (1991) comment that 'as in all laboratory experiments, the motivation of subjects is questionable' also rings true in this research. Whilst it could be observed that some experts genuinely wanted to help with the experiment, others were not as keen. Among the latter, they had started off being either curious about VIS modelling, or pressured by their peers to participate. Nonetheless, the influence due from the reservations above was allayed partially by adopting a repeated measures experimental design for the research, and exposing every expert to all experimental conditions. In so doing, the variation in scores between conditions that is due to the random differences between different participants is reduced dramatically. These differences include those pertaining to motivation, among others. Therefore, *ceteris paribus*, a repeated measures design will always be more sensitive than a non-repeated measures design (for instance, a between-groups design), and will be more likely to detect any differences that exist between conditions. For most of the data analysis (Chapter 9 and 10), the benefit from the repeated measures design's enhanced sensitivity indeed appears sufficient to offset the ramification from the loss in power due to a small sample and the experts' dubious motivation. This is because the conclusions drawn from the series of two-way repeated measures ANOVA tests and pairwise comparisons that were carried out for testing the hypotheses generally corroborate with each other. However, the same cannot be said for the borderline case of testing Hypothesis 3, where the conclusion drawn from the two-way repeated measures ANOVA test contradicts that drawn from the pairwise comparison at an identical significance level (Section 10.4.2). The perplexing contradiction in the latter could be avoided if more experts were available to participate in the experiment, which should increase the tests' statistical power to detect an effect and chance of reaching corroborating conclusions.

Furthermore, in spite of making every effort to ensure that the information offered in the game models match those present in the real world (Chapter 7), there are still some deviations. For instance, an expert may occasionally use the sense of touch to determine whether the engines have been tested. If an engine was tested, then it would feel hot. However, this information could not be incorporated into the game models realistically. In another instance, a third group of diesel engines with a capacity of 2.2 litres was introduced into the Puma engine assembly line mid-way through the experiment. Hence, the game models' currency was affected, albeit it might not alter the experts' decision-making demeanour which was used to evaluate Proposition 1.

Also, it was learnt that an expert will try to maintain the engine type that is being handled by each operational hot-test cell (Section 5.4). This is to minimise the amount of unproductive changeover time that is lost when a hot-test cell changes from testing one type of engines to another. As such, there will be more time to test the engines, thereby addressing the expert's objective to maximise the number of engines tested by the end of his shift. More importantly, this also serves to keep the hot-test operators happy as fewer changeovers imply less work; otherwise, the operators may lodge their complaints against the expert or even abuse him verbally. However, even though the latter is a very effective incentive for the expert to maintain the engine type in the real world, it could not be reproduced for the experiment. Consequently, the decisions made by the expert in the experiment may not be as faithful to those made in reality as originally thought.

Lastly, as Saunders *et al.* (2006) have warned, since the research is based on a case study, the findings that are made in this thesis are specific to the context studied and may not be generalised to other research settings. They can, however, act as indicators for those who are researching in this domain and applying these ideas.

11.4 SUGGESTIONS FOR FUTURE RESEARCH

Several areas with potential for future research have been identified during the investigation detailed in this thesis. Their origins range from the literature review at the beginning of the thesis, to the findings at the end.

O'Keefe and Pitt (1991) show that the preference for a visual representation mode (Section 4.3.1) can be explained partially by cognitive style. The latter can be measured with well-established instruments such as the Myers-Briggs Type Indicator (Myers, 1977). Thus, following their lead, a probable avenue for future research lies in studying if and how an expert's cognitive style may influence the effectiveness of a visual representation dimension on VIS-based knowledge elicitation, or even VIS-based decision-making.

Next, it is noted in Figure 9.2 that Subject G's decision-making demeanour in the 2D game bears an uncanny resemblance to that in the real world. The significance of this observation was later confirmed by the χ^2 test results in Table 9.9. However, the same cannot be said for Subject A, B, C or H. A review of each expert's profile (Table 5.3) reveals that apart from Subject G's prior experience of using the computers and game consoles, he also has the least experience in switch operations (< 1 year). Perhaps it may be suggested that Subject G's relative inexperience means his knowledge of switch operations has not become too tacit for VIS-based elicitation. If this is true, then another probable avenue for future research lies in determining the correct means (Section 2.4) for eliciting knowledge of varying levels of tacitness.

Further to this, considering that the experts are better placed to process the information from the 3D-VIS model, which has a higher visual fidelity and runs slower than the 2D-VIS model (Section 4.3.1), the findings for Proposition 2 and 3 suggest that a higher visual fidelity may actually make it more difficult for the experts to process information. Hence, if technological advances in the future can improve the 3D-VIS model's run speed to be on par with a 2D-VIS model's, then a similar experiment can be performed to find out whether case quantity will deteriorate as predicted by this deduction. Moreover, taking into account the contrasting views that a 3D-VIS model is a better communication tool than a 2D-VIS model, and also that it is easier to uncover inaccuracies in a 2½D-VIS or 3D-VIS model than in a 2D-VIS model (Akpan and Brooks, 2005a and b), it may be theorised that the mental processes used by the experts for understanding and validating the VIS models are different from those used for responding to VIS models. This may provide another basis for future research.

As well, a speculation may be made from a confluence of the findings from Proposition 2 and 5. It is first found that no differences can be detected between the case quantity from a 2D-VIS model, and a slower-running 3D-VIS model which has a higher visual fidelity. Then, it is found that case quantity is also not affected by the range of scenes that are developed in the VIS model. Therefore, it might be that there is a constraint on the quantity of example cases which can be collected in a knowledge elicitation session, and it is likely to be imposed by some inherent factors in the expert. These factors may include an onset of mental fatigue or limited attention span, and understanding their influence on the expert's participation may provide another avenue for future research.

Last but not least, it is worthwhile to extend this investigation on VIS-based knowledge elicitation by repeating this experiment in various case studies involving different contexts and hopefully, even more experts. By doing so, the findings from these case studies might be threaded together to premise a more general and valid conclusion. At this point, it is helpful to establish the contexts where VIS-based knowledge elicitation is expected to be useful, and also discuss the adaptations that might be needed when adopting the methodology used in this research. Using Mintzberg's (1973) classification of management roles in Table 5.1, VIS is deemed useful for eliciting knowledge from real-world decision-makers whose responsibilities include handling disturbances and/or allocating resources, and in contexts where information that is vital for decision-making can be represented entirely and meaningfully by dynamic visual objects such as iconic animation and dynamically changing graphic (Section 4.3.1). An apt example would be the case study used by Robinson *et al.* (2005), in which they investigated how decision-makers in an engine assembly plant actually use various information (such as manpower availability, expected time required to remedy the machine breakdowns and dependence of the plant's operations on the broken machines) to prioritise the repair and maintenance needs of its operations.

In general, the methodology outlined in Section 4.5 can be applied to the contexts described above after making a few sensible adaptations. These adaptations are normally expected in the preliminary processes of the methodology as they help lay the context-specific groundwork for VIS-based knowledge elicitation and the experiment. They are: understanding the case study, and designing the experiment. The methodology begins with gaining an in-depth appreciation of the case study by using a suitable selection of complementary elicitation techniques documented in Section 2.4. The selection that is used eventually is expected to differ from that used in this research, as it depends on the information that is needed and the resources that are made available to the knowledge engineer in the context studied. Following this, the methodology continues with designing an efficient experiment and planning a bias-free elicitation schedule. Notwithstanding the actual number of participants in the experiment, a repeated measures experimental design should always be used when feasible, as it is more sensitive and likely to detect any differences that exist between experimental conditions than other designs. Finally, the order by which the experts are exposed to the different experimental conditions in the elicitation schedule should be randomised as far as possible. Nevertheless, it is acknowledged that the latter is not always possible (as shown in this research) since it is contingent on the flexibility of the experts' availability. In this case, mitigating measures that are specific to the context studied should be implemented to reduce any systematic effects that might arise.

