Every Picture tells a Story: An Investigation of Data Models as Tools of Communication

A thesis presented in partial
Fulfillment of the requirements for the degree of
Master of Information Sciences
At Massey University, Palmerston North, New Zealand

SOMRUTAI MALAIPONG

2000

Abstract

Data models are important in information system (IS) development, particularly as tools for expressing and communicating information business requirements. The ability to understand the information content of data models is a fundamental skill required by anyone involved with them. The aim of data modelling is usually the creation of a database design but without everyone having a clear understanding of what the data model 'says', the quality of the design may suffer. When we build a model, we obviously want it to be understood but that understanding is dependent on the ability of the model to communicate its meaning. The better the model is as a vehicle of communication the clearer the understanding will be.

This research report explores this important aspect of data models and investigates the research that has been undertaken in this area including the NaLER (Natural Language for E-R) technique for reading data models. It describes an experiment conducted to explore this aspect. Subjects were tested on their ability to accurately and comprehensively interpret or 'read' a data model both before and after learning the NaLER technique. Measurement was done by using a questionnaire which consisted of two types of questions. The results show that when the subjects used NaLER, they improved their scores on the difficult questions but not the simple and medium questions. In addition, the results show that after learning NaLER, subjects' confidence in their ability to understand a model was increased, even though they actually scored less well overall.

Acknowledgements

I would like to express my deep gratitude to my supervisor, Ms. Clare Atkins, for her invaluable guidance, advice and encourage provided throughout the course of my study.

I would like to thank Mr. Barry Jackson for his support and advice.

I would like to thank to my friends for their support and commiseration.

I would also like to thank to staff of the Information System Department for their supports and advice.

A special thanks to my husband and my family for their support and encouragement they have given me throughout my study.

Finally, I would like to thank to the New Zealand Government for their financial support during my study and to the Thai Government for granting me study leave.

Table of Contents

List of Tables

List of Figures

Chapter One - Introduction

A data model is an abstraction of the real things people deal with in their business and is a very powerful tool for expressing and communicating business requirements. As both business requirements and database technology become more sophisticated, the need to understand data models is increasing. Modern technologies and computer-aided software engineering (CASE) tools now include data modelling as a key phase in the development life cycle, and a data model has taken its place, alongside the functional specification, as one of the major deliverables of system analysis (Whitten *et al.*, 1998). Data modelling, or at least, understanding data models, is now a required skill for virtually anyone involved in information systems analysis and design.

A data model is a relatively small part of the systems specification but has a high impact on the quality and useful life of the system (Simsion, 1994). A well-designed data model can make programming simple and cheaper while a small change to a data model may have a major impact on the system as a whole (Simsion, 1994). Since clarity and accuracy of the analytical documentation at the early stages of the development life cycle is critical to the success of the final product, undetected ambiguity, or inaccurate analysis, can result in faulty systems. Therefore communication is one of the most important factors to develop the most appropriate system and a data model is often the tool, which is used to communicate between different groups of people involved in the development life cycle.

According to Atkins and Patrick (1998), the ability to understand the semantics or information content of data models is a fundamental skill required by any person involved in system development. Although the ultimate aim of data modelling is usually the production of a "good" database design, Campbell (1992, p.12) points out that "data modelling diagrams are being used for a number of purposes including; as a user communication tool during Data Analysis (documenting the data needs of the business), as a Database Design tool (at this stage we are designing the database which will support the requirements identified above), as part of Systems Documentation (the ultimate purpose of which is to communicate requirements and design) and as part of a schema definition (along with data dictionary entries) which a CASE tool will use to generate data definition language statements for a particular database management system." This is supported by Hitchman (1995) who says that the analysis and representation of the persistent data is an activity fundamental to the creation of any information system development. Some form of data representation or data model is generally considered to be a desirable, and in some cases, an essential element of an information system.

The first step of database design is the development of the conceptual data model of the application. A conceptual data model is an abstraction of the real world (organisation) data pertinent to an enterprise. The process of deriving and analysing the data inherent in a business situation and of mapping objects of this understanding of reality in a conceptual model representation constitutes the discovery phase (Batra *et al.*, 1990). The discovery phase consists of two parts. The first part involves the elicitation of the information requirements, which is traditionally a result of an interview between the user and the analyst. The second part, in the data modelling context, involves the representation of the information requirements into the form of a conceptual data model (Batra *et al.*, 1990).

My research concentrates on the second part of this phase, that is, the translation of the understanding of the business situation into a representational form. This schema is usually represented in a diagrammatic form, and serves as a communication tool between developers and users. Once approved by users, it is converted into a specific database schema, depending on the data model and the DBMS used for implementation. However, the diagram can be misunderstood or misread by the users and in order to explore this issue, two experiments were conducted. The purpose of this study is to compare the understandability of models by students before and after learning a technique NaLER (Natural Language for Entity Relationship Models) designed to assist their 'reading' of E-R (Entity Relationship) models.

Data models are important in information system development, particularly as tools for expressing and communicating business requirements. Since the importance of data models within the creation of information systems has increased, the range of people involved with their use has grown. The group includes data modellers, database designers, administrators, system analysts and end users.

Basically, a data model is the way of a set of perceiving, organising and describing data. It consists of a set of concepts such as items, groups and relations or entities and attributes, which facilitate the specification of data in a precise format (Klein and Hirschheim, 1987). The ability to understand the information content of data models is a fundamental skill required by everyone involved with them. Thus, a data model is an important element in information system development. However, there has been little research on the use of data models as tools of communication.

Many researchers state that data models are utilised to communicate with users during the database design process (e.g. Simsion, 1994; Jih *et al.*, 1989; Batini *et al.*, 1992). Many authors also write that system success is dependent on effective communication between the system designers and the users. (e.g. Avison and Fitzgerald, 1995; Burch, 1992; Brodie *et al.*, 1984; Campbell, 1992). Campbell (1992) states that "to satisfy the user communication aspect, there must be an appropriate way of showing and demonstrating the developing system to the users". He also highlights the fact that data models are useful in adopting a user's view of the world but it is important that the data analyst should explain the model in a language that the user is familiar with. However, Connolly *et al.* (1996) suggest that the adoption of another modelling formalism does not appear to be warranted solely to enhance analyst/user communication. The process of converting the model from one formalism to another may introduce more errors that the suspected analyst/user mis-communication. Also, it may be worthy of note that many users may have come into contact with an E-R Model in past projects and trying to teach them another formalism may confuse them even more.

According to Burch (1992), "conceptual data models play a substantial and essential communication role; one of the major reasons for creating models is to facilitate communication with the user" (p.35). Burch also adds that "the model must serve as a communication medium between users and systems professionals." This is supported by Batra and Davis (1992) who say that "the conceptual model is usually easy to understand and can from the basis for communication with users" (p.83). Burch, 1992, is in agreement, writing that "written, oral and graphic communication are all necessary for communicating documented deliverables effectively to all participants in the development of the new system"(p.35). Therefore, to some degree, evaluating a data model as a tool of communication tool is also dependent on how well the analysts present the models to the users.

When we build a model we obviously want it to be understood, but that understanding is dependent not only on how well the analyst presents the model but also on the ability of the model to communicate its meaning. This research focuses on this last issue and is thus based on the following questions:

1. How accurately are ER models understood using the standard techniques?
2. How comprehensively are ER models understood using the standard techniques?
3. How accurately are ER models understood using the NaLER technique?
4. How comprehensively are ER models understood using the NaLER technique?
5. Is there a higher level of user confidence in their understanding when using the NaLER technique?

Based on the previous questions, three hypotheses have been defined for this study:

H 1: ER models read using the NaLER technique will yield a more accurate understanding than ER models read using the standard techniques.

H 2: ER models read using the NaLER technique will yield a more comprehensive understanding than ER models read using the standard techniques.

H 3: Users using the NaLER technique will have a higher level of confidence in their understanding.

Two experiments were conducted to address the research questions. The subjects were recruited from a final year undergraduate course in Data Modelling, thus they had some data modelling and database background. Therefore, the subjects could in some ways, be considered to be 'sophisticated users'. In the first experiment, the subjects were tested on the standard approach to reading an ER model. The subjects were randomly assigned to two groups. They were given a copy of a data model and were asked to write a short description of what it represented. They were also given one of two sets of questions to answer. These questions were asked them to indicate whether a statement about the data model was true or false and were designed to assess three levels of understanding (simple, medium and deep). Finally, the subjects were asked how confident they were in their understanding of the data model.

Over the next few weeks, the subjects were taught another technique, NaLER. The second experiment was conducted after the subjects had learned this new technique. In the second experiment, the subjects were given the same data model as the first experiment but they were given the alternative set of questions to answer. They were also asked to write the NaLER sentences to state how they 'read' the model by using this method. Finally, they were again asked to indicate their confidence in their understanding of the data model.

Since the experiments required the participation of humans, approval was sought from the university's Ethical Committee.

The major limitations of this research stem from the difficulties of creating a 'real world' environment in an experimental situation. As Galliers, 1992 says the empirical studies of the experimental research type have inherent limitations, for example the isolation of such situations from most of the variables that are found in the real world. Due to the fact that students are not actually sophisticated users they have their limited business knowledge. Also, these students represent a self-selected sample because they have chosen both an Information System major and then a Data Modelling course. Another limitation was time, since the experiments were done by using the volunteer students as subjects, and both experiments took place in tutorial time to avoid interrupting students' normal learning process. In addition, the researcher herself had a limited time to complete her Masters in Information Science, so there is no control group in this experiment. Then, the method (NaLER) itself, is not a formal method, and more precise guidelines might have been helpful.

Although not a limitation, a real problem for this research is the ambiguous terms related to a data model itself. Therefore, the definitions and a glossary are provided to help in establishing a consistent use of terminology throughout this research report.

This research report consists of six chapters. Chapter Two presents related research which is discussed in order to provide some background to the understanding of data models. The aim of this chapter is to examine areas relevant to the research objectives and to highlight areas not extensively covered in the literature. A definition of terms and a basic terminology also given. Chapter Three discusses two techniques for reading E-R models; the standard technique and the NaLER technique. Chapter Four describes the research methodology and experiment background. The reason why an experimental approach was considered the most suitable for this research is also discussed. Results from the experiments will be discussed in Chapter Five. Finally, Chapter Six provides an overall conclusion and recommendations for further research are discussed.

Chapter Two - Context

This chapter presents the research related to the general area of data modelling. It is separated into three sections. The terminology of some basic terms are described in the first section. The second section discusses a number of topics which are related to data models. The final section highlights data modelling research and includes general studies, comparative studies and data modelling as communication tools.

2.1 Terminology

There are a number of terms relevant to data models that are ambiguous or are unclear. Therefore, this section will help in establishing a consistent use of some of the terminology used throughout this research report. It focuses on the terms which are particularly contentious. A final definition of these terms and others are included in the glossary.

2.1.1 Data Model

The literature presents a number of quite varied definitions regarding what constitutes a 'data model'. Simsion (1994) recognizes the term 'data model' as one used to represent two distinct concepts when applied by practitioners or academics. "Practitioners use the term to refer to the representation of data required to support a particular function or set of functions, whereas academics use it to refer to a particular method, or formalism, for representing data. Academics use data model to describe a particular way of representing data: for example, in tables, hierarchically, or a network. Hence they talk of the relational model, the entity relationship model or the network model"(p.22). Obviously, there are two conflicting meaning of data model and it is not always clear which one an author is using.

Connolly *et al.* (1996) describes a 'data model' as "an integrated collection of concepts for describing data, relationship between data and constraints on the data in an organisation"(p.59). They continue, "a data model represents the organisation itself, it should provide the basic concepts and notations that will allow database designers and end users to unambiguously and accurately communicate their understanding of the organisational data"(*ibid.* p. 59). The purpose of the data model is thus to represent data and make the data understandable. Connolly *et al.* (1996) identify three related data models:

> "1. An external data model, to represent each user's view of the organisation, sometimes called the Universe of Discourse (UoD),
>
> 2. A conceptual data model, to represent the logical (or community) view that is DBMS independent,
>
> 3. An internal data model, to represent the conceptual schema in such a way that it can be understood by the DBMS" (p.60).

This is the view of the data model as a 'blueprint' and is generally how it is used by practitioners. However, the use of data models as 'architecture' is commonly focused in many text books (e.g. Avison, 1992; Kroenke, 1992; McFadden and Hoffer, 1994; Ricardo, 1990; Sanders, 1995).

The use of the term data model is therefore ambiguous and Benyon-Davies (1996) highlights the differences between 'architecture' and 'blueprint'. He discusses, the data model as architecture, as referring to set of general principles for handling data, e.g. the relational data model, the hierarchical data model or the object-oriented data model. The data model as blueprint, is used to refer to an integrated, but implementation-independent set of data requirements for some application. A data model in this sense is an important component part of any information systems specification.

In this research report, the term 'data model' will be used to mean a 'blueprint', while the 'architecture' will generally be referred to as a modelling formalism.

2.1.2 Data Modelling

Data modelling is the activity of creating a data model and occurs during the system development process. The process of describing, analysing data and designing the required data structures can be referred to as data modelling, while the output of the process are known as data models. Data modelling is a structured set of techniques for defining and recording business information requirements. It is a depiction of the end user's view of the data needs of the organisation recorded in a consistent and rigorous fashion. Therefore, it is an activity considered by many researchers as an essential skill for a successful information systems development (Hitchman, 1995; Batra and Davies, 1992; Simsion, 1994).