The End

REFERENCES

Akpan, J.I. (2005) *An Empirical Study of the Impact of Virtual Reality on Discrete-Event Simulation*, PhD Thesis, Lancaster University, United Kingdom.

Akpan, J.I. and Brooks, R.J. (2005a) 'Experimental investigation of the impacts of virtual reality on discrete-event simulation', *in* Kuhl, M.E., Steiger, N.M., Armstrong, F.B. and Joines, J.A. (eds), *Proceedings of the 2005 Winter Simulation Conference*, pp. 1968-1975.

Akpan, J.I. and Brooks, R.J. (2005b) 'Practitioners' perception of the impacts of virtual reality on discrete-event simulation', *in* Kuhl, M.E., Steiger, N.M., Armstrong, F.B. and Joines, J.A. (eds), *Proceedings of the 2005 Winter Simulation Conference*, pp. 1976-1984.

Alifantis, T. (2006) *Knowledge Based Improvement: Simulation and Artificial Intelligence for Understanding and Improving Decision Making in an Operations System*, PhD Thesis, University of Warwick, United Kingdom.

Bachi, R. (1962) 'Standard distance measures and related methods for spatial analysis', *Papers in Regional Science*, 10:1, pp. 83-132.

Baines, T.S. and Kay, J.M. (2002) 'Human performance modelling as an aid in the process of manufacturing system design: A pilot study', *International Journal of Production Research*, 40:10, pp. 2321-2334.

Banks, J., Carson II, J.S., Nelson, B.L. and Nicol, D.M. (2005) *Discrete-Event System Simulation* (4th edn), Prentice Hall.

Barfield, W. and Furness, T.A. (1995) *Virtual Environments and Advanced Interface Design*, Oxford University Press.

Barnes, M. (1996) 'Virtual reality and simulation', *in* Charnes, J.M., Morrice, D.J., Brunner, D.T. and Swain, J.J. (eds), *Proceedings of the 1996 Winter Simulation Conference*, pp. 101-110.

Barnes, M.R. (1997) 'An introduction to QUEST', *in* Andradóttir, S., Healy, K.J., Withers, D.H. and Nelson, B.L. (eds), *Proceedings of the 1997 Winter Simulation Conference*, pp. 619-623.

Barrett, A.R. and Edwards, J.S. (1995) 'Knowledge elicitation and knowledge representation in a large domain with multiple experts', *Expert Systems with Applications*, 8:1, pp.169-176.

Bell, P.C. and O'Keefe, R.M. (1994) 'Visual interactive simulation: A methodological perspective', *Annals of Operations Research*, 53, pp. 321-342.

Bell, P.C. and O'Keefe R.M. (1995) 'An experimental investigation into the efficacy of visual interactive simulation', *Management Science*, 41:6, pp. 1018-1038.

Boddy, D. (2005) *Management: An Introduction* (3rd edn), Prentice Hall.

Breuker, J. and Wielinga, B. (1987) 'Use of models in the interpretation of verbal data', *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 17-44.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth International.

Buchanan, B.G., Barstow, D., Bechtal, R., Bennett, J., Clancey, W., Kulikowski, C., Mitchell, T. and Waterman, D.A. (1983) 'Constructing an expert system', *in* Hayes-Roth, F., Waterman, D.A. and Lenat, D.B. (eds), *Building Expert Systems*, Addison-Wesley Publishing Company, pp. 127-167.

Byrd, T.A. (1995) 'Expert systems implementation: Interviews with knowledge engineers', *Industrial Management and Data Systems*, 95:10.

Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research*, Houghton Mifflin Company.

Caprihan, R., Kumar, A. and Stecke, K.E. (2006) ‘A fuzzy dispatching strategy for due-date scheduling of FMSs with information delays’, *International Journal of Flexible Manufacturing Systems*, 18, pp. 29-53.

Chau, P.Y.K. and Bell, P.C. (1995) ‘Designing effective simulation-based decision support systems: An empirical assessment of three types of decision support systems’, *Journal of the Operational Research Society*, 46, pp. 315-331.

Chryssolouris, G., Lee, M., Domroese, M. (1991) ‘The use of neural networks in determining operational policies for manufacturing systems’, *Journal of Manufacturing Systems*, 10:2, pp. 166-175.

Clancey, W.J. (1986) ‘Cognition and expertise’, *1st AAAI Workshop: Knowledge Acquisition in Knowledge Based Systems*, Canada.

Clustan (2007) ‘Welcome to Clustan’ [online] (cited 26 June 2007) Available from <URL:http://www.clustan.com>.

Coffey, J.W. and Hoffman, R.R. (2003) ‘Knowledge modelling for the preservation of institutional memory’, *Journal of Knowledge Management*, 7:3, pp. 38-52.

Computerworld (2002) ‘QuickStudy: System Development Life Cycle’ [online] (cited 22 April 2007) Available from <URL:http://www.computerworld.com/developmenttopics/development/story/0,10801,71151,00.html>.

Cordingley, E.S. (1989) 'Knowledge elicitation techniques for knowledge-based systems', *in* Diaper, D. (ed), *Knowledge Elicitation: Principles, Techniques and Applications*, Ellis Horwood Limited, pp. 89-175.

Cunningham, P. (1998) *11th IEA Conference*, Berlin: Springer.

D'Agostino, R.B. and Stephens, M.A. (1986) *Goodness-of-fit Techniques*, Marcel Dekker.

Darlington, K. (2000) *The Essence of Expert Systems*, Prentice Hall.

Davis, R. (1985) 'Interactive transfer of expertise', *in* Buchanan, B.G. and Shortliffe, E.H. (eds), *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley Publishing Company, pp. 171-205.

Dennis, A. and Wixom, B.H. (2003) *Systems Analysis and Design* (2nd edn), John Wiley and Sons.

Duan, Y. and Burrell, P. (1995) 'A hybrid system for strategic marketing planning', *Marketing Intelligence & Planning*, 13:11, pp. 5-12.

Edwards, J.S. (1991) *Building Knowledge-based Systems: Towards a Methodology*, Pitman.

Feigenbaum, E.A. (1980) *Knowledge Engineering in the 1980's*, Dept. of Computer Science, Stanford University, Stanford, CA.

Field, A. (2006) *Discovering Statistics Using SPSS* (2nd edn), Sage Publications.

Field, A. and Hole, G. (2006) *How to Design and Report Experiments*, Sage Publications.

Fildes, J. and Ranyard, J.C. (1997) 'Success and survival of operational research groups – A review', *Journal of the Operational Research Society*, 48:4, pp. 336-360.

Firlej, M. and Hellens, D. (1991) *Knowledge Elicitation: A Practical Handbook*, Prentice Hall.

Flitman, A.M. and Hurrion, R.D. (1987) 'Linking discrete event simulation models with expert systems', *Journal of the Operational Research Society*, 38:8, pp. 723-733.

Fox, J., Myers, C.D., Greaves, M.F. and Pegram, S. (1987) 'A systematic study of knowledge base refinement in the diagnosis of leukaemia', *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 73-90.

Gammack, J.G. (1987) ‘Different techniques and different aspects on declarative knowledge’, *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 137-163.

Gower, J.C. (1971) ‘A general coefficient of similarity and some of its properties’, *Biometrics*, 27, pp. 857-872.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., and Tatham, R.L. (2006) *Multivariate Data Analysis* (6th edn), Pearson Prentice Hall.

Hart, A. (1987) ‘Induction and knowledge elicitation’, *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 165-189.

Hayes-Roth, F., Waterman, D.A. and Lenat, D.B. (1983) ‘An overview of expert systems’, *in* Hayes-Roth, F., Waterman, D.A. and Lenat, D.B. (eds), *Building Expert Systems*, Addison-Wesley Publishing Company, pp. 3-29.

Hines, W.W., Montgomery, D.C., Goldsman, D.M. and Borror, C.M. (2003) *Probability and Statistics in Engineering* (4th edn), John Wiley and Sons.

Hollocks, B.W. (2006) ‘Forty years of discrete-event simulation – A personal reflection’, *Journal of the Operational Research Society*, 57:12, pp. 1383-1399.

Howell, D.C. (2007) *Statistical Methods for Psychology* (6th edn), Thomson Wadsworth.

Hurrion, R.D. (1976) *The Design, Use and Required Facilities of an Interactive Visual Computer Simulation Language to Explore Production Planning Problems*, PhD Thesis, University of London.

Hurrion, R.D. (1980) ‘An interactive visual simulation system for industrial management’, *European Journal of Operational Research*, 5, pp. 86-93.

Hurrion, R.D. (1986) ‘Visual interactive modelling’, *European Journal of Operational Research*, 23, pp. 281-287.

Hurrion, R.D. (1991) ‘Intelligent visual interactive modelling’, *European Journal of Operational Research*, 54, pp. 349-356.

Jackson, P. (1985) ‘Reasoning about belief in the context of advice-giving systems’, *in* Bramer, M.A. (ed), *Research and Development in Expert Systems*, Cambridge University Press.

Jackson, P. (1999) *Introduction to Expert Systems* (3rd edn), Addison-Wesley.

Jeffrey, P. and Seaton, R. (1995) ‘The use of operational research tools: A survey of operational research practitioners in the UK’, *Journal of the Operational Research Society*, 46:7, pp. 797-808.

Jeong, K.Y. (2000) ‘Conceptual frame for development of optimised simulation-based scheduling systems’, *Expert Systems with Applications*, 18, pp. 299-306.

Jeong, K.C. and Kim, Y.D. (1998) ‘A real-time scheduling mechanism for a flexible manufacturing system: Using simulation and dispatching rules’, *International Journal of Production Research*, 36:9, pp. 2609-2626.

Johannsen, G. (1989) ‘Knowledge analysis in power plants’, *in* Singh, M. G. (ed.), *Systems and Control Encyclopaedia, First Supplement*, Pergamon Press, pp. 366-373.

Johannsen, G. and Alty, J.L. (1991) ‘Knowledge engineering for industrial expert systems’, *Automatica*, 27:1, pp. 97-114.

Johnson, L. and Johnson, N.E. (1987) ‘Knowledge elicitation involving teachback interviewing’, *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 91-108.

Kelly, G.A. (1955) *The Psychology of Personal Constructs*, Norton.

Kidd, A. and Welbank, M. (1984) ‘Knowledge acquisition’, *in* Fox, J. (ed), *Expert Systems*, Infotech State of the Art Report, Pergamon Infotech Ltd.

Kowalski, R. and Westen, D. (2005) *Psychology* (4th edn), John Wiley and Sons.

Krzanowski, W.J. (2005) *Principles of Multivariate Analysis*, Oxford University Press.

Kunnathur, A.S., Sundararaghavan, P.S. and Sampath, S. (2004) ‘Dynamic rescheduling using a simulation-based expert system’, *Journal of Manufacturing Technology Management*, 15:2, pp. 199-212.

Lanner (2007) ‘WITNESS Suite’ [online] (cited 28 June 2007) Available from <URL:http://www.lanner.com/en/simulation_professionals/witness_vr.php>.

Liang, T.P. (1992) ‘A composite approach to inducing knowledge for expert systems design’, *Management Science*, 38:1, pp. 1-17.

Liang, T.P., Moskowitz, H. and Yih, Y. (1992) ‘Integrating neural networks and semi-markov processes for automated knowledge acquisition: An application to real-time scheduling’, *Decision Sciences*, 23:6, pp. 1297-1313.

Lyu, J. and Gunasekaran, A. (1997) ‘An intelligent simulation model to evaluate scheduling strategies in a steel company’, *International Journal of Systems Science*, 28:6, pp. 611-616.

Madni, A.M. (1988) ‘The role of human factors in expert systems design and acceptance’, *Human Factors*, 30, pp. 395-414.

Mak, R.W.T., Gupta, S.M. and Lam, K. (2002) 'Modelling of material handling hoist operations in a PCB manufacturing facility', *Journal of Electronics Manufacturing*, 11:1, pp. 33-50.

McDermott, J. (1983) *Building Expert Systems*, CMU report, Carnegie-Mellon University, Pittsburgh.

Moody, J.W., Blanton, J.E. and Will, R.P. (1998) 'Capturing expertise from experts: The need to match knowledge elicitation techniques with expert system types', *Journal of Computer Information Systems*, 39:2, pp. 89-95.

Motulsky, H.J. (1999) *Analysing Data with GraphPad Prism*, GraphPad Software Inc, San Diego.

Mintzberg, H. (1973) *The Nature of Managerial Work*, Harper & Row.

Myers, I.B. (1977) *The Myers-Briggs Type Indicator*, Consulting Psychologists Press.

Naylor, J.B., Griffiths, J. and Naim, M.M. (2001) 'Knowledge-based system for estimating steel plant performance', *International Journal of Operations & Production Management*, 21:7, pp. 1000-1019.

Negnevitsky, M. (2005) *Artificial Intelligence: A Guide to Intelligent Systems* (2nd edn), Addison Wesley.

Newell, A. (1969) ‘Heuristic programming : Ill-structured problems’, *in* Aronofsky, A. (ed), *Progress in Operations Research Vol 3*, John Wiley and Sons, pp. 360-414.

Newell, A. and Simon, H.A. (1972) *Human Problem Solving*, Prentice Hall.

O’Brien, J.J. and Griffiths, J.F. (1967) ‘Choosing a test of normality for small samples’, *Meteorology and Atmospheric Physics*, 16:2-3, pp. 267-272.

Ogborn, J.M. and Johnson, L. (1984) ‘Conversation theory’, *Kybernetes*, 13, pp. 177-181.

O’Keefe, R.M. (1986) ‘Simulation and expert systems – A taxonomy and some examples’, *Simulation*, 46:1, pp. 10-16.

O’Keefe, R.M. (1989) ‘The role of artificial intelligence in discrete-event simulation’, *in* Widman, L.E., Loparo, K.A. and Neilsen, N.R. (eds), *Artificial Intelligence, Simulation and Modelling*, Wiley, pp. 359-379.

O’Keefe, R.M. and Bell, P.C. (1992) ‘Findings from behavioural research in visual interactive simulation’, *in* Swain, J.J., Goldsman, D., Crain, R.C. and Wilson, J.R. (eds), *Proceedings of the 1992 Winter Simulation Conference*, pp. 751-755.

O’Keefe, R.M. and Pitt, I.L. (1991) ‘Interaction with a visual interactive simulation, and the effect of cognitive style’, *European Journal of Operational Research*, 54, pp. 339-348.

Ören, T.I. (1994) 'Artificial intelligence in simulation', *Annals of Operations Research*, 53, pp. 287-319.

Pidd, M. (2003) *Tools for Thinking* (2nd edn), John Wiley and Sons.

Pidd, M. (2005) *Computer Simulation in Management Science* (5th edn), John Wiley and Sons.

Pierreval, H. and Ralambondrainy, H. (1990) 'A simulation and learning technique for generating knowledge about manufacturing systems behaviour', *Journal of the Operational Research Society*, 41:6, pp. 461-474.

Powell, S.G. (1995) 'The teachers' forum: Six key modelling heuristics', Interfaces, 25:4, pp. 114-125.

Preece, J. (1994). *Human-Computer Interaction*, Addison-Wesley.

Raghunathan, S. and Tadikamalla, P. (1992) 'The use of stochastic simulation in knowledge-based systems', *Decision Sciences*, 23:6, pp. 1333-1356.

Rehn, G.D., Lemessi, M., Vance, J.M. and Dorozhkin, D.V. (2004) 'Integrating operations simulation results with an immersive virtual reality environment', *in* Ingalls, R.G., Rossetti, M.D., Smith, J.S. and Peters, B.A. (eds), *Proceedings of the 2004 Winter Simulation Conference*, pp. 1713-1719.