Davydov (1994) describes data modelling as an activity that involves the creation of the abstractions that represents a data oriented image of a given application. The data modelling process distinguishes between three major types of representations:

(1) The conceptual representation which identifies entities and their interrelationships,

(2) the logical representation which results from the translation of the conceptual representation into one that is compatible with the chosen DBMS, and

(3) the physical representation which assignment of files to blocks and access paths, which lead to efficient performance of the system.

It is generally believed that this 'three-schema' approach provides considerable benefits in the design process (McFadden and Hoffer, 1994). However, there is some confusion between data modelling as an analysis or as a design activity. Simsion (1994) highlights the practitioner perceptions that data modelling is very much a design activity. On the other hand, from the academic point of view, data modelling can be seen as either an analysis or a design activity or both. For example Shoval and Frumermann (1994) state that "data modelling is an analysis and design activity" (p.28). Benyon (1997) always use the term modeller interchangeably between an analyst and a designer. Current textbooks on information systems development also support this view (e.g. Burch, 1992; Gibson and Hughes, 1994; Hawryszkiewycz, 1997).

The focus of this research report is on the first model in the triad: the conceptual data model.

2.1.3 Conceptual Data Model

The literature presents a number of quite varied definitions of a conceptual data model. McFadden and Hoffer (1994) define a conceptual data model as " a detailed model that captures the overall structure of organisational data, while being independent of any database management system or other implementation consideration. A conceptual data model includes the relevant entities, relationships, and attributes, as well as the business rules and constraints that define how the data are used" (p.87).

Connolly *et al.* (1996) note that the conceptual schema is the 'heart' of the database. It supports all the external views and is, in turn, supported by the internal schema. They define conceptual modelling or conceptual database design as "the process of constructing a model of the information used in an enterprise that is independent of the implementation details, such as the target DBMS, application programs, programming languages or any other physical considerations" (p. 63-64).

This principle also related to the ANSI schema (ANSI/X3/SPARC, 1975) which reports the three levels of abstraction. The external schema, which represents the individual user's view of data, is not mapped directly to the internal schema, the physical database structures in which the data was stored. Instead it is mapped to the conceptual schema, which provides a global or community view of all the data in the information system. The conceptual schema thus acts as a buffer to hide modifications in the other two levels from each other and protect the external schema from any changes to the physical data structures, providing physical data independence.

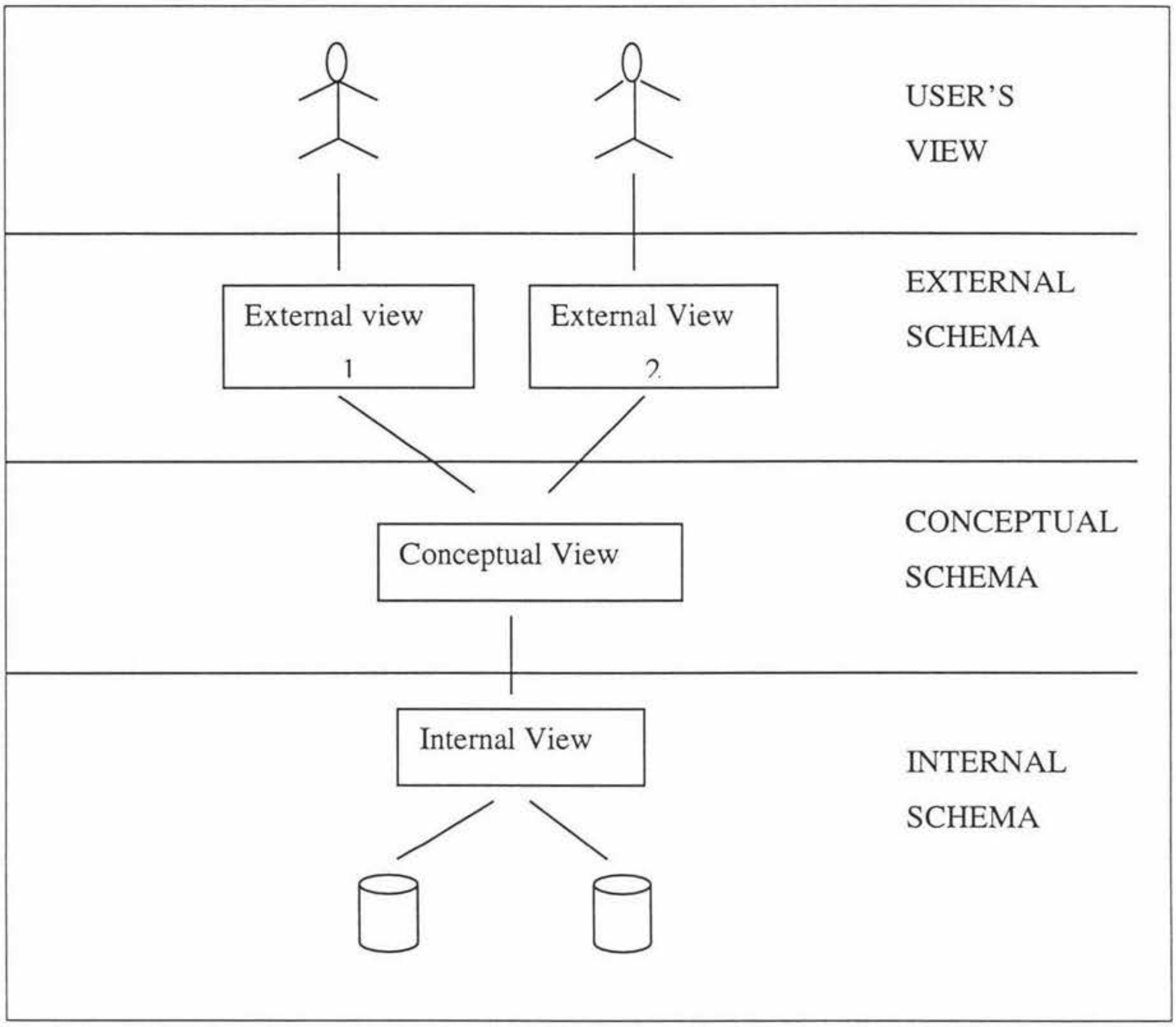


Figure 1 ANSI/X3/SPARC Architecture

The purpose of the three level architecture is to separate the way the database is physically represented from the way the users think about it. There are several reasons to support this separation:

(1) Different users need different views of the same data,

(2) The way a particular user needs to see the data may change over time

(3) Users should not have to deal with the complexities of the database storage structures,

(4) The DBA should be able to change the overall logical structure of the database without affecting all users'

(5) The DBA should be able to change data and file structure without affecting the overall logical structure or users' view,

(6) Database structure should be unaffected by changes to the physical aspects of storage, such as changes to storage devices (Ricardo, 1990).

According to Batini *et al.* (1992), conceptual design is the most critical phase of database design. They define a conceptual model as a tool for representing reality at a high level of abstraction. Using conceptual models, one can build a description of reality that is easy to understand and interpret. Figure 2 presents Batini *et al.*'s placement of conceptual data modelling within the data-driven approach to information system design.

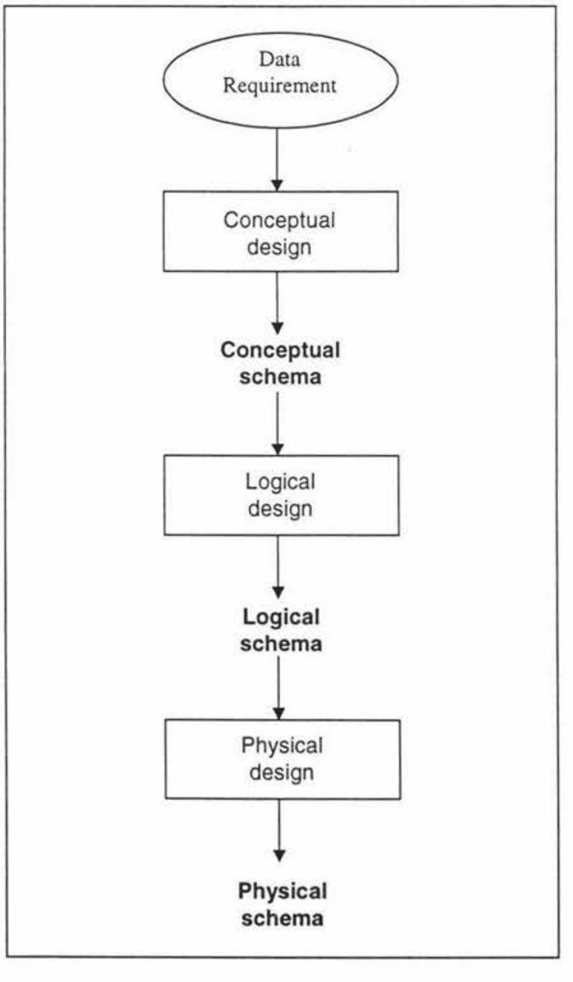


Figure 2 Data-driven approach to information systems design (adapted from Batini *et al.*)

Batra and Marakas (1995) note that there is a seemingly general consensus among academics regarding the value of conceptual data modelling (CDM), they state that "when CDM is properly and rigorously performed, the delivered system is expected to be functionally richer, less error-prone, more fully attuned to meet user needs, more able to adjust to changing user requirements and less expensive" (p.185).

As Simsion (1993) observes, data models are the plans that communicate to the system designers what the system requires. In order for conceptual data models to fulfil their role as a design tool, they must be able to map to a logical design as they are the foundation upon which the logical model is built. Interestingly, Whitten *et al.* (1998) claim that while a systems analysis data model effectively identifies data requirements, it doesn't usually represent a good database design. They also note that good design means creating a model that represents the system or database requirements. Like Campbell (1992), Whitten *et al.* (1998) identify that the same representation of the conceptual data requirements, isn't always suitable for the two different audiences who receive the conceptual data model.

The ISO report on the conceptual schema (Griethuysen, 1983) identifies four major objectives of a conceptual data modelling:

(1) Suitable modelling facilities for static and dynamic assertions about the enterprise,

(2) Ease of use and understandability,

(3) Easy adaptation to changes in the abstraction of the organisation, and

(4) Provisions for a common language for communication between users and the information processor

Further, a conceptual data model should facilitate communication and ease user validation of requirements (Juhn and Naumann, 1985).

It would seen then that there is some consensus in the literature that the conceptual data model should be:

(1) Understood by the subject area experts whose enterprise reality it is intended to represent (Kim and March, 1995; Batini *et al.*, 1992),

(2) A precise, formal and complete specification of the information requirements (Tsichritzis and Lochovsky, 1982; Juhn and Naumann, 1985; McFadden and Hoffer, 1994;),

(3) Free of architectural or implementation bias (Olle, 1993; Ram, 1995; Connolly *et al.*, 1996), and

(4) Transferable into a logical design for specific implementation (Tsichritzis and Lochovsky, 1982; Elmasri *et al.*, 1985; Teorey *et al.*, 1986; Simsion, 1994).

2.1.4 The Entity-Relationship Model

When considering data modelling, it is useful to first understand what each formalism contains. It is therefore important to identify what is meant by the Entity-Relationship (E-R) Model, and to distinguish it from the Chen Model (originally called the Entity-Relationship Model), the Extended Entity-Relationship (EER) Model and E-R/Relational Hybrid (E-R/R).

The original Entity-Relationship Model has been in use since its publication in 1976 (Chen, 1976). The Chen E-R model is a graphical model that seeks to identify data relevant to an information system by categorizing such data into 'entities', 'relationships' and 'attributes' and representing these categories in a graphical form using standard notation. It is termed by some authors the E-R Model (e.g. Connolly *et al.*, 1996; Hawryszkiewycz, 1998) or the Chen model (e.g. Simsion, 1994). There appears to be an area of inconsistency in the literature, what some authors label the E-R Model others call the Extended E-R Model, and what some call the Chen Model other call the E-R Model. This confusing use of terminology shows yet another split between academics and practitioners.

Hawryszkiewycz (1998) defines the Entity-Relationship (E-R) model as "a model that represents system data by entity and relationship sets"(p.182). This definition is supported by many authors (Connolly *et al.*, 1996; Simsion, 1994; Kendall and Kendall, 1999). However, the Chen convention for relationships (the diamond symbol) is rarely used by practitioners (Hitchman, 1995). Simsion (1994) gives some reasons for this, "first, it simply puts too many objects on the page...diamonds make this much harder, and a practical Chen model can be quite overwhelming...Second, most CASE tools don't support the diamond contention...Third, many of the people who contribute to and verify the model will also need to see the final database design. End users may access it through query languages, and analysts will need to specify processes against it. If the final design uses the same concepts as the verified model, these people don't have the problem of coming to grip with two different views of their data"(p. 74).