Robinson, S. (1994) *Successful Simulation: A Practical Approach to Simulation Projects*, McGraw-Hill.

Robinson, S. (2003) 'Modelling human decision-making', *in* Al-Dabass, D. (ed), *Proceedings of the 17th European Simulation Multiconference*, SCS, Delft, pp. 448-455.

Robinson, S. (2004) *Simulation: The Practice of Model Development and Use*, John Wiley and Sons.

Robinson, S. (2005) 'Discrete-event simulation: From the pioneers to the present, what next?', *Journal of the Operational Research Society*, 56:6, pp. 619-629.

Robinson, S. (2007) 'Modelling human interaction in organisational systems', *in* Fishwick, P.A. (ed), *Handbook of Dynamic System Modelling*, CRC Press.

Robinson, S., Edwards, J.S., Yongfa, W. (1998) 'An expert systems approach to simulating the human decision maker', *in* Medeiros, D.J., Watson, E.F., Manivannan, M. and Carson, J. (eds), *Proceedings of the 1998 Winter Simulation Conference*, San Diego, CA: The Society for Computer Simulation, pp. 1541-1545.

Robinson, S., Alifantis, T., Edwards, J.S., Hurrion, R.D., Ladbrook, J. and Waller, T. (2001) 'Modelling and improving human decision making with simulation', *in* Peters, B.A., Smith, J.S., Medeiros, D.J. and Rohrer, M.W. (eds), *Proceeding of the 2001 Winter Simulation Conference*, Piscataway, New Jersey: Institute of Electrical and Electronic Engineers, pp. 913-920.

Robinson, S., Alifantis, T., Edwards, J.S., Ladbrook, J. and Waller, T. (2005) 'Knowledge based improvement: Simulation and artificial intelligence for identifying and improving human decision-making in an operations system', *Journal of the Operational Research Society*, 56:8, pp. 912-921.

Rogerson, P.A. (2006) *Statistical Methods for Geography* (2nd edn), Sage Publications.

Royce, W.W. (1970) *Managing the Development of Large Software Systems*, IN Proc. WESTCON, San Francisco.

Rugg, G., McGeorge, P. and Maiden, N. (2000) 'Method fragments', *Expert Systems*, 17:5, pp. 248-257.

Rugg, G., Eva, M., Mahmood, A., Rehman, N., Andrews, S. and Davies, S. (2002) 'Eliciting information about organisational culture via laddering', *Information Systems Journal*, 12, pp. 215-229.

Saunders, M., Lewis, P. and Thornhill, A. (2006) *Research Methods for Business Students* (4th edn), Prentice Hall.

Shaw, M.J. (1989) 'A pattern-directed approach to flexible manufacturing: A framework for intelligent scheduling, learning, and control', *International Journal of Flexible Manufacturing Systems*, 2, pp. 121-144.

Shaw, M.L.G. and Gaines, B.R. (1987) 'An interactive knowledge elicitation technique using personal construct technology', *in* Kidd, A.L. (ed), *Knowledge Acquisition for Expert Systems: A Practical Handbook*, Plenum Press, pp. 109-136.

Sherman, W.R. and Craig, A.B. (2003) *Understanding Virtual Reality: Interface, Application, and Design*, Morgan Kaufmann Publishers.

Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences* (2nd edn), McGraw-Hill Book Company.

Stuart, R. (2001) *The Design of Virtual Environments*, Barricade Books.

Tan, A.A.W. (2003) *The Reliability and Validity of Interactive Virtual Reality Computer Experiments*, PhD Thesis, Technische Universiteit Eindhoven, The Netherlands.

Tan, A.A.W., Timmermans, H.J.P. and de Vries, B. (2000) 'Investigation of human behaviour in a virtual environment', *in Proceedings of the VWsim'00 Conference, 2000 Virtual Worlds and Simulation Conference*, San Diego, California.

Toothaker, L.E. and Newman, D. (1994) 'Nonparametric competitors to the two-way ANOVA', *Journal of Educational and Behavioural Statistics*, 19:3, pp. 237-273.

Turban, E., Aronson, J.E. and Liang, T.P. (2005) *Decision Support Systems and Intelligent Systems* (7th edn), Prentice Hall.

Vince, J. (1998) *Essential Virtual Reality Fast: How to Understand the Techniques and Potential of Virtual Reality*, Springer-Verlag.

Waller, A.P. and Ladbrook, J. (2002) 'Experiencing virtual factories of the future', *in* Yücesan, E., Chen, C.H., Snowdon, J.L. and Charnes, J.M. (eds), *Proceedings of the 2002 Winter Simulation Conference*, pp. 513-517.

Waterman, D.A. (1986) *A Guide to Expert Systems*, Addison-Wesley Publishing Company.

Weitzel, J.R. and Kerschberg, L. (1989a) 'Developing knowledge-based systems: Reorganising the system development life cycle', *Communications of the ACM*, 32:4, pp. 482-488.

Weitzel, J.R. and Kerschberg, L. (1989b) 'A system development methodology for knowledge-based systems', *IEEE Transactions on Systems, Man, and Cybernetics*, 19:3, pp. 598-605.

Wenzel, S., Bernhard, J. and Jessen, U. (2003) ‘A taxonomy of visualisation techniques for simulation in production and logistics’, *in* Chick, S., Sánchez, P.J., Ferrin, D. and Morrice, D.J. (eds), *Proceedings of the 2003 Winter Simulation Conference*, pp. 729-736.

Williams, T. (1996) ‘Simulating the man-in-the-loop’, *OR Insight*, 9:4, pp. 17-21.

Willemain, T.R. (1994) ‘Insights on modelling from a dozen experts’, *Operations Research*, 42:2, pp. 213-222.

Wishart, D. (2001) ‘k-means clustering with outlier detection, mixed variables and missing values’, *in* Schwaiger, M. and Opitz, O. (eds), *Exploratory Data Analysis in Empirical Research*, pp. 216-226, Springer-Verlag.

Wishart, D. (2006) *ClustanGraphics Primer* (4th edn), Clustan, Edinburgh.

Wu, S.Y.D. and Wysk, R.A. (1990) ‘An inference structure for the control and scheduling of manufacturing systems’, *Computers and Industrial Engineering*, 18:3, pp. 247-262.

APPENDICES

A Pre-Experiment Questionnaire

The questionnaires that were used to establish the experts' profiles in Table 5.3 are provided below.

A.1 ABOUT YOURSELF

i. Name : ____________________

ii. Age (years) : □ 30 – 39 □ 40 – 49 □ 50 – 59 □ 60 – 69

iii. Experience as Switch Operator (years) : ____________

		Very familiar				**Not familiar**
		5	**4**	**3**	**2**	**1**
iv.	Please rate your familiarity with computers.	□	□	□	□	□
v.	Please rate your familiarity with video game consoles.	□	□	□	□	□

A.2 LEARNING STYLE

Please tick the option that best describes you.

		Most like me				**Least like me**
		5	**4**	**3**	**2**	**1**
i.	I remember what I see better than what I hear.	□	□	□	□	□
ii.	I understand information by visualising pictures.	□	□	□	□	□
iii.	I use different colours to highlight, select and organise when writing or reading.	□	□	□	□	□
iv.	I make notes using drawings, spacing, symbols etc.	□	□	□	□	□
v.	I recall written information by visualising text pages, notes or study cards.	□	□	□	□	□
vi.	I would rather read a story than listen to it.	□	□	□	□	□
vii.	I understand numeric problems that are written better than those that I hear.	□	□	□	□	□
viii.	A graph of numbers is easier for me to understand than written numbers.	□	□	□	□	□
ix.	Written numeric problems are easier for me to solve than oral ones.	□	□	□	□	□
x.	I learn better by reading than listening.	□	□	□	□	□
xi.	Seeing a number makes more sense to me than hearing a number.	□	□	□	□	□
xii.	I use visual cues to recall information.	□	□	□	□	□

B A Comprehensive Overview of the Example Case

An example case is a unit of episodic knowledge, a collection of which is recorded during each knowledge elicitation session conducted for the research. Every example case recorded describes the scene when the expert interacted with the Visual Interactive Simulation (VIS) model, and is made up of two parts: a decision element and an attribute element. Briefly, the decision element is a set of decisions made by the expert when he interacts with the VIS model, whilst the attribute element is a corresponding set of attributes that describes the state in the VIS model when the interaction takes place. These elements and their relationship are described in more detail in Section 5.4.

On the one hand, the decision element is recorded as a set of 21 binary variates, which represents a set of 21 decisions made by the expert when he interacted with the VIS model. These decisions include selecting the conveyor path for newly assembled engines that are entering the hot-test operations as well as for untested engines that have been rerouted in the hot-test operations (one variate), and switching on/off each of the 20 hot-test cells/stands in the hot-test operations (20 variates).

On the other hand, the attribute element is recorded as a set of 551 variates and contains values measured on a mixture of binary, nominal and ratio scales. Together, these values represent the attributes that describe the state of the VIS model when the expert interacted with it, and are believed to influence the decisions made by him during the interaction. The attributes are:

i. Quantity of 2*l* engines on each section of conveyor (87 variates);

ii. Quantity of 2.4*l* engines on each section of conveyor (87 variates);

iii. Quantity of repaired engines on each section of conveyor (87 variates);

iv. Quantity of empty platens on each section of conveyor (87 variates);

v. Type of engine (2*l*, 2.4*l*, faulty or empty platen) currently in each hot-test cell (80 variates);

vi. Type of engine (2*l*, 2.4*l*, faulty or empty platen) currently parked on each waiting stand (80 variates);

vii. Operational status of each hot-test cell (20 variates);

viii. Quantity of engines tested in each hot-test cell (20 variates);

ix. Total quantity of engines tested in the hot-test operations (1 variate); and

x. Shift period/break when an interaction takes place (2 variates).

Nonetheless, many variates are later found to be redundant and removed. In addition, several variates are actually components of various attributes or cover large ranges of values, and they are recoded or rescaled respectively. The number of variates in each attribute element is reduced to 184 eventually. The work carried out to clean and consolidate these variates is described in more detail in Section 10.2.

C A Framework for Data Analysis

A repeated measures experimental design was used in this research, whereby two independent factors of interest (visual representation dimension, and model parameters) were investigated, and eight experts were exposed to all combinations of these factors (experimental conditions) in each complete experiment trial.

Hines *et al*. (2003) comment that when two factors are being studied in an experiment, both the main effect of each factor and the interaction effects between the factors need to be considered. The main effect of a factor is defined as the change in an expert's behaviour in response to a change in the treatment level of the factor. However, there might also be an interaction effect between the factors, such that the difference in behaviour between treatment levels of one factor is not the same at all other treatment levels of the other factor. As a significant interaction effect can mask the significance of main effects, it is important to ascertain its influence. This is done by determining and interpreting the main effect of a factor in relation to specific treatment levels of the other factor, as opposed to its general leverage across all values of the other factor. Hence, an analytical framework adopted for a repeated measures experiment must be able to serve the following two purposes:

i. Look at the main effects of individual factors (the independent variables) on a behavioural measure (the dependent variable); and where possible

ii. Provide insights on how the factors interact with each other, and what interaction effects these have on the dependent variable.

There are broadly two means to carry out an analysis: parametric and non-parametric. The decision to use either means for the analytical framework will depend on whether the data fulfil the parametric criteria. If the data meet the criteria, then a framework based on parametric tests is suitable, otherwise one that is based on non-parametric tests will be more appropriate. In general, parametric tests are preferred over non-parametric tests as the former are believed to have more statistical power over the latter, though this is not always true (Toothaker and Newman, 1994; Field, 2006).

The analytical framework that was used eventually for testing the hypotheses in this thesis is outlined in Figure C.1, and discussed in more detail below. It is based on parametric tests only, since the data were found to meet the parametric criteria. As well, the various tests used in the framework are also elaborated further in the sections following next.

C.1 AN OVERVIEW OF THE PARAMETRIC ANALYTICAL FRAMEWORK

Before subjecting any data to a parametric analysis, they must first meet four main criteria:

i. The data for each experimental condition must be from a normally distributed population (criterion of normality);

ii. The dependent variable should be measured on at least an interval scale (criterion of minimum measurement level);

iii. The data for each experimental condition must have the same variance as one another (criterion of homogeneity of variance); and

iv. The data are independent of each other (criterion of independence).

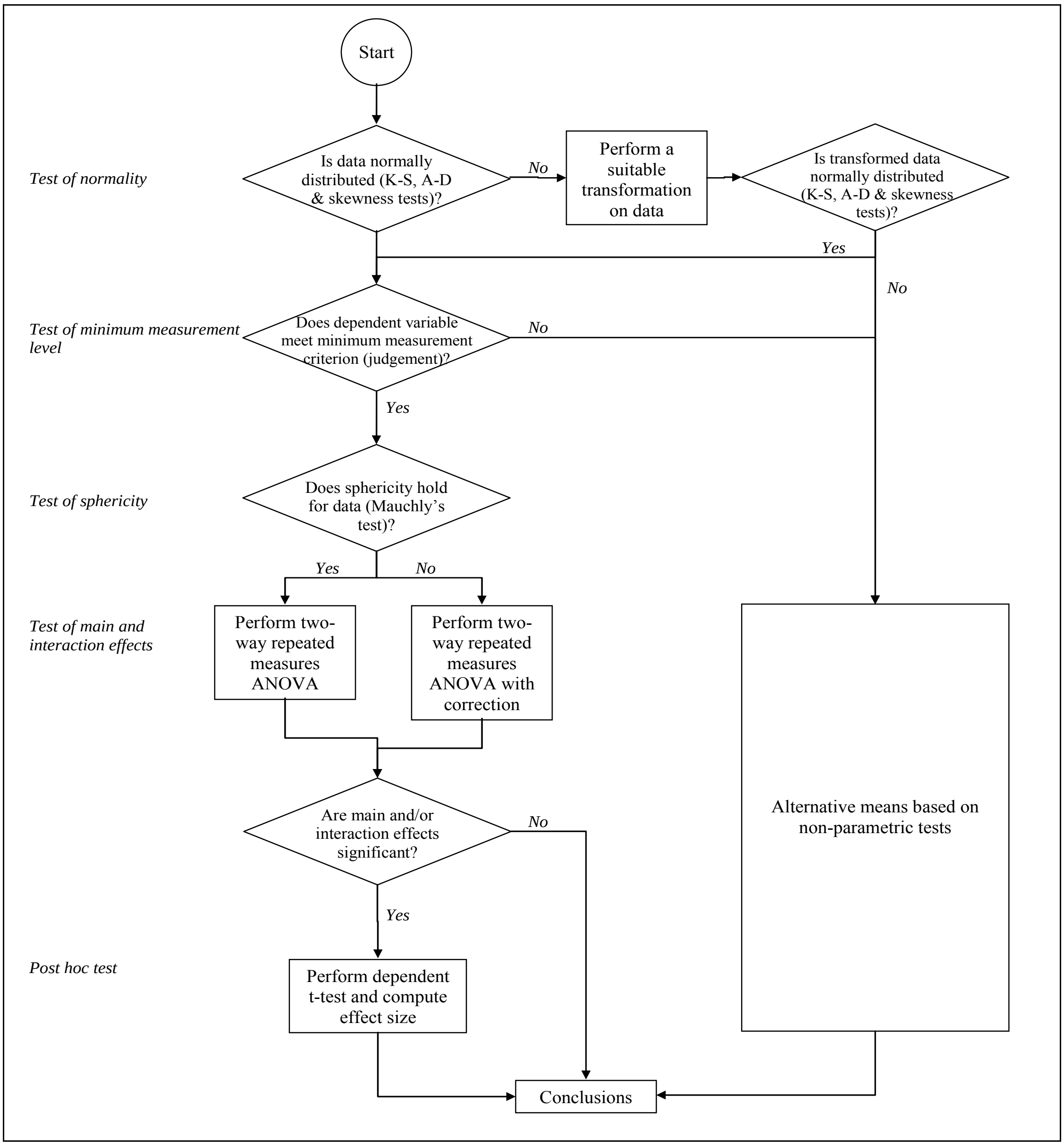

Start
Test of normality
Is data normally distributed (K-S, A-D & skewness tests)?
No
Perform a suitable transformation on data
Is transformed data normally distributed (K-S, A-D & skewness tests)?
Yes
No
Test of minimum measurement level
Does dependent variable meet minimum measurement criterion (judgement)?
No
Yes
Test of sphericity
Does sphericity hold for data (Mauchly's test)?
Yes
No
Test of main and interaction effects
Perform two-way repeated measures ANOVA
Perform two-way repeated measures ANOVA with correction
Alternative means based on non-parametric tests
Are main and/or interaction effects significant?
No
Yes
Post hoc test
Perform dependent t-test and compute effect size
Conclusions