Initially, the E-R model provided a representation of data, which did not reflect the constraint of the physical database structures. A number of early studies state that the E-R model provide a clear and intuitive way of communicating with non-technical users (Konsynski, 1979; Brodie *et al.*, 1984; Teorey *et al.*, 1986; Batra *et al.*, 1990). However, with the advent of commercial relational database management systems (RDBMS), the E-R model utilised some constructs that mapped naturally to relational objects, i.e. entities to relational tables and attributes to columns. Alongside these developments, in many situations, the users view the database as a collection of tables. As a result, both technical and non-technical users came to utilise some form of the E-R notation as a useful means of the recording and understanding the physical database as well (Atkins, 2000).

As Atkins (1996) recognizes "When the E-R model was originally proposed, it was intended to express data requirements in a way that could be transformed to hierarchical, network or relational structures. However as relational structures have increasingly become the only target, so a number of relational constructs have found their way into the E-R conventions. The result is that many E-R techniques can perhaps be better described as Relational Modelling using E-R graphical notation" (p.36).

There is evidence in the literature to confirm this view of E-R/ Relational hybrid models. Marche (1993) refers to data models, especially 'normalised entity relationship' ones while Batra and Zanakis (1994) specifically state that an E-R diagram developed with the relational representation as the target must obey the normalisation concepts. However, if an E-R has to conform to normalisation rules it cannot claim to be a user's view of the world without any consideration of how data is stored on computerised files (Atkins, 1996) and must be less useful in providing a basis for mapping to non-relational implementation paradigm (Jarvenpaa and Machesky, 1989).

By now, it would seem that the use of the E-R relational hybrid model is widespread in the practitioner areas. Many practitioner models conform to the rules of the Relational Model but are represented by an E-R diagram. This E-R relational hybrid model is what is often meant by the term E-R model. This research also focuses on the data model in term of the E-R relational hybrid model, which is now widely used by practitioners in the conceptual and logical design stages.

2.1.5 The O-O model

An object-oriented model is becoming more prevalent in Information System Development. Whitten *et al.*, (1998) define object modelling as " a technique for identifying objects within the systems environment and the relationships between those objects"(p.286). In this sense, "an object is something that is or is capable of being seen, touched, or otherwise sensed, and about which users store data and associate behaviour" (p.287). This view is very similar to the definition of an E-R model with the term 'object' instead of 'entity' being used to represent something in the real world.

Ricardo (1990) describes the concept of an object-oriented model as "an object-oriented model begins with the notion of physical entities that exist in the real world, and defines an object as a collection of properties that describe an entity. Objects have names, which are the same as the names of the entities represented. The collection of properties chosen for an object must be sufficient to describe the entity, that is it must include those properties that users need to know. In the real world there are entity classes, which are sets of entities, and entity instances, which are occurrences of entities. Similarly, there are object classes and object instances, which represent particular objects" (p.417).

However this research focuses on data models rather than object models which include behaviour. Data models enable the users to more extract the information and permit them through the abstraction to model and view the data in a manner which is consistent with how people view the world.

2.2 Related Research Areas

In this section, information from the literature, which relate to this research areas is discussed under three broad headings.

2.2.1 Infological vs Datalogical

The ISO/CT97 identified two principal purposes of the conceptual schema, firstly to describe the Universe of Discourse (UoD), i.e. to provide an 'abstraction of reality' (Tsichritzis and Lochovsky, 1982) and secondly to control the descriptions in the database; i.e. to prescribe the major data structures that will be implemented (Griethuyzen, 1983). The first purpose implies that the conceptual schema, formulated independently of DBMS considerations, should be understandable to UoD experts. The second purpose requires the representation to be a reflection of the potential database structures. While recognising these two distinct purposes it does not recognise any tension between them nor any specific difficulties in accommodating both within one representation. However, as Shoval (1985) points out, there is a clear "distinction between (the) two levels in that schema, each of which is dedicated to one of the purposes" (p.417).

Tsichritzis and Lochovsky (1982) recognise that the first purpose, which they term 'infological', is primarily used as an aid to understanding and thus validation. The infological view is concerned with the content or the semantics of a system's information need (Hirschheim *et al.*, 1995). The second, 'datalogical' purpose, on the other hand requires a very detailed and technically oriented representation. Much research has sought to provide modelling facilities that reconcile these very different purposes. However, as Tsichritzis and Lochovsky (1982) observe, provided that there is a means of mapping from one form of the representation to the other, then there is no reason why both purposes should be fulfilled by the same formal representation. Later researchers such as Elmasri *et al.* (1985) agree. Nevertheless, the E-R Model continues to be generally used to satisfy both needs, although not with equal success (Elmasri *et al.*, 1985). To extend Simsion's (1994) architect analogy, data modellers are attempting to describe both the detailed design specifications and the artist's impression with the same representation (Atkins, 2000).

It would seem that, initially at least, the E-R Model fulfilled the datalogical role adequately and in the infological role partly filled a difficult gap. However, even though conceptual models are still usually built only as a precursor to database design, increasing levels of abstraction and an increased breadth of scope have made it increasingly important that non-IS specialists are involved in their construction. As Atkins (2000) points out, the users and sponsors of the system must understand and be able to verify the data structures and business rules (Veryard, 1994) while auditors may use the conceptual schema as an indication of database integrity (Amer, 1993). Campbell (1992) also points out that the data modelling diagrams are being used for a number of purposes which include "as a user communication tool (...documenting the data needs of the business) and ...as part of Systems Documentation (the ultimate purpose of which is to communicate requirements and design)" (p.12). It is clear that because of this role as a communication tool, the infological representation is becoming increasingly important (e.g. Shanks, 1997; Shoval and Frumermann, 1994; Siau *et al.*, 1995). The infological role is often associated with research focussed on either improving the E-R Model or finding a better alternative (Blaha *et al.*, 1988; Campbell, 1992; Coad and Yourdon, 1991; Elmasri *et al.*, 1985; Hammer and McLeod, 1981; Schenck and Wilson, 1994; Smith and Smith, 1977; Teorey *et al.*, 1986).

In practice, however, this infological role is not always explicitly recognised. Atkins (2000) comments that, this allows Campbell (1992,p.13) to accuse IS practitioners of "technological arrogance" in expecting users to become familiar with, and think of their data in, the constrained structures that the use of any modelling formalism, based on datalogical requirements, must use (Kent, 1978). Campbell seems justified in his comment that "there is a conflict between the desire to address design issues and the need to create a form of data model with which the business user is comfortable" (Campbell, 1992 p.13). As Olle (1993) observes, the conceptual schema must be represented in a form that is "assimilatable by subject area experts who are not familiar with informatics oriented representation forms"(p.53).

However, the datalogical function of the model almost always appears to be regarded as paramount and most would agree that the focus of the conceptual model is the database design process. After all very few conceptual data models are built unless there is an ultimate goal of creating some slice of virtual reality, i.e. an electronic database (Tsichritzis and Lochovsky, 1982).

We can summarise this discussion by noting that although a data model is often defined in terms of its role as a "communication medium between professional analysts/designers and users" (Shoval and Frumermann, 1994 p.28), it seems that in practice it is almost always built to fulfil its primary, datalogical role (de Carteret and Vidgen, 1995). It is generally agreed that a useful infological model needs to reflect very closely the way users view data. Again Atkins (2000) comments that, users may have their own idiosyncratic ways of viewing their data (Raymond *et al.*, 1989) and that these may differ from person to person and none of the views may be useful in terms of building a database. These considerations have led some researchers to investigate using some form of controlled natural language as the basis for the infological representation. Natural language also provides the only way in which the "connection between database and the reality about which statements are to be represented, can be established" (Biller and Neuhold, 1978 p.11).

2.2.2 Understanding graphical designs

Batini *et al.* (1992) claim that the basic features of conceptual design and of conceptual data models are relatively simple, and their understanding does not require much technical knowledge about database system as a prerequisite. Thus, they mention that users can easily learn enough about conceptual design to guide designers in their decisions and even to design simple database by themselves. However, the lack of clarity and detail concerning some semantic constructs in some data modelling texts would support the view that information concerning some semantic constructs is not readily available. (Barker, 1990; Veryard, 1992). In addition, there is no research which directly addresses this issue.

Atkins and Patrick (1998) observe that apart from the need of understanding a conceptual data model to assist in tasks such as decision making and strategic planning, there are a number of specific activities for which a clear and comprehensive understanding of its information content is essential. They also mention that the ability to understand the semantics or information content of E-R models, i.e. to 'read' them, is a fundamental skill required by any person involved with them. They suggest that the ability to read a data model can be seen to be similar to reading a map. A map may be consulted by a variety of people for any one of a number of reasons. Each individual user needs to have either a working knowledge of the map symbols or access to a suitable annotated legend.

It seems reasonable to suggest that by using a map reading technique, people gain a better and easier understanding when they read a model. MacEachren (1995,p.v) points out that "Maps are powerful tools, and have been for centuries, because they allow us to see a world that is too large and too complex to be seen directly... what we see when looking at a map is not the world, but an abstract representation that we find convenient to use in place of the world."

Reading a graphical data model can be likened to map reading and the ability to read a graphical data model can be seen to be similar to reading a geographically based map (Atkins, 2000). A map may be used by a variety of people for any one of a number of reasons. Each user needs to have either a working knowledge of the map symbols or access to a suitable annotated legend. With this knowledge, the user is able to extract useful information (Atkins, 2000). Therefore, it is worth using a map reading technique to apply to read a graphical data model.

The idea that data models are poor at capturing meaning is also an important point to look at. As Hirschheim *et al.* (1995) point out, 'meaning' is related to human understanding: through meaning we make sense of our feelings, thoughts, and the world around us. Therefore, it can be noted that the meaning of a model is related to a human understanding i.e. an analyst's and user's perspective of a system.

This leads to an interesting question. Is it that a data model is poor at capturing meaning or is it that the people are poor at understanding that meaning? This leads us to consider whether we should concentrate on finding better tools to build the models or better tools to read them?

2.2.3 Communication

This research focuses on data models as tools of communication and concentrates on a tool to better read data models. Generally speaking, communication is the action or process of communicating but the problem is how can we be sure that the meaning that a 'reader' takes from the model, is the one that the 'builder' intends to convey? A business user with strict time constraints and a less than whole hearted focus on drawing pictures, may not be wish to either learn how to use, or even read a new tool. Secondly, unless they are often required to interact with such 'pictures' (i.e. data model diagrams), they are unlikely to acquire the skills of 'reading' them instinctively, though familiarity. When we build a data model, we obviously want it to communicate its meaning. The better the model is as vehicle of communication the clearer the understanding will be.

A simple definition of communication could be: "the transmission from one person to another of a message which is understood by the receiver as the sender intended." (Pinner, 1994 p.1) They describe five levels of communication, namely:

> "1. Intrapersonal, which occurs within ourselves;
> 2. Interpersonal, two-way, face-to-face;
> 3. Mediated, two-way, not face-to-face;
> 4. Person-to-group, one way, face-to-face;
> 5. Mass communication, one way, not face-to-face" (p.8)

Bittner (1988) questions how we define the term 'communication'. He applies the definition of communication by returning to the original source, the dictionary for its definitions of communication:

> "1.To make known; impart; transmit;
> 2. To have an interchange, as of thoughts or ideas;
> 3. Something communicated;
> 4. A system of sending and receiving messages, as by telephone, television or computer;
> 5. A connection;
> 6. A channel or conduit for information; and
> 7. Movement of messages from sender to receiver" (p.9).

He also notes that "Communication is a dynamic process. It is constantly in motion, constantly changing, and constantly being adapted by both the sender (encoder) and receiver (decoder) of communication. Other words, in addition to transmit, transfer, and transact, frequently creep into definition of communication. Among those words are: codes, ideas, intentions, conscious, unconscious, face to face, words, meaning, passage, sender, symbols, interactive, messages, channel, connection, receiver, information, irreversible, and interact"(p. 9).

Moreover, Ruben *et al.* (1982) note that "Communication is the process by which humans and other organisms create, send, receive and interpret information in order to adapt to and affect their environment. Among human beings, the ability to communicate has been extended through the development and use of symbols and a variety of means for their production, dissemination, reception, and storage across time and space. It is through the process of communication that societies are developed, maintained, and changed" (p. 255).

When creating a data model, a data modeller is trying to communicate his view of the data that should be represented in a system. A person (a user) who views the model is interpreting the modeller's meaning. If they are using different rules or beginning from a different position, the two interpretations can de different. Emmert and Donaghy (1981) give an idea about two points of view on the definition of communication:

Definitions based on the intent to influence:

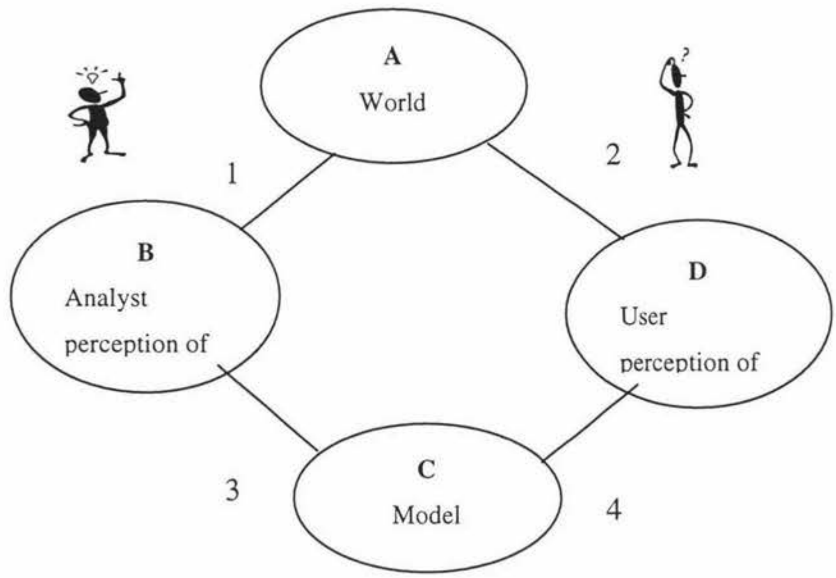
> " Communication... is... all the procedures by which one mind may affect another. This, of course, involves not only written and oral speech, but also music, the pictorial arts, the theater, the ballet, and in fact, all human behavior.
>
> ...the primary goal of most communication is to influence the behaviour of the receiver in some way.
>
> ...what is normally thought of as perfect or effective communication implies the control and manipulation of a receiver of messages by the sender."

Definitions based on perceived meaning:

> "All behaviour-oral, verbal, written verbal, tonal, postural, contextual, tactile- is communicative. Moreover, the behaviour does not have to be intentional for communication to occur. The term 'communication' is broad enough to include unintentionally sent stimuli which receive responses.
>
> ...communication occurs whenever an individual assigns significance or meaning to an internal or external stimulus" (p.45-46).

Smithson (1984), in agreement, notes that "Good communication has two objectives. The first is to obtain total understanding when any message is passed from one individual (or group) to another. The second objective, which is more difficult, is to obtain the response that is required. This is central to the communication process. The first objective can be achieved relatively easily if the individual chooses the most appropriate medium and the most suitable language, whereas the second objective requires the correct approach" (p.3).

Kroenke (1992) makes an important note about data modelling that "databases do not model the real world". Rather, "database are models of users' model of the world"(p.117). This idea may be related to the users' perception of the world (**D**) which shown in Figure 3. The question to ask when evaluating data models is not "Does this design accurately represent the real world?" but rather, "Does this design accurately represent the users' model of his or her environment?" "The goal is to develop a design that fits the user's mental conception of what he or she wants to track. Thus, a computer system needs to model and represent the users' communication with one another. They do not model anything other than that system of tokens and communications"(p. 117).



1 = Analyst perception of the world

2 = User view of the world

3 = Analyst interpretation of his/her world into a model

4 = User interpretation of models as his/her perception of the world

Figure 3 An Analyst/User perception of the world

The diagram at figure 3 illustrates an analyst/user perception and interpretation of the world and highlights the difference that can exist between **B** and **D.**

This leads me to my research, which is interested in discovering how well users understand data models and whether this understanding can be improved. However, there is little research that investigates how the interpretation of a data model is undertaken.

2.3 Data Modelling research

This section discusses previous research which has focussed on various aspects of data modelling.

2.3.1 General studies

Many studies have been completed and reviewed about user's views of data models. Hitchman (1995) finds that aspects of the entity-relationship data modelling method, for the analysis of systems, are not well understood in the commercial domain. He states that "among practitioners in the UK, many do not seem to understand some available semantic constructs of the entity-relationship model"(p. 35).

Hitchman (1997) also discusses the literature on the effectiveness of data modelling in the practitioner domain. He found evidence to suggest that modelling is problematic. Furthermore, he found that current research is limited in what it tells us about modelling in practice. In response to this problem, he proposes DEKAF (A domain expert knowledge approach framework), which is described in the context of overcoming problems of research generalisability. "DEKAF provides both a useful way of understanding and thinking about the data modelling process and a way of making assumptions explicit in a particular practitioner domain"(p.186). He puts forward an argument for adopting a 'softer approach'. "DEKAF proves successful in removing the pressure to model reality and providing the opportunity to think about and explore the domain. It changes from concentrating on the domain, to concentrating on the domain expert–knowledge of what is being modelled not some reality"(p.188).

Siau *et al.* (1995) investigate the importance of the relationship concept for information modelling through a series of psychological experiments. Their study aims to gain a better understanding of the communication problem between end users and designers. It was found that the expert subjects resolve any uncertainty in their understanding of the domain by simply using optional relationships. End users, on the other hand, are not familiar with the intricacy of the relationship concepts. It is likely, therefore, that they will spend more effort and time deciding whether to use mandatory or optional relationships.

Srinivasan and Te'eni (1995) concentrate on data modelling which involves the representation of different types and their relationships. They conduct two laboratory studies in which subjects engage in data modelling to solve a complex problem. Their objective is to explore the problem solving process with an emphasis on the role of problem representation. By using a 'think aloud' process tracing methodology, they examine the data modelling behaviour as a set of activities that are managed by several heuristics. They found that some heuristics were effective in reducing the complexity across levels of abstraction in the problem representation. Overall, these observations help to explain how people deal with complexity in data modelling. The results also suggest that it may be advantageous to design systems that support work at various levels of abstraction and support transitions among those levels.

Atkins (1996) writes that although conceptual models are usually built as a precursor to database design, increasing levels of abstraction and an increased breadth of scope have made it increasingly important than non IS specialists are involved in their construction. The users and sponsors of the system must understand and be able to verify the data structures and business rules while auditors may use the conceptual schema as an indication of database integrity (Atkins quoting Amer, 1993).

Amer (1993) conducts an experiment that examined two conceptual database modelling representations; the entity relationship representation and the relational representation. His study determines their effectiveness in modelling and documenting an accounting database. His results indicate the superiority of the entity relationship data modelling representation over the relational representation in facilitating the audit review of accounting database conceptual models. This superiority, which is due to the graphical nature of the entity relationship representation, suggests that using the entity relationship conceptual modelling representation for database documentation will result in more effective audits in accounting database processing environment.

Firns (1990) points out two potential conflicting objectives when developing a data model. These are a) to adequately and accurately represent in an understandable manner, real world phenomena and relationships that may exist between them and b) to develop the basis of a database structure in which specific instances of the real world phenomena may be represented in the form of data values. He mentions that Ferguson (1988) makes a similar observation and deduces that a data model must satisfy two conditions: the model must adequately support the behaviour and constraints of objects in the real world and the model must be mapped to data structures and process that manipulate these structures, in the information system. (Ferguson, 1988 cited in Firns, 1990)

Olle (1993) states that models must be representation independent. He recommends four forms of representation: diagrams, language, narrative prose, and structured tables. He claims the benefit being, the modeller can select the representation the user is most likely to understand. Likewise, Goldstein and Storey (1990, p.23) find that data models are only useful once the user properly understands them (cited in Hitchman, 1995).

Urquhart and Young (1994) state that ineffective communication during any stages of the development cycle was negatively correlated with project success. They also continue that establishing effective communication with users is essential to providing them with the system they want and point out that a good design based on incorrect requirements produces a good solution to the wrong problem.

Simsion (1994) agrees that technical users of data model also need this accurate understanding in order to specify functions against the data or design appropriate physical data structures. The need is not confined to the users, data modellers too need to be able to recognise the different information implication inherent in alternative conceptual data structures and make informed choices between them.

Jih *et al.* (1989) investigate the effects of relational and entity-relationship data models on the query performance of end users. They point out that in database systems the end user interacts with the database at the external schema level. At this level the user sees only the logical structure of the database that is relevant to his/her work. There is an argument as to whether the relational or the entity relationship model is superior when used in the end user environment. Their basic assumption is "that if one data model was superior to the other, then the superiority of the model would be reflected in the user's query writing performance"(p.257). In addition, this superiority would be demonstrated on both simple and complex tasks. In their study, query writing performance was measured by three variables: number of syntax errors, number of semantic errors and amount of time to complete queries. They found that "subjects using the relational model made fewer syntax errors, but required more time to complete a query. No significant differences in the number of semantic errors were found between the two data models"(p.257). They conclude that, "based on these results, neither the relational nor the entity-relationship data model was clearly superior when used as the interface between a database system and the end user"(p.257).

Batra and Marakas (1995) discuss and compare the perspectives of academic and practitioner communities regarding the application of conceptual data modelling. The focus of their paper is on the conceptual representation, which identifies entities and their interrelationships. From the academic perspective, the empirical finding suggest three points:

1. Conceptual data modelling is unlikely to be used for enterprise-wide models;
2. It is likely that conceptual data modelling will be used for specific applications;
3. The use of a conceptual data model (like E-R) leads to better quality design.

From the practitioner perspective, it was found that authors from the practitioner community have extolled the virtues of conceptual data modelling. However, there is a general lack of any substantive evidence, to suggest that they are being widely used in the applied design environment. From this finding, they suggest that the academic and the practitioner communities can benefit by understanding what each has to offer and appreciating each other's problems and constraints. Therefore, the academic can provide theoretical knowledge and an objective approach, while the practitioner can assess this knowledge in a practical setting and provide feedback and direction for effective research.

2.3.2 Comparative Studies

Tsichritzis and Lochovsky (1982) state that "There has been a good deal of controversy about what is a 'good' or what is the 'best' data model. The debate is similar to arguing the relative merits of programming languages. It is very difficult to argue persuasively that one data model is best uniformly. Each data model has its advantages, depending on who is doing the schema design and the realm in which one is working"(p.17). However, this has not prevented a number of researchers from attempting to compare different formalisms.

2.3.2.1 Creating better data structures

There has been an increasing interest in determining the quality of data structures built with different formalisms. This section intends to provide a summary of these studies illustrated in Table 1.

Authors	Tools	Quality Characteristic	Measurement
Shoval & Even-Chaime, 1987	Normalisation NIAM	1) best result 2) ease of use 3) preferred method	1) correct solution 2) time to complete 3) subjective survey
Jarvenpaa & Machesky, 1989	LDS Relational	1) accuracy 2) time 3) notational understanding	1) correct solution 2) time to complete 3) comprehension question
Batra *et al.*, 1990	EER Relational	1) accuracy in capturing semantic 2) ease of use	1) semantics of correct question subjective survey
Amer, 1993	ER Relational	Understandability by oversight users	Discrepancy checking number of mistakes made in identifying errors
Marche, 1993	Relational	Model stability	Change in relations & attributes over versions
Batra & Antony, 1994	ER	reasons for errors in modelling relationships	Protocol analysis, supported by use of correct solution
Hitchman, 1995	ER	Comprehension of semantics of ER	Survey questionnaire Correct solution
Kim & March, 1995	EER ORM	1) correct 2) consistent 3) complete 4) comprehensible	1) correct solution 2) correct solution 3) correct solution 4) number of correct answer, number of type of the identified errors
Shanks, 1997	ER	1) correctness 2) completeness 3) innovation 4) flexibility 5) understandability 6) overall quality	1) correct syntax 2) correct solution 3) new nouns as entities 4) subjective assessment 5) subjective 6) subjective
Shoval, 1997	EER OO	1) design quality 2) time to complete 3) designer's preference of models 4) comprehension	1) correct of design 2) time to complete 3) preference of model by designers 4) number of correct answer

Table 1 Data modelling studies
(adapted from Atkins, 1998)

Jarvenpaa and Machesky (1989) seek to provide insight into the ease of learning logical data modelling among novices. They performed an experimental study using two tools to examine the three learning experiments: the Logical Data Structure (LDS), which is based on the entity-relationship concept, and the Relational Data Model (RDM).

The conceptual framework in this study is on the point that data modelling requires the analyst to:

1. Partition the original problem into a collection of sub-problems with manageable data structures,
2. Derive the relevant information objects,
3. Understand and represent the relationship among objects, and

4. Formulate questions to refine and discover omissions or inconsistencies in objects and relationships.

For a series of trials, naïve analysts were asked to generate data models using one of the tools. Feedback, regarding the correct model was provided after each trial. Jarvenpaa and Machesky (1989) found that when given structured tasks in a learning environment with repeated feedback, novices are able to rather quickly learn to produce small, good quality models using either the Logical Data Structure or Relational Data Model tools. Comparatively, the LDS tool promoted significantly more top-down directed analysis and resulted in more accurate data models than the RDM tool. The significant differences are explained in terms of the visual appearance of the notation associated with the tools. The results have implications for data modelling training and for the development of data modelling tools.

Batra *et al.* (1990) compare user performance using the relational and the extended entity relationship (EER) models in a representation task. They found that the EER model led to significantly better user performance in modelling binary one-many, binary many-many and ternary one-many-many relationships. There was some evidence that the EER model was better in modelling ternary many-many many relationships and that the relational model was better in modelling unary relationships. In general, subjects found it difficult to model ternary relationships.

Batra and Srinivasan (1992) review the usability of data management environments by synthesising the literature that pertains to data modelling as a representation medium and as a query interface. The primary motivation is to identify a useful framework that would enable a researcher to effectively conceptualise this area and design appropriate research agenda. They highlight that researchers have generally only defined and examined small pieces of the overall problem. Furthermore, they note that the human factor studies of data modelling suggest the following:

1). Among the hierarchical, network and relational data models, it is not clear which data model is the easiest for the best user performance. However, the improved performance is strongly linked to the access language that is normally used with relational systems.

2). Between the relational model and semantic data models, the latter have been found to lead to better performance in many situations.

3). Modelling relationships seem to be the difficult component in a conceptual modelling task. Modelling entities and attributes is relatively easy.