Figure C.1: The analytical framework for testing the hypotheses in this research

However, Field (2006) points out that as the experiment used a repeated measures design with full participation from each expert, the last two criteria (criterion of homogeneity of variance and independence) do not need to be tested. Thus, as long as the data satisfy the first two criteria (criterion of normality and minimum measurement level), a parametric analytical framework based on the two-way repeated measures ANalysis Of VAriance (ANOVA) can be employed.

Instead, Field (2006) and Howell (2007) suggest that the data be assessed for another criterion known as sphericity. It determines if the relationships between different pairs of treatment levels of a factor are similar. In effect, it can be likened to the criterion of homogeneity of variance required in between-group ANOVA. However, the sphericity criterion is not relevant if there are only two treatment levels in the factor. It is because, by definition, the factor needs at least a third treatment level before a comparison can be made between different pairs of treatment levels. Subsequently, the results from this assessment are used to decide whether it is necessary to revise the critical values that are used for assessing the test statistics from the ANOVA that follows.

Under the parametric analytical framework, a two-way repeated measures ANOVA is performed on the data to work out concurrently if any of the factors affects the dependent variable, and if the factors interact at all. Field (2006) explains that the latter is known as a two-way ANOVA because two factors are being analysed in a single test, and it is a repeated measures ANOVA (also known as within-subjects ANOVA) because all factors are measured using the same experts. Also, Field (2006) and Howell (2007) note that ANOVAs are omnibus, as they test for an overall experimental effect due to a factor and do not provide any clues on the nature of the effect. As an illustration, consider a single factor experiment with more than two treatment levels. Whilst ANOVA may detect that the single factor affects the dependent variable, it does not provide any insights on whether a particular treatment level has a greater effect on the dependent variable than the rest, or a combination of treatment levels of the factor has an unequal effect on the dependent variable.

Further to this, Field (2006) suggests that a series of planned or post-hoc comparisons are carried out on the treatment levels, in order to ascertain the specifics of a factor's effect. On the one hand, planned comparisons, which are also known as planned contrasts, are performed when the alternative hypotheses are specified to test some *a priori predictions* about the data. Likewise, as planned comparisons were not used in the eventual analysis, they shall not be discussed further. On the other hand, post-hoc comparisons, which are also known as pairwise comparisons, are performed when the alternative hypotheses are specified to explore the data for *any differences* due to treatment levels in the factors. A series of post-hoc/pairwise comparisons are carried out by performing a dependent t-test on every pair of treatment levels of the factor. As such, they are not applicable if the factor has only two treatment levels. In the event that a treatment level is found to have a statistically significant effect relative to other treatment levels, the effect's materiality and importance are determined by computing its size.

Siegel and Castellan (1988) point out that one may find the test for an overall experimental effect due to a factor to be redundant in light of the post-hoc comparisons that follow. However, they argue the former can be justified by adopting the view that it is only when there are positive results supporting an overall effect's presence, then will the latter be carried out subsequently. The formal tests for normality, sphericity, main and interaction effects are described in more detail in the following sections. In addition, the post-hoc test and the computation of effect size are explained. However, the test for minimum measurement level will not be elaborated upon, as it is effectively done by classifying each variable as nominal, ordinal, interval or ratio.

C.2 TEST OF NORMALITY

To begin, a histogram of the data is plotted to give an initial assessment of its distribution. This is then followed by executing statistical tests to support the initial assessment. There are two statistical tests that can be used to check whether a distribution deviates from a comparable normal distribution; they are the one-sample Kolmogorov-Smirnov (K-S) test, and the Anderson-Darling (A-D) test. On the one hand, the K-S test compares the data to a normally distributed set of values with the same mean and standard deviation as the data. On the other hand, the A-D test is a modification of the K-S test that gives more weight to the tails than does the latter. Also, the A-D test is a more sensitive test than the K-S test.

An appropriate pair of null and alternative hypotheses for each K-S or A-D test executed are given as follows:

H_0 : The data are normally distributed;

H_a : The data are not normally distributed.

Using a 5% level of significance, the result of the K-S or A-D test is insignificant if the p-value is greater than 0.05. This implies that the data are normally distributed. Conversely, the result is significant if the p-value is less than 0.05, which implies that the data are not normally distributed.

However, Motulsky (1999) comments that the K-S test might have little statistical power to determine if the data are from a normal distribution, when the sample size is small (less than 12 in size). It is because a small sample does not contain enough information for inferring the shape of the population's distribution. Also, D'Agostino and Stephens (1986) comment that the A-D test might be unsuitable when the sample size is smaller than eight. Instead, O'Brien and Griffiths (1967) suggest that the skewness test appears to be sufficient for detecting departures from normality when the sample is less than 100 in size. This is despite the fact that the skewness test is strictly a test of symmetry, which is not able to distinguish a normal distribution from other symmetrical distributions. Notwithstanding, the latter reservation is not a major issue, as the ANOVA is quite robust; it is able to accommodate moderate departures from normality (Howell, 2007). Therefore, in view of the above, the skewness test will be used alongside the K-S and A-D tests to ascertain whether it is appropriate to apply ANOVAs on the data.

An appropriate pair of null and alternative hypotheses for each skewness test executed are given as follows:

H_0 : The data are symmetrically distributed;

H_a : The data are not symmetrically distributed.

The test statistic $\left(Z_{skewness}\right)$ for the skewness test and its distribution are provided by Field (2006) as below:

$$Z_{skewness} = \frac{Skewness}{StdError_{skewness}} \sim N(0,1)$$

Using a 5% level of significance, the result of a skewness test for small samples is insignificant if the absolute value of the test statistic is less than 1.96. This implies that the data are symmetrically distributed. Conversely, the result is significant if the absolute value is greater than 1.96, which implies that the data are not symmetrically distributed.

Field (2006) and Hair *et al.* (2006) further remark that if the data's distribution is too skewed to be deemed not normal by the K-S and A-D tests, nor symmetrical by the skewness test, the problem might be corrected by using one of the following three transformations: logarithm, square-root or reciprocal. These transformations have an effect of reducing the impact of outliers that causes the skewness, and yet still manages to retain the relationships within variables. That is, the relative differences between experts for a given variable stay the same. Then, the transformed data are put through another cycle of the K-S, A-D and skewness tests to assess for normality. In the event that there are more than one transformation that work, the optimal transformation will be the one that reduces the data's skewness by an extent that is just adequate for it to be assessed as being normally distributed.

C.3 TEST OF SPHERICITY

When there are at least three treatment levels, the simplest way to check for sphericity is to compute the differences between pairs of data in all combinations of the treatment levels, and compare the variances of these differences between treatment levels. If these variances are approximately equal, then sphericity is present. As an illustration, consider a single factor experiment with three treatment levels labelled as A, B and C.