Although the results of these studies are not always clear-cut or consistent, there is a tendency to agree that EER is superior to other, record-based and conceptual models. One of their general conclusions was that usability of data models should be evaluated by their ability to model relationships.

Shoval (1997) works on this conclusion by conducting two experiments, a user comprehension experiment and a design quality experiment. The user comprehension experiment involved a group of students with similar educational backgrounds and divided them into two groups. One group was taught object-oriented modelling, the other extended entity-relationship modelling. Two equivalent schema diagrams were prepared (one for each modelling formalism) and provided to subjects in the appropriate group. Subjects were then provided with a survey designed to elicit their understanding of semantic content of the model. The results showed that the EER model was slightly more comprehensible that the OO model.

The design quality experiment was undertaken in a similar way to the comprehension experiment. A group of information systems students (44 in number) were all trained to use the two modelling formalisms, they were then divided into two groups, subjects in one group designed the first task with OO model and the second task with EER model. Subjects in the other group switched models and tasks. The correctness of each schema was rated using the facet evaluation schema described by Batra *et al.* (1990). In addition, subjects were provided with a survey designed to determine their perceptions of the modelling system. The results showed that complex relationships were designed more correctly in EER, whereas there was no significant differences when designing other modelling constructs. The results also showed that the time taken to produce an

EER model was less than that taken to produce an OO model. The survey also showed that designers preferred using the EER model.

This study agrees with the finding from Bock and Ryan (1993). Bock and Ryan undertake an experimental analysis of the accuracy of two modelling formalisms. This was done with a classical post-test-only experimental design. The 38 subjects from a graduate MIS course on database modelling and design were randomly assigned to two groups. One group was given training in object-oriented modelling, while the other was given training in extended entity relationship modelling. All subjects were then provided with the same textual, natural language description of the application task. The correctness of the models was measured following the protocol outlined by Batra *et al.* (1990).

The results showed that the EER model significantly improved performance in modelling attribute identifiers, one to one relationships and many to many relationships. For all other facets the entity-relationship model performed equally with the object-oriented model.

Shanks (1997) explores the differences between conceptual data models designed by expert and novice data modelling practitioners using the entity relationship modelling formalism. The data models are evaluated using a number of quality factors synthesised from previous empirical studies and frameworks for quality in conceptual modelling. The focus of this study is to extend previous studies by using practitioners as participants and using a number of different quality factors in the evaluation. In this study, each participant was required to read a case study description and prepare a conceptual data model. Each participant was also required to complete a brief questionnaire after data modelling task. The data model qualities in this study are in these criteria: correctness, completeness, innovation, flexibility, understandability and overall quality.

The result showed that data models built by expert data modellers are more correct, complete, innovative, flexible and better than those built by novices. The greatest differences between the models were found in completeness, innovation, flexibility and the naming of relationships. The results suggested that further research into aspects of expertise that lead to such differences and how training courses can narrow the gap between expert and novice performance is required.

However, as Tsichritzis and Lochovsky (1982) suggest "there has been a good deal of controversy about what is a 'good' or what is the 'best' data model"(p.246) and it is very difficult to argue persuasively that one data model is best uniformly. Each formalism has advantages, depending on who is doing the schema design and the realm in which they are working.

2.3.2.2 Cognitive Process (easy of use)

Anderson (1982) has proposed a three-stage learning model for cognitive skills. (Figure 4)

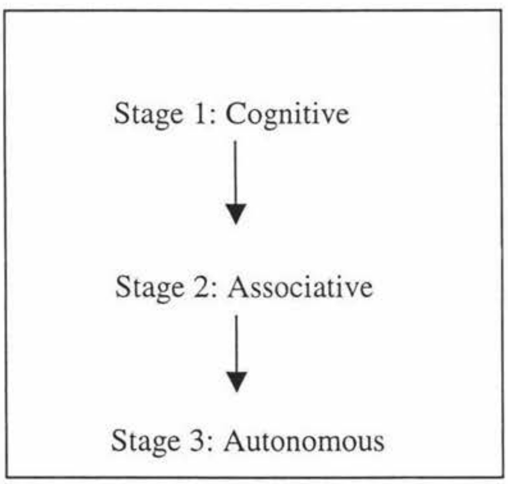


Figure 4 Three stage learning model of cognitive skills

(adapted from Jarvenpaa and Machesky, 1989)

In the first or cognitive stage, the instruction for the skill being taught is encoded as a set of declarative statements about the skill. This is called *declarative knowledge*. In the second or associative stage, a smooth procedure is developed to perform the skill as the compiled statements reveal their procedural form. This is called *procedural knowledge*. In the third or autonomous stage, the procedural form of the undergoes a process of continual refinement, which results in increased speed and accuracy in performance of the skill.

Context problem solving skills are required to successfully specify data requirements and represent them into a data model. Data modelling requires the analysts to

1). Partition the original problem into a collection of subproblems with manageable data structure,

2). Derive the relevant information objects,

3). Understand and represent the relationships among objects, and

4). Formulate questions to refine and discover omissions or inconsistencies in objects and relationships. (Jeffries *et al.,* 1980; Borgida *et al.*, 1985).

Jarvenpaa and Machesky (1989) note that the above cognitive skills and the way the modelling tools facilitate the adoption of these skills are the key questions concerns the determinants of the learning behaviour of the novice and the expert performance.

Batra and Antony (1994) investigate the common errors made by novice (student) modellers. Their findings suggest that the subjects tended to propose a solution based on their initial understanding of the problem, and that once an initial solution was put together, the basic structure of the model rarely changed. They suggest that a significant reason for this lack of adjustment to an initial solution was that since there is no mechanism to inform the designer that the solution is incorrect, there is little motivation to modify the initial solution.

Batra and Sein (1994) work on these finding and recognise that logical database design does not lend itself to self-monitoring by the designer and attempt to provide feedback to novice modellers via a design aid called SERFER (Simulated ER based Feedback system for Relational database) to improve the quality of conceptual and logical relational database. SERFER provides feedback by creating a series of structured natural language statements and a graphical (E-R) form based on their input description of a relation. However, the natural language sentences which comprise the feedback are related not to the information content of the model but to the syntactical rules that are being invoked. They describe the design of SERFER and test its effectiveness in a laboratory experiment using the 'hidden operator' method. Their results showed that this feedback can help users detect and correct certain types of database design errors in modelling ternary relationship. However, no improvement seems possible in the case of unary relationships. The experiment could not determine whether errors could be corrected in modelling binary relationships, since the subjects were reasonably adept and rarely committed serious errors in this experiment.

However, Atkins and Patrick (1998) feel that the nature of the feedback proposed by Batra and Sein is inappropriate. They mention that for novice modellers struggling to come to terms with the differences implicit in various forms of representation, natural language sentences of this form may be confusing. As an alternative, they propose the NaLER (A Natural Language Method for Interpreting Entity-Relationship Models), which they argue encourages a form of self-monitoring by providing modellers with a procedure for creating their own feedback in terms of the information content of their model.

Atkins and Patrick (1998) believe that a more useful form of feedback is to present to the modeller a natural language translation of the information content of the data structure in its modelled form. In this way, a modeller in the act of creating a data model can judge whether a statement that has been in the modelling language says what is was intended to say. They argue that NaLER provides a description of the modelled world that can be easily compared with the real world. The NaLER technique is discussed more fully in the next chapter.

2.3.2.3 Comprehension

Some researchers mention that semantic data models were developed to provide a precise and unambiguous representation of organisational information requirement. These models serve as a communication vehicle between analysts and users. (e.g. Shoval and Frumermann, 1994; Shoval,1997) However, determining correct, consistent and complete information requirement is a difficult task. This section discusses data model studies which focus on their comprehension.

Kim and March (1995) study the effects of different data modelling formalisms on the modelling and validation phases. They compare two semantic data models; the extended entity relationship (EER) and the Natural Language Information Analysis Method (NIAM) model (Halpin, 1995). In this study, the user validation performance consists of two measures; comprehension, which measures syntactic performance and discrepancy checking, which measures semantic performance. The findings of this research are encouraging for IS practitioners. With a small amount of training, users were able to read and validate application data models of nontrivial size and complexity in both methods. They also suggest that when more users become data model-literate (capable of validating an application data model produced by analysts), the analyst's job of producing a complete and correct representation of user information requirements will be made much easier.

Shoval (1997) examines user comprehension on two data models (EER and OO). He conducts an experiment by using two groups of users, each studying a different model. The study emphasized comprehension of the schema diagram. The purpose was to find out not only if there was a difference in overall comprehension between the two schema diagrams, but also if there was a difference in the specific construct types. Each subject was given a schema diagram, according to the model she/he studied and a questionnaire. Each user was expected to review the diagram and mark each statement 'true' or 'false'. The statements were classified according to different construct types: a) attributes of entities/objects, b) binary-relationships, c) two binary relationships, d) ternary-relationship and e) other facts - with no direct relationships. The level of comprehension was measured by counting the number of correct answers within each for group of the categories of statements, and for all statements together. He concludes that ERR ternary relationships are easier to understand, whereas for OO, only non-specific facts are easier to understand, with no significant difference for the other constructs.

2.3.2.4 Type of user and/or modeller (expert/novice)

The development of a conceptual data model is a design activity influenced by various kinds of knowledge including application domain knowledge, data modelling representation and process knowledge. When designing a conceptual data model, expert data modellers develop a holistic understanding of a problem, categorise problem descriptions into standard abstractions and reuse generic data models from their previous experience (Batra and Davis, 1992). Novice data modellers follow simples guidelines such as "make a list of all the nouns..." (Veryard, 1984, p.15).

Batra and Davis (1992) explore the similarities and differences between experts and novices designers engaged in a conceptual data modelling in data base design. The study employed the 'thinking aloud' protocol methodology for data gathering and analysis. Each subject was given the case, which was semantically rich, and included the following facets: entity, category, identifier, descriptor and the following types of relationship: unary many-many, binary many-many, and ternary many-many-many. The study focused on the translation of user requirements (external view level) into a conceptual data model. The subjects seemed to operate at three distinct levels of abstraction: enterprise, recognition and representation level. Then, they iterated among these levels over the time they worked on the task. These levels were applicable to both experts and novices through the frequency and pattern of iterations and the total time spent at each level differed between experts and novices. This study found that the experts focused on generating a holistic of the problem before developing the conceptual model. The experts were able to categorise problem descriptions into standard abstractions. The novices tended to have more errors in their solutions largely due to their inability to integrate the various parts of the problem description and map them into appropriate knowledge structures. This study also found that the experts and novices behaviour were similar in terms of modelling facets like entity, identifier, descriptor and binary relationship. However, there were some differences in modelling ternary relationship, but many differences in the modelling of unary relationships.

Shanks (1997) also investigates the differences between conceptual data models designed by expert and novice data modelling practitioners using the entity relationship formalism. The data models are evaluated using a number of quality factors synthesised from previous empirical studies and frameworks for quality in conceptual modelling. However, this study extended the previous studies (Juhn and Naumann, 1985; Jarvenpaa and Machesky, 1986; Shoval and Even-Chaime, 1987, Batra et al, 1990; Batra and Davis, 1992; Kim and March, 1995) in several ways. For example, all participants in this study were information systems practitioners rather than students, the evaluation of the models are more comprehensive and based on several quality factors synthesised from a number of quality frameworks and the ER formalism used was the Martin (1987) standard, which is widely used in practice, rather than the Chen standard commonly used in other studies. (Hitchman, (1995) found that only 5% of practitioners used the Chen notation)

This study found that data models built by expert data modellers are more correct, complete, innovative, flexible and better understood than those built by novices.

2.3.3 Data Models as Communication Tools

Many authors write that system success is dependent on effective communication between the systems designers and users(e.g. Burch, 1992; Batra and Davis, 1992; Shanks, 1997; Shoval and Frumermann, 1994). However, Campbell (1992) points out that "there is a conflict between the desire to address design issues and the need to create a form of data model with which the business user is comfortable"(p. 13). He also suggests that "in concentrating on the design issues, and requiring users to think in these terms, we are practicing a form of technological arrogance. We should, instead, keep the form of the model appropriate to the task in hand. When we are communicating with users, we should not be involving them in our thoughts on design issues"(p.13). It is suggested that this will result in better communication at the analysis stage and reduce the likelihood of costly errors reaching the design stage.

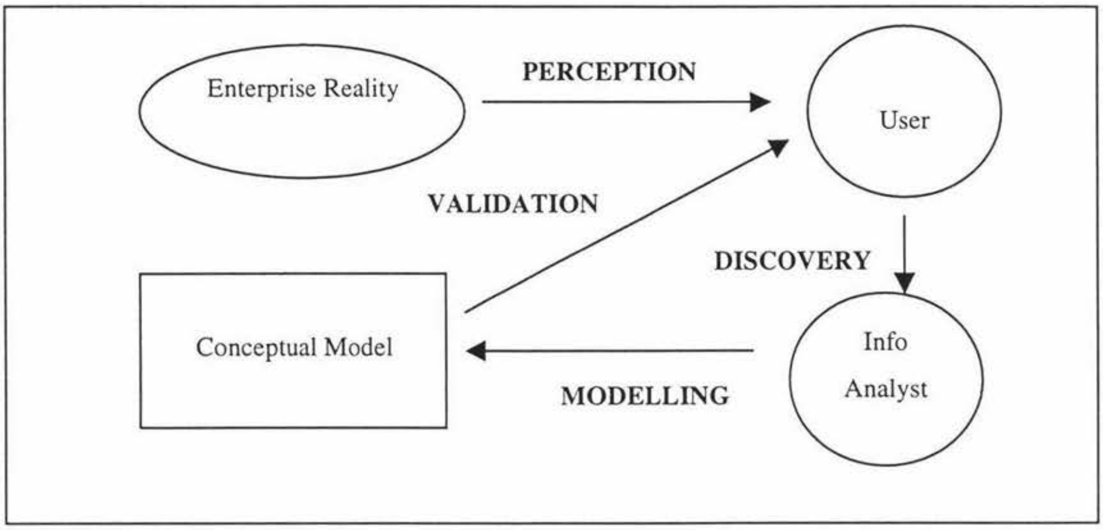


Figure 5 A process model of information requirement determination

(adapted from Kim and March, 1995)

Figure 5 shows a four-phase process model for requirements determination:

1. **Perception** –Users perceive the enterprise reality. The same enterprise reality may be perceived differently by different users (inconsistency). Any one user may perceive only a part of the reality (incompleteness),
2. **Discovery** – Analysts interact with users to elicit their perceptions,
3. **Modelling** – Based on the information identified in the discovery phase, analysts build a formal, conceptual model (representation) of the enterprise reality. This model serves as a communication vehicle between analysts and users.
4. **Validation** – Before concluding the model is correct, consistent, and complete, it must be validated.

Users must comprehend or understand the meaning of the model in order to identify discrepancies between the model and their knowledge of reality. (Kim and March, 1995)

Campbell (1992) states that to satisfy the user communication aspect, there must be an appropriate way of showing and demonstrating the developing system to the users. It is clear that because of this role as a communication tool, research is often focussed on either improving the E-R model or finding a better alternative (Blaha *et al.*, 1988;

Campbell, 1992; Coad and Yourdon, 1991; Elmasri *et al.*, 1985; Hammer and McLeod, 1981; Moody and Osianlis, 1996; Schenck and Wilson, 1994; Smith and Smith, 1997; Teorey *et al.*, 1986)

Campbell (1992) also highlights the fact that data models are useful in helping analysts to "adopt the users' view of the world"(p.13) but it is important that the data analyst should explain the model in a language that the user are familiar with. He discusses the use of ER models as user communication tools. He contends that a large number of the errors that occur in database design are a result of poor analyst/user communication, and that there is a conflict between the desire to address design issues and the need to create a form of data model with which the business user is comfortable. In response to this problem, Campbell proposes a resource-activity diagram, which he describes as being free from design-specific issues, and as being more intuitive to the business user. He states that there is a need for a form of data modelling tuned to analyst/user communication, and that the resource-activity diagram involves relatively minor changes which are designed to make the process more intuitive to the user. These changes involve transferring entities into resources and relationships into activities, as well as the inclusion of symbols for the representation of super-type/sub-type and aggregation relationships.

However, Connolly *et al.* (1996) suggest that, the adoption of yet another modelling formalism does not appear to be warranted solely to enhance analyst/user communication. The process of converting the model from one formalism to another may introduce more errors that the suspected analyst/user mis-communication. It may also be worthy of note that many users may have come into contact with the E-R Model in past projects and as such trying to teach them another formalism may confuse them even more.

According to Burch (1992), "conceptual data models play a substantial and essential communication role; one of the major reasons for creating models is to facilitate communication with the user"(p.35). Burch also adds that the model must serve as a communication medium between users and systems professionals.

Batra and Davis (1992) write that "the conceptual model is usually easy to understand and can from the basis for communication with users"(p.83). Burch, 1992, is in agreement, and suggests that written, oral and graphic communication are all necessary for communicating documented deliverables effectively to all participants in the development of the new system.

Thus, based on the discussion in the previous sections and the literature survey, it can be noted that the ability to understand the semantics or information content of data models, especially to 'read' them is become an interesting topic in the information system area. Despite the increased need, from an increasing range of people to have this skill, little attention has been given to it. This research not only investigates an important aspect of data models as communication tools but also reports the experimental design and results of a study that was conducted to test modeller's ability to 'read' a data model, both before and after the introduction of a 'reading' technique, NaLER, which is discussed in the next chapter.

Chapter Three - Reading E-R Models

3.1 Standard Technique

E-R models are built on three basic concepts: entities, attributes and relationships (Chen, 1976). The process of designing an appropriate set of entities and relationships to meet a business problem is called entity relationship modelling (or E-R modelling). A data model in this format is called an E-R model, and the diagram an E-R diagram (Simsion, 1994).

An **entity** is a class of persons, place, objects, events or concepts about which we need to capture and store data (Whitten *et al.,* 1998; Ricardo, 1990).

An entity can be represented by box with rounded corner (Simsion, 1994; Whitten *et al.,* 1998) or by a rectangular box in some textbooks (Hawryszkiewycz, 1998; Connolly *et al.*, 1995).

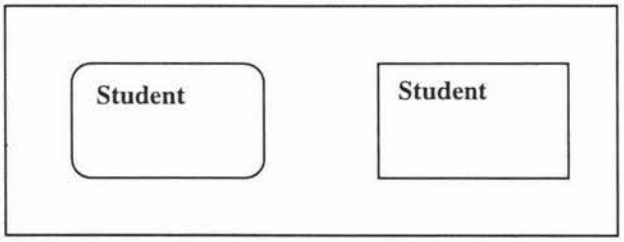


Figure 6 Alternative Entities Notations

An **attribute** is a descriptive property or characteristic of an entity (Whitten *et al.*, 1998; Ricardo, 1990).

An attribute can be represented in various graphical forms as illustrated in Figure 7.

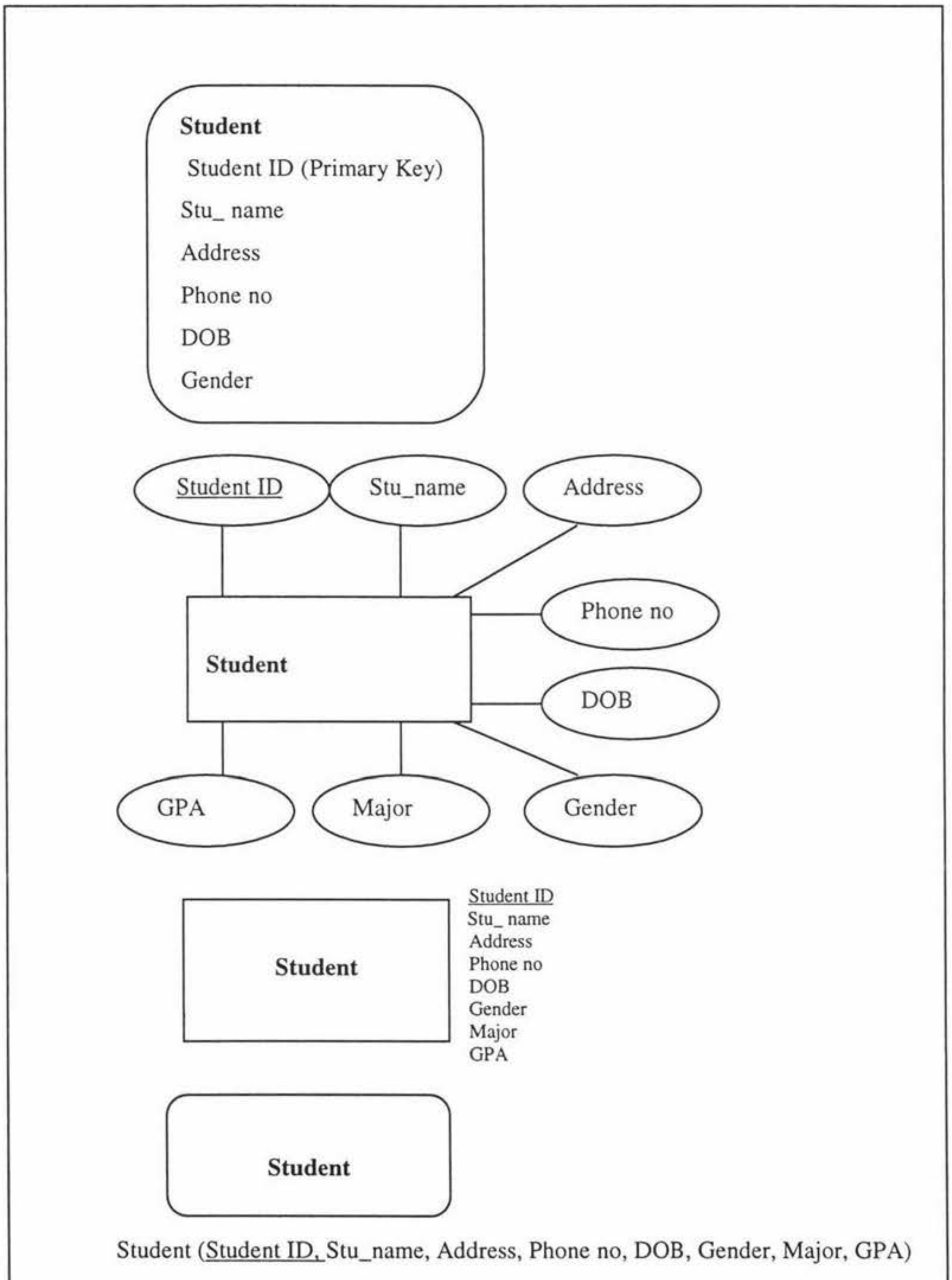


Figure 7 Alternative Attributes Notations

A relationship is a natural business association that exists between one or more entities. The relationship may represent an event that links the entities or merely a logical affinity that exists between the entities (Whitten *et al.*, 1998; Ricardo, 1990).

A relationship can be represented in different forms either as a diamond shape or as a line links between entities, as illustrated in Figure 8.

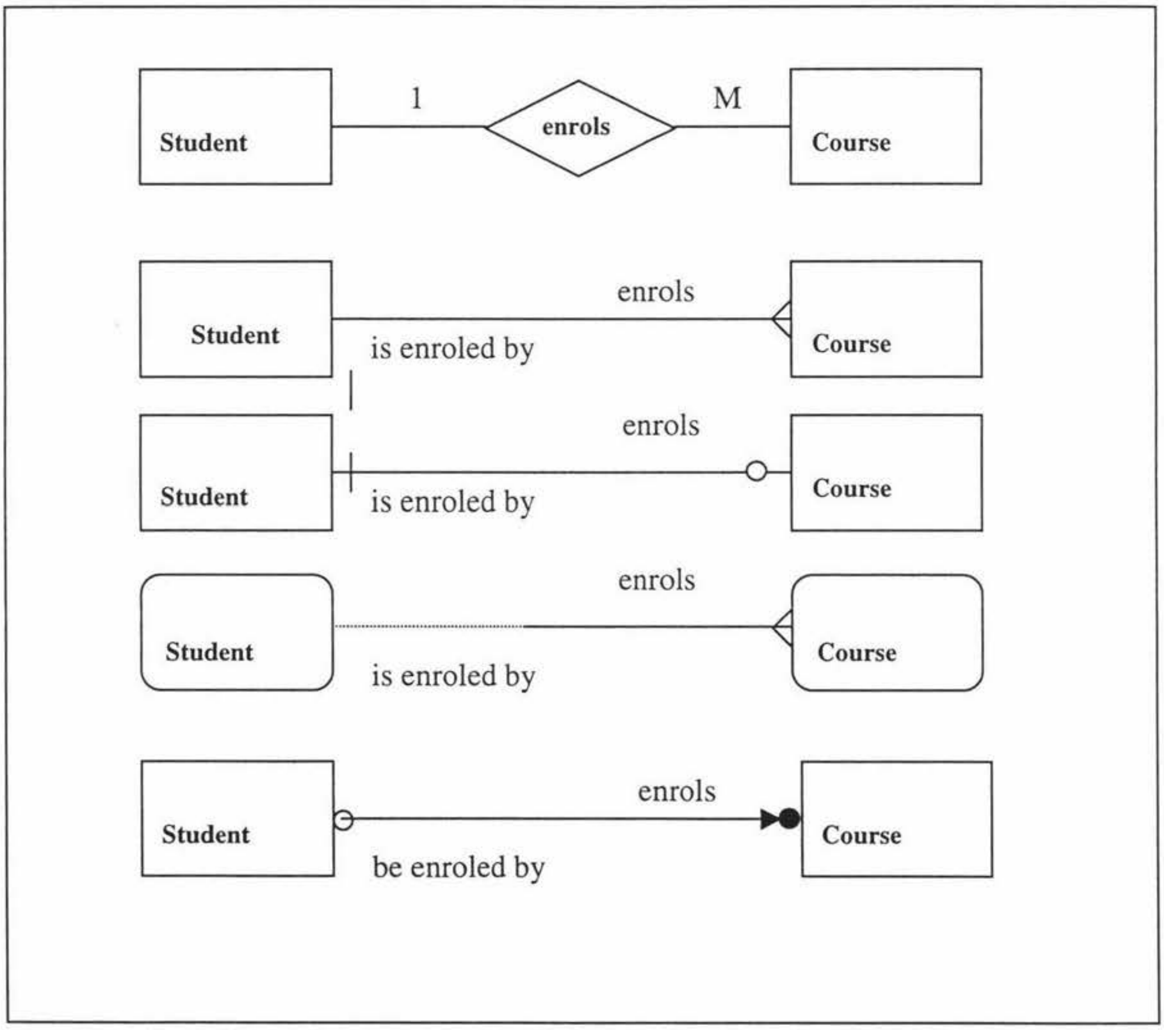


Figure 8 Alternative Relationship Notations

It should be noted that there is wide variety of notations used to represent E-R models, but there is no formally taught method for reading them. It is assumed that if data modellers can build an E-R model they should be able to 'read' it. So the ability to read a data model relies on the modellers' intuition and experience.