Hence, three sets of differences comprising 'A – B', 'B – C' and 'A – C' are computed. As such, sphericity for this experiment will hold if:

$\text{Variance}_{\text{A-B}} \approx \text{Variance}_{\text{B-C}} \approx \text{Variance}_{\text{A-C}}$

A test known as Mauchly's test can be used to assess whether the variances of the differences between treatment levels are equal. An appropriate pair of null and alternative hypotheses for each Mauchly's test executed are given as follows:

H_0 : The variances of differences are not different;

H_a : The variances of differences are different.

Using a 5% level of significance, the result of Mauchly's test is insignificant if the p-value is greater than 0.05. This implies that the variances of differences are not different and sphericity holds for the experiment. Conversely, the result is significant if the p-value is less than 0.05, which implies that the variances of differences are different, and sphericity does not hold for the experiment.

C.4 TEST OF MAIN AND INTERACTION EFFECTS

A two-way repeated measures ANOVA is performed on the data to work out concurrently if any of the factors affects the dependent variable, and if the factors interact at all. After it is performed, the test statistics are compared against the relevant critical values, which are dependent on the results from the sphericity test above. If the sphericity criterion for a factor is found to be violated in the previous test, then a revision factor known as the Greenhouse-Geisser correction will be applied onto the corresponding degree of freedom to revise the critical value for the factor.

An appropriate pair of null and alternative hypotheses for each ANOVA executed to assess the main effect of visual representation dimension or model parameters are given as follows:

H_0 : The factor (by itself) does not affect the dependent variable;

H_a : The factor (by itself) affects the dependent variable.

Also, an appropriate pair of null and alternative hypotheses for each ANOVA executed to assess the interaction effect between visual representation dimension and model parameters are given as follows:

H_0 : There is no interaction between visual representation dimension and model parameters associated with it;

H_a : There is interaction between visual representation dimension and model parameters associated with it.

Using a 5% level of significance and depending on the pair of hypotheses used, the result of the test is insignificant if the relevant p-value is greater than 0.05. This implies that there is no evidence of any main/interaction effect being present. Conversely, the result is significant if the p-value is less than 0.05, which implies that there is evidence of main/interaction effect being present.

C.5 POST-HOC TEST

A post-hoc test consists of a series of pairwise comparisons that is designed to compare all different combinations of treatment levels of a factor. In essence, a dependent t-test is performed on every pair of treatment levels of the factor. As there is a particular danger of inflating the overall Type I error rate (also known as familywise error rate) across the tests carried out in a series of pairwise comparisons, a revision known as the Bonferroni adjustment is applied to the error rate. In the SPSS output of the two-way repeated measures ANOVA, a Bonferroni adjustment is already incorporated into the p-values to control the overall Type I error rate to a 5% level of significance.

An appropriate pair of null and alternative hypotheses for each pairwise comparison executed are given as follows:

H_0 : The pair of treatment levels are not different in their effects on the dependent variable;

H_a : The pair of treatment levels are different in their effects on the dependent variable.

Using an overall 5% level of significance, the result of a pairwise comparison is insignificant if the relevant p-value is greater than the value given by '0.05/number of comparisons'. This implies that the pair of treatment levels are not different in their effects on the dependent variable. Conversely, the result is significant if the p-value is less than the aforementioned value, which implies that the pair of treatment levels are different in their effects on the dependent variable. In the latter case, the identity of the treatment level with a larger effect can be determined from the sign of the 'mean difference' in the statistical output. As an illustration, if the mean difference between treatment level 1 and treatment level 2 is positive, then it can be concluded that level 1 has a greater effect on the dependent variable, and *vice versa*.

C.6 EFFECT SIZE

An effect size is a standardised measure of the magnitude of the statistically significant effect observed between two treatment levels. A common effect size measure is Pearson's correlation coefficient (r), which can be computed using relevant values in the SPSS output generated from performing a planned comparison between the two treatment levels. It is expressed as:

$$r_{treatment\,1\,vs\,treatment\,2} = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}}$$

where df_R is the residual degrees of freedom, and $F(1, df_R)$ is the F-ratio for the contrast between treatment level 1 and 2.

As a guideline, Field (2006) suggests that:

i. If $0.10 \leq r < 0.30$, then the effect is small; or

ii. If $0.30 \leq r < 0.50$, then the effect is medium; or

iii. If $r \geq 0.50$, then the effect is large.

D SPSS ‘Frequencies’ Output

The values that were used for computing the test statistics for the skewness tests in Section 10.3.2 and 10.4.2 are provided in the tables below.

D.1 SPSS ‘FREQUENCIES’ OUTPUT FOR TRANSFORMED CASE QUANTITY

Statistics

		Transformed case quantity (2D, Unadjusted)	Transformed case quantity (2D, Adjusted)	Transformed case quantity (2.5D, Unadjusted)	Transformed case quantity (2.5D, Adjusted)	Transformed case quantity (3D, Unadjusted)	Transformed case quantity (3D, Adjusted)
N	Valid	8	8	8	8	8	8
	Missing	0	0	0	0	0	0
Mean		1.745960	1.572306	1.531244	1.415165	1.792893	1.840189
Std. Deviation		.3225074	.2376271	.2955743	.1932988	.3189897	.2212127
Skewness		-.305	-.084	.334	.175	.384	-1.271
Std. Error of Skewness		.752	.752	.752	.752	.752	.752

D.2 SPSS ‘FREQUENCIES’ OUTPUT FOR TRANSFORMED COLLECTION RATE

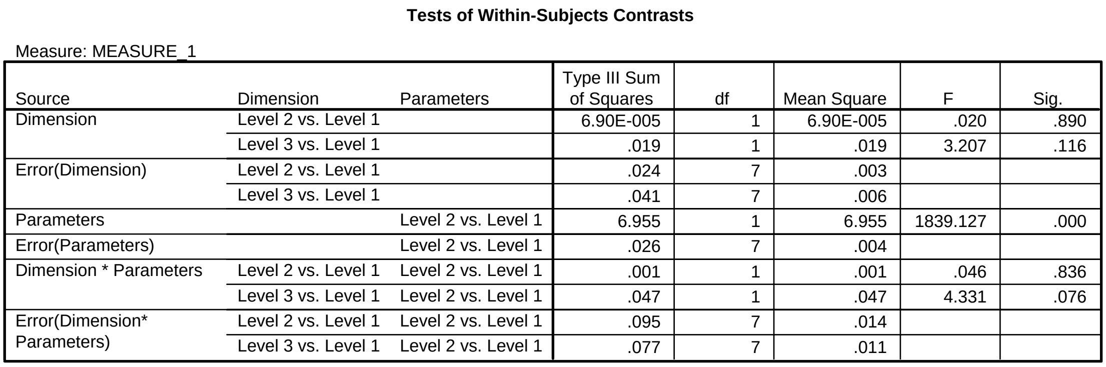
Statistics

		Transformed rate (2D, Unadjusted)	Transformed rate (2D, Adjusted)	Transformed rate (2.5D, Unadjusted)	Transformed rate (2.5D, Adjusted)	Transformed rate (3D, Unadjusted)	Transformed rate (3D, Adjusted)
N	Valid	8	8	8	8	8	8
	Missing	0	0	0	0	0	0
Mean		.168916	.015546	.043479	-.059152	-.076375	-.107280
Std. Deviation		.2127060	.2043404	.2568476	.1717542	.2781715	.1970505
Skewness		.157	-.546	.705	.011	.096	-1.356
Std. Error of Skewness		.752	.752	.752	.752	.752	.752

E SPSS ‘Tests of Within-Subjects Contrasts’ Output

The values that were used for computing the effect sizes in Section 10.2.4, 10.3.2 and 10.4.2 are provided in the tables below.