A formal syntax to read an E-R model is described by Barker (1990) using two-way sentences as follows.

Each (and every) ENTITY-A { may be / must be } end- name-1

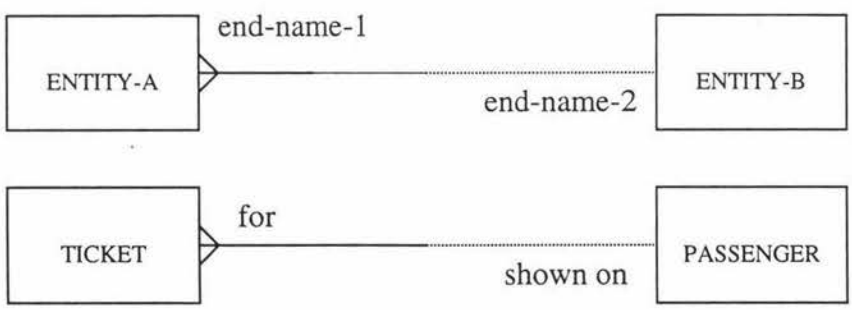
{ ONE AND ONLY ONE / ONE OR MORE } { ENTITY-B (ever). / ENTITY-B (plural). }

And conversely:

Each (and every) ENTITY-B { may be / must be } end-name-2

{ ONE AND ONLY ONE / ONE OR MORE } { ENTITY-A (ever). / ENTITY_A (plural). }

e.g.

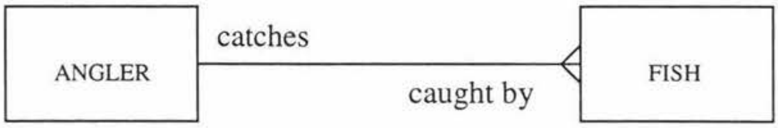


Each TICKET must be for one and only one PASSENGER.

Each PASSENGER may be shown on one or more TICKETS.

The above syntax is good for some purposes, but to really check out the details perhaps as a means of constructing a description of the model, it may be ambiguous.

e.g.



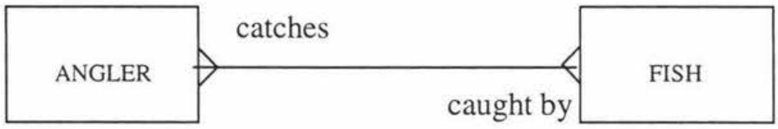
Each ANGLER may catch one or more FISH.

Each FISH must be caught by one ANGLER.

Reading this E-R diagram by using the formal syntax it looks correct, but if we give Fish_id as the primary key for Fish. At this stage, we will find some ambiguous meaning because it will not make any sense that Fish has Fish_id as the primary key. But we will not notice this incorrect semantic unless we put the primary key together in the sentence.

However, if we change the primary key from Fish_id to Fish_code_type. Perhaps it may provide a better meaning that can be compared with the 'real' world.

The diagram will be changed to



Each ANGLER may catch one or more FISH.

Each FISH may be caught by one or more ANGLERS.

Giving a primary key together with the entity when reading a model may help the modeller and the user to check the meaning of the information content of the models. However, there is no formally taught when using the standard technique.

Also relationships are not always named at both end, and sometimes just in one direction or sometimes not named at all. This may lead the users more difficult to read a model.

Nordbotten and Crosby (1996) note that it is assumed that the users can correctly interpret the information provided in the graphical model. However, the users may not always view the model in the way anticipated by the system designers. Nordbotten and Crosby (1996) also point out that "very little is known about how graphic models are read or how the graphic style of the model, might influence its perception and comprehension" (p.43).

This leads to the question that what will happen when we ask a user to 'read' a model? As there is a lack of formal instruction for reading a model, the users may not always 'read' a model in the way that we anticipate and perhaps the details may be missed. Atkins and Patrick (1998) suggest that these details may be missed because the users' deep understanding is lacking. Therefore, they propose the NaLER technique in order to help the users to focus on the details of the information content of the models.

3.2 NaLER Technique

In order to try and address this lack of formal instruction for 'reading' E-R models, the NaLER (Natural language for E-R) technique has been developed. It was originally conceived as a pragmatic way of encouraging final year undergraduate students, taking an advanced data modelling course at Massey University, New Zealand, to judge whether the data models that they had created made the statements that they had intended (Atkins & Patrick, 1998).

3.2.1 Uses of the NaLER technique

The creation of the NaLER technique had two objectives. It was intended as an aid to the modeller within the design process and as a means of presenting the information content of the design model to users.

The first objective, was to provide a self-monitoring mechanism whereby modellers could create their own feedback in terms of the information content of their created structures. Based on the premise that "modelling is essentially making statements in some language" (Lindland *et al.*, 1994 p.43) and that "all representation is an act of knowledge construction" (MacEachren, 1995 p.vii), a useful form of feedback would be to present a natural language translation of the information content of the data structure to the modeller. In this way a designer, in the act of creating a data model, can judge whether a statement created in the modelling language, says what it was intended to say (Atkins, 2000).

The second objective was to provide a means of constructing a description of the modelled world that can be compared with the 'real' word. NaLER provides a means to create sentences for the verification of the semantics of the E-R model. In this way, it can be confirmed that user requirements are being supported (Atkins, 2000).

3.2.2 The NaLER Construction

NaLER was designed for use with a relational data model, represented diagrammatically by an ER diagram and supported by a data dictionary, as produced by contemporary CASE tools.

3.2.2.1 Assumptions

According to Atkins (2000), there are a number of additional desirable elements, it is recognised that not all CASE tools provide the same level of documentation and when used in support of the modelling process, not all information may have been recorded. Therefore, while full documentation is recommended, if the model is incomplete in some way there are a number of assumptions that can be made. Atkins and Patrick (1998) note that these assumptions are,

A1 – Relationships are represented by the lines between entities.

A2 – Relationships are named in at least by one direction.

A3 – Cardinality is shown on the diagram in some form.

A4 – Relationships are mandatory unless clearly marked as optional.

A5 – Entities are named.

A6 – Entities have a unique identifier or primary key.

A7 – A 1 to 1 mandatory relationship is implicit in the position of an attribute in an entity but if the attribute is described as nullable the relationship is optional.

A8 – Attributes with the same name but placed in different entities may relate to the same 'real world' concept. If it seems reasonable that they do then if one such attribute is a primary key in one entity then it is a foreign key in any others.

A9 – An entity whose primary key consists of two or more foreign keys may be resolving a many to many relationship between the entities of which those attributes are foreign.

These are assumptions that should not be required where the modeller is using NaLER has against his/her own model. However, when 'reading' another's model, there could be times where some of these conditions are not met. As a result, Atkins and Patrick (1998) propose some additional assumptions which can assist in maximising the amount of useful information that can be extracted. These are,

A10 – For unnamed relationships

1) 1-M, allocate the name 'has' unless a more intuitive one is apparent.
2) 1-1, allocate the name 'has' unless the two entities have an identical primary key in which case name the relationship 'is'.
3) M-M, it is not usually possible to allocate a meaningful name with any confidence, although 'has' may be tried.

A11 – If an entity is not named but has attributes assigned to it, it may be possible to deduce a name from the attributes, particularly from the name of the primary key attribute(s).

A12 – If a foreign key attribute exists without an existing relationship line, then the relationship should be treated as missing and unnamed. It may be useful to create the relevant sentence using 'has' as the relationship name and a cardinality of 1 – M.

Atkins and Patrick (1998) note that any adjustments that are made on these assumptions should be clearly marked in the final natural language description. The usual method is to bold or underline them.

3.2.2.2 Procedure

This section describes the steps of the NaLER technique proposed by Atkins and Patrick (1998). For reference, all the statements are numbered as they are constructed, however, the ordering of the statements is not significant.

Figure 9 shows a simple ER diagram on which most of the examples in this section are based.

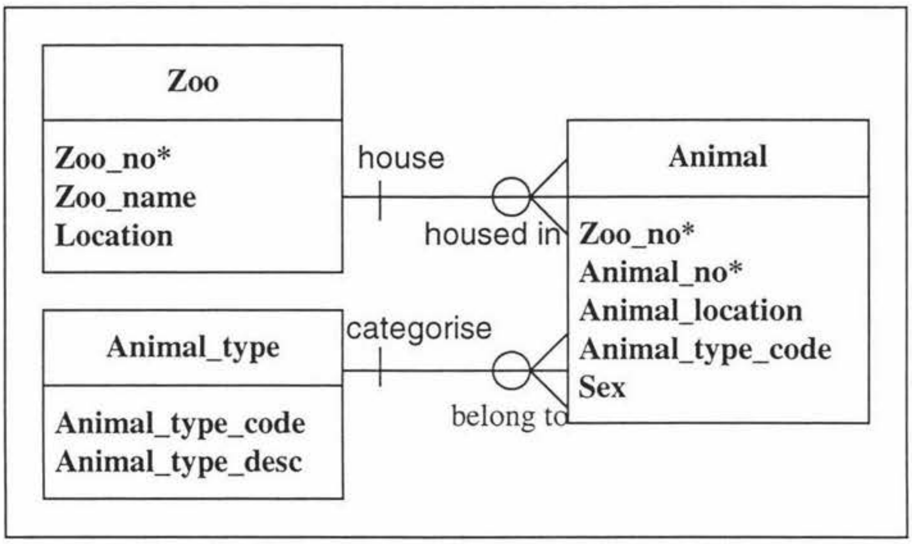


Figure 9 Example ER diagram (Atkins and Patrick, 1998)

Step 1 – Identify and document the diagram conventions.

This step is to record and clarify what notation has been used to construct the diagram and data dictionary. Any inconsistent use of notation in the model should also be noted.

Step 2 – Perform a syntax check of the model in accordance with the above assumptions and make (and mark) any necessary adjustments.

This step is to identify any areas in the model where assumptions from the above list need to be made. Any assumptions that are made should be recorded. This step encourages the designers to double check their own work for syntactic completeness and correctness as well any model they 'read'.

Step 3 – For each entity,

Construct a sentence for the primary key attribute(s) as:

Each <entity-name> is uniquely identified by <primary key>.

e.g. S1: Each Zoo is uniquely identified by zoo-no.

This task is intended to focus on the appropriateness of the chosen primary key and the entity name.

3.2 Construct sentence for each attribute in the entity that is not a primary or foreign key as:

One <entity name> identified by <PK> must have one <attribute name>.

e.g. **S2: One Zoo identified by zoo-no must have one Zoo name.**

or **One Zoo (zoo-no) must have one Zoo name.**

This step works on the assumption that the model is in at least First Normal Form. This step is also valuable in checking the level of normalisation by providing an intuitive means of thinking and talking about functional dependencies.

3.3 Construct a sentence for each binary relationship, including recursive ones, that the entity participates in as:

One <entity-name1> identified by <PK> <optionality> <relationship-name> <cardinality> <entity-name2> identified by <foreign key>.

e.g. **S3: One Zoo identified by zoo-no may house one or more Animals identified by zoo-no, animal-no.**

or **One Zoo (zoo-no) may house one or more Animal (zoo-no, animal-no).**

The form of these sentences may be familiar as a variation of the business rules used in some methodologies. However, the inclusion of the key attribute(s) makes a significant contribution to the understanding of the model (Atkins, 2000).

Step 4 – Populate the sentences from step 3.2 and 3.3 with valid examples

Example from 3.2:

Zoo 111 has the name 'regents Park'.

Zoo 222 has the name San Diego. etc.

Example from 3.3:

Zoo 111 houses animal 111, 90.

Zoo 222 houses animal 222, 100. etc.

This task is based on a technique in the NIAM-CSDP (Nijssen & Halpin, 1989), where the 'fact types' are derived from elementary natural language 'facts', such as those created here. This step reverses the procedure and allows the 'reader' to relate the model's statements directly to examples in the UOD.

Atkins and Patrick (1998) note that these four steps are sufficient for the 'reading' of a significant proportion of the most E-R/Relational models. Although, they also propose some additional steps for more complex entities. However, there were not required in the experiments for this research and are therefore not discussed here.

Step 5: Produce a NaLER Description by listing all the constructed sentences illustrating them with examples where appropriate.