E.1 SPSS ‘TESTS OF WITHIN-SUBJECTS CONTRASTS’ OUTPUT FOR STANDARD DISTANCE

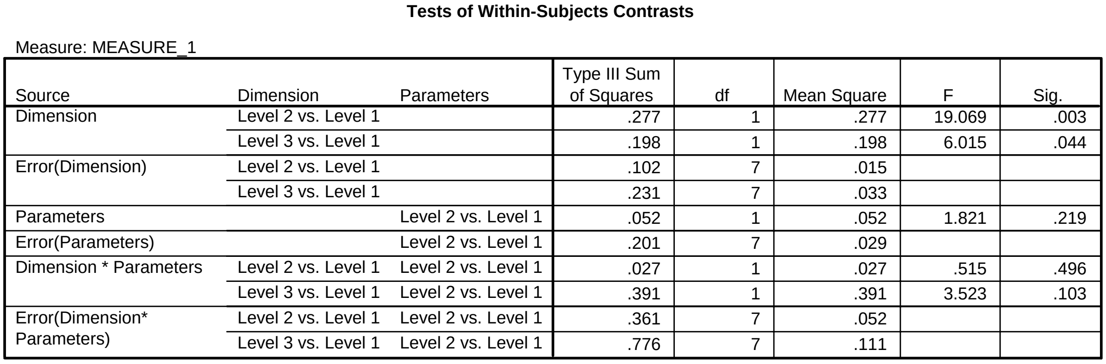


Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Dimension	Parameters	Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Level 2 vs. Level 1		6.90E-005	1	6.90E-005	.020	.890
	Level 3 vs. Level 1		.019	1	.019	3.207	.116
Error(Dimension)	Level 2 vs. Level 1		.024	7	.003		
	Level 3 vs. Level 1		.041	7	.006		
Parameters		Level 2 vs. Level 1	6.955	1	6.955	1839.127	.000
Error(Parameters)		Level 2 vs. Level 1	.026	7	.004		
Dimension * Parameters	Level 2 vs. Level 1	Level 2 vs. Level 1	.001	1	.001	.046	.836
	Level 3 vs. Level 1	Level 2 vs. Level 1	.047	1	.047	4.331	.076
Error(Dimension* Parameters)	Level 2 vs. Level 1	Level 2 vs. Level 1	.095	7	.014		
	Level 3 vs. Level 1	Level 2 vs. Level 1	.077	7	.011		

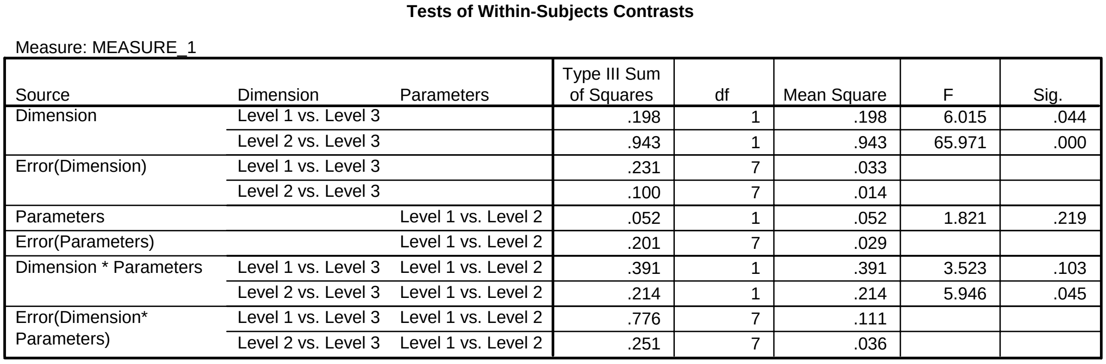
E.2 SPSS ‘TESTS OF WITHIN-SUBJECTS CONTRASTS’ OUTPUT FOR TRANSFORMED CASE QUANTITY



Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Dimension	Parameters	Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Level 2 vs. Level 1		.277	1	.277	19.069	.003
	Level 3 vs. Level 1		.198	1	.198	6.015	.044
Error(Dimension)	Level 2 vs. Level 1		.102	7	.015		
	Level 3 vs. Level 1		.231	7	.033		
Parameters		Level 2 vs. Level 1	.052	1	.052	1.821	.219
Error(Parameters)		Level 2 vs. Level 1	.201	7	.029		
Dimension * Parameters	Level 2 vs. Level 1	Level 2 vs. Level 1	.027	1	.027	.515	.496
	Level 3 vs. Level 1	Level 2 vs. Level 1	.391	1	.391	3.523	.103
Error(Dimension* Parameters)	Level 2 vs. Level 1	Level 2 vs. Level 1	.361	7	.052		
	Level 3 vs. Level 1	Level 2 vs. Level 1	.776	7	.111		

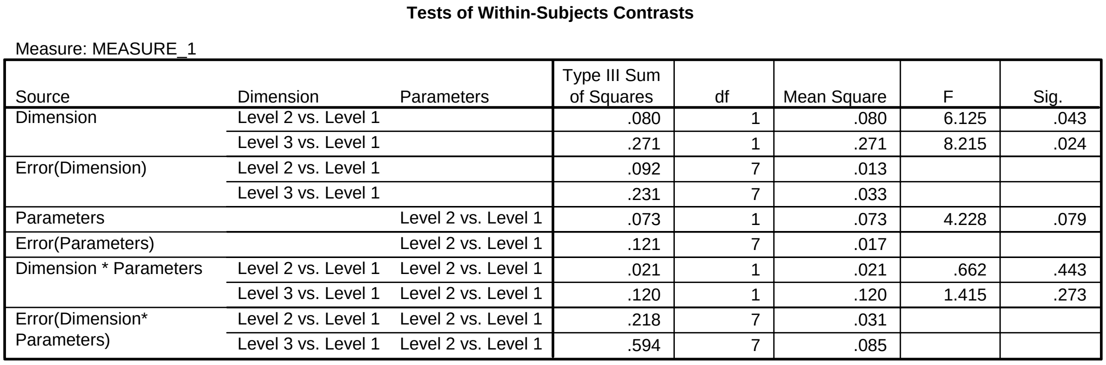


Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Dimension	Parameters	Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Level 1 vs. Level 3		.198	1	.198	6.015	.044
	Level 2 vs. Level 3		.943	1	.943	65.971	.000
Error(Dimension)	Level 1 vs. Level 3		.231	7	.033		
	Level 2 vs. Level 3		.100	7	.014		
Parameters		Level 1 vs. Level 2	.052	1	.052	1.821	.219
Error(Parameters)		Level 1 vs. Level 2	.201	7	.029		
Dimension * Parameters	Level 1 vs. Level 3	Level 1 vs. Level 2	.391	1	.391	3.523	.103
	Level 2 vs. Level 3	Level 1 vs. Level 2	.214	1	.214	5.946	.045
Error(Dimension* Parameters)	Level 1 vs. Level 3	Level 1 vs. Level 2	.776	7	.111		
	Level 2 vs. Level 3	Level 1 vs. Level 2	.251	7	.036		

E.3 SPSS 'TESTS OF WITHIN-SUBJECTS CONTRASTS' OUTPUT FOR TRANSFORMED COLLECTION RATE



Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Dimension	Parameters	Type III Sum of Squares	df	Mean Square	F	Sig.
Dimension	Level 2 vs. Level 1		.080	1	.080	6.125	.043
	Level 3 vs. Level 1		.271	1	.271	8.215	.024
Error(Dimension)	Level 2 vs. Level 1		.092	7	.013		
	Level 3 vs. Level 1		.231	7	.033		
Parameters		Level 2 vs. Level 1	.073	1	.073	4.228	.079
Error(Parameters)		Level 2 vs. Level 1	.121	7	.017		
Dimension * Parameters	Level 2 vs. Level 1	Level 2 vs. Level 1	.021	1	.021	.662	.443
	Level 3 vs. Level 1	Level 2 vs. Level 1	.120	1	.120	1.415	.273
Error(Dimension* Parameters)	Level 2 vs. Level 1	Level 2 vs. Level 1	.218	7	.031		
	Level 3 vs. Level 1	Level 2 vs. Level 1	.594	7	.085		