S1: Each Zoo is uniquely identified by one zoo-no.
S2: One Zoo (zoo-no) must have one location.
S3: One Zoo (zoo-no) must have one Zoo-name.
S4: One Zoo (zoo-no) may house one or more Animal (zoo-no, animal-no).
S5: Each Animal is uniquely identified by one zoo-no, animal-no.
S6: One Animal (zoo-no, animal-no) must have one animal-location.
S7: One Animal (zoo-no, animal-no) must be housed in one Zoo (zoo-no).
S8: One Animal (zoo-no, animal-no) must belong to one Animal-Type (animal-type-code).
S9: Each Animal-Type is uniquely identified by one animal-type-code.
S10: One Animal-Type (animal-type-code) must have one animal-type-location.
S11: One Animal-Type (animal-type-code) may categorised one or more Animal (zoo-no, animal-no).

Figure 10 NaLER sentences for E-R diagram in Figure 9(Atkins & Patrick, 1998)

This step formalises the previous work into a standard description by listing all constructed sentences.

Using the NaLER technique, might encourage the users to focus on the details of the information content of the model. It might also help them to develop their deep understanding of a model's semantics when they 'read' a model. The NaLER technique described in this chapter would be used as the treatment for the experiment, which is discussed in the next chapter.

Chapter Four - The Experiment

4.1 Section 1 – Research Methodology

4.1.1 Overview

This research is focused on the conceptual data model, which can be considered as the cornerstone of information systems analysis and design (Siau *et al.*, 1995). In addition, the conceptual model facilitates the communication between the designers and the users by providing them with a conceptual representation of the enterprise that does not include many of the details of how that data is physically stored. There are a number of researchers in information systems who use experimental research to compare different data models (Juhn and Naumann, 1985; Jarvenpaa and Machesky, 1989; Shoval and Even-Chaime, 1987; Jih *et al.*, 1989; Batra *et al.*, 1990; Amer, 1993; Shoval and Frumermann, 1994; Siau *et al.*, 1995; Kim and March, 1995; Shoval, 1997; Shanks, 1997). Most of the researches compared the EER model with the classical models, mainly the relational, and with other semantic models. However, there has been little research on the use of data models as tools of communication.

My research concentrates on the translation of an understanding of the business situation into a representational form for a conceptual database. This schema is usually represented in a diagrammatic form, and serves as a communication tool between developers and users. Once approved by users, it is converted into a specific database schema, depending on the data model and the DBMS used for implementation. However, the diagram can be misunderstood or misread by the users and in order to explore this issue, two experiments were conducted. The purpose of this study is to compare the understandability of models by students before and after learning a technique NaLER (Natural Language for Entity Relationship Models) to assist their 'reading' of E-R (Entity Relationship) model (Atkins and Patrick, 1998).

The case description (see Appendix 1) and the data model used in this research were adapted from Batra *et al.*, 1990. However, the original model from Batra *et al.* was an

EER (Extended Entity Relationship) model and utilised the Chen notation. This was transformed into an ER/R representation with which the students were more familiar. (see Appendix 2)

The reason that an Entity Relationship model was chosen to represent the conceptual data model for this research was from the result of several empirical studies (Batra and Srinivasan, 1992; Jarvenpaa, and Machesky, 1989; Jih *et al.*, 1989) generally indicate that the E-R model, compared with other models, is easier to understand by the users and they perform better when using the E-R model.

4.1.2 Reason why an experiment approach was considered the most suitable

Experimental research is set up in order to test an hypothesis or a theory. The experiment is usually designed so that the researcher has as much control as possible over the conditions under which the actual tests take place. The principle of the experimental research is the identification of the precise relationships between variables in a designed, controlled environment using quantitative analytical techniques. The advantage is that the researcher can isolate and control a small number of variables, which may then be studied intensively. By experimenting it is possible to establish the validity of theories before taking positive action on them. This can not only save time, effort and resources, but in many cases the experiment can throw light on important aspects which need to be understood if the theory is to be fully exploited (Moore, 1983; Galliers, 1992).

In this research, two laboratory experiments were conducted to address the research questions. Traditionally, the laboratory has been the setting for most human factor research. In fact, the laboratory-based research strategy has been used in many data modelling studies comparing user performance (Juhn and Naumann, 1985; Jarvenpaa and Machesky, 1989; Shoval and Even-Chaime, 1987; Jih *et al.*, 1989; Batra *et al.*, 1990; Amer, 1993; Shoval and Frumermann, 1994; Siau *et al.*, 1995; Kim and March, 1995; Shoval, 1997; Shanks, 1997). The main advantages of the laboratory research method are high interval validity, precise definition and manipulation of independent variables over a feasible range and control for nuisance or extraneous variables.

The data from the experiment was analysed quantitatively (i.e. by the collection and analysis of data in numeric form). Quantitative research tends to emphasize relatively large-scale and representative sets of data and is presented or perceived as being about the gathering of 'facts' (Blaxter *et al.*, 1996).

4.1.3 Research model

The broad purpose of this research is to compare the subjects' understanding of a data model before and after learning the NaLER technique.

The research model is shown in Figure 11. The independent, dependent and control variables are explained as follow.

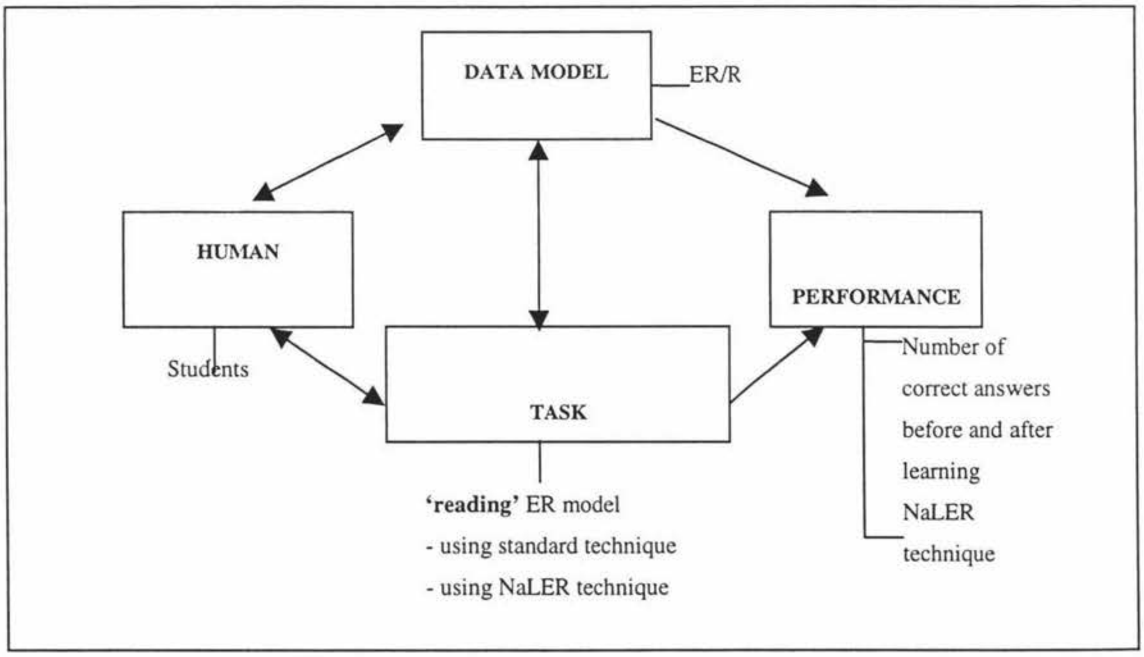


Figure 11 Research model (adapted form Batra *et al.*, 1990)

4.1.4 Independent Variable

NaLER was selected as the treatment for the experimental groups. NaLER was a new technique, which was proposed to help in reading an ER model by using natural language to interpret them. It has been described in Chapter three.

4.1.5 Dependent variable

The performance variable, i.e. the understandability of the subjects using two different techniques, would be measured by the number of correct answers given to each set of questions. The experiment would be interested in two levels of understanding; accurate understanding and comprehensive understanding. Accurate understanding would be measured by the number of correct answers given to a set of the simple questions while comprehensive understanding would be measured by the number of correct answers given to set of medium and difficult questions.

The confidence of the subjects would be assessed in two ways; confidence that questions had been answered correctly and confidence that the model had been understood completely. Confidence would be measured by the mean score for each confidence question. Subjects would be asked to self-assess their level of confidence using a five point likert scale.

4.1.6 Control variable

Task and Human were selected as the control variables in the experiment. Since the purpose of the research was to compare the level of understanding of the subjects before and after learning the NaLER technique. The task of each subject would be to 'read' a data model by using the standard technique in the first experiment and the NaLER technique in the second experiment. Each subject would be required to write a brief description of the data model they were given in the first experiment, and to write the NaLER sentences in the second. They would then be asked to answer a questionnaire that consisted of a set of 'true' and 'false' statements about facts in the data model. Finally, each subject would be asked to rate their confidence levels.

4.2 Section 2 – Research Process

4.2.1 The questionnaire

The questionnaire consisted of twenty 'true and 'false' statements about facts in the ER/R model. The reasons for using 'true' or 'false' statements were because it was easy to mark and to grade the correct answers and it required less time for the subjects to complete. The idea of this questionnaire was based on the work of Shoval (1997), and the purpose was to find out not only if there was a difference in overall understanding before and after using NaLER technique, but also if there was a difference in the specific question type (simple, medium and difficult). In order to avoid bias in the use of questionnaire, two sets of questionnaires were prepared based on a common set of thirty-four statements (see Appendix 3). Each questionnaire (either Set A[1] or Set B[2]) consisted of ten statements to test simple understanding, five statements to test medium understanding and five statements to test deep understanding. Therefore, both sets of the questionnaire consisted of all three types of questions in equal proportion. In addition, there were six questions, which testing simple understanding, which appeared in both questionnaires.

The three types of the statements (simple, medium and difficult) were measured by the degree of the understanding with which the subjects 'read' the data model. The criteria designed for each type of statement is shown in Table 2. Simple statements could be interpreted directly by looking at only one entity, while the medium statements were more complicated to interpret and required to look at more than one entity. Finally, the difficult statements were the most complicated and could not be read directly by only looking at the diagram, the subjects had to be able to look for the detail of more than two entities and link them together. For example, the statement "Each employee must have an employee #." was graded as a simple statement because it could be interpreted directly by looking at just one entity. While the statement "An employee working in a

---

[1] See Appendix 4

[2] See Appendix 5

particular city must be working on a project." was graded as a difficult statement because it could not be interpreted directly. It needed to be focused on the information content of the model in detail.

Type of Statement	Criteria
Simple	Can be interpreted directly by looking at only one entity and its attributes.
Medium	Can be interpreted by looking at one or two entities and their attributes.
Difficult	Can be interpreted by looking at the details of the information content of more than two entities and their attributes.

Table 2 The criteria of 3 types statements

The likert scales were used to check the subjects confidence on their understanding of the data model that they were given. The first measure (confidence correct) asked the subjects how accurately they felt they had understood the model. The second (confidence complete) asked the subjects to assess how completely they felt they had understood the model.

4.2.2 Pilot study

In order to test out the questionnaire, a pilot study was conducted with a group of postgraduate students taking a conceptual modelling course. They were asked to read the data model, which was to be used in the actual experiments. The main purpose of the pilot study was to identify any problems either with the data model or the questionnaire.

The problems identified in the pilot study included several ambiguous statements and some confusion over the notation used in the graphical representation of the ER model. Based on the feedback from the pilot study subjects, the documents in the real experiment were refined and improved. The data collected from the pilot study was not used for analysis.

4.2.3 Select the subjects

A random assignment procedure was used to assign the subjects into two different groups. The purpose of using the random assignment was to provide an unbiased method for placing the subjects into groups so that there was a high probability of characteristics of each group being about the same (Sproull, 1988; Graziano and Raulin, 1993).

The random assignment was done by using the last digit of the student ID. The subjects who had the last digit of their student ID. in the range 0 - 4 were assigned to group one, while the students who had the last digit of their student ID. In the range 5 - 9 were assigned to group two.

4.2.4 The procedure

Since the experiments require the participation of humans, approval was sought from the university's ethical committee (see Appendix 6). The participants' informed consent was gained through a participant information sheet (see Appendix 7) and consent form (see Appendix 8). This form advised the participants that they might ask questions at any point during the experiment, and might refuse further participation at any stage.

Prior to the experiments, the subjects were given the information sheet and the consent form to fill and were informed that the experiments were conducted with the approval of the university's ethics committee.

In the first experiment, the subjects were given the model and were asked to write a brief description[3] of it. This was simply intended to focus their attention on the model and to gauge their initial understanding. Then they were asked to answer the questions that consisted of a set of 'true' and 'false' statements about the facts represented in the data model. The subjects had no way of knowing which questions were considered to

[3] Transcripts of these descriptions can be found at Appendix 9

be simple, medium or difficult. The subjects in group one were given the questionnaire set A while the subjects in group two were given the questionnaire set B.

Over the next few weeks, the subjects were taught NaLER as part of the usual content of their data modelling course.

The second experiment was conducted after the subjects had been taught this new technique. In the second experiment, the subjects were given the same data model but they were given the alternative set of questions to answer (i.e. the subjects in group one were given the questionnaire set B and the subjects in group two were given the questionnaire set A). They were also asked to write the NaLER sentences[4] before they answered the questions. This would help the subjects to investigate the semantics of the model in detail and perhaps improve their ability to focus on what the model was actually recording in terms of information content. Finally, they were asked to indicate their confidence on their understanding of the data model. The procedure described above is illustrated at Figure 12.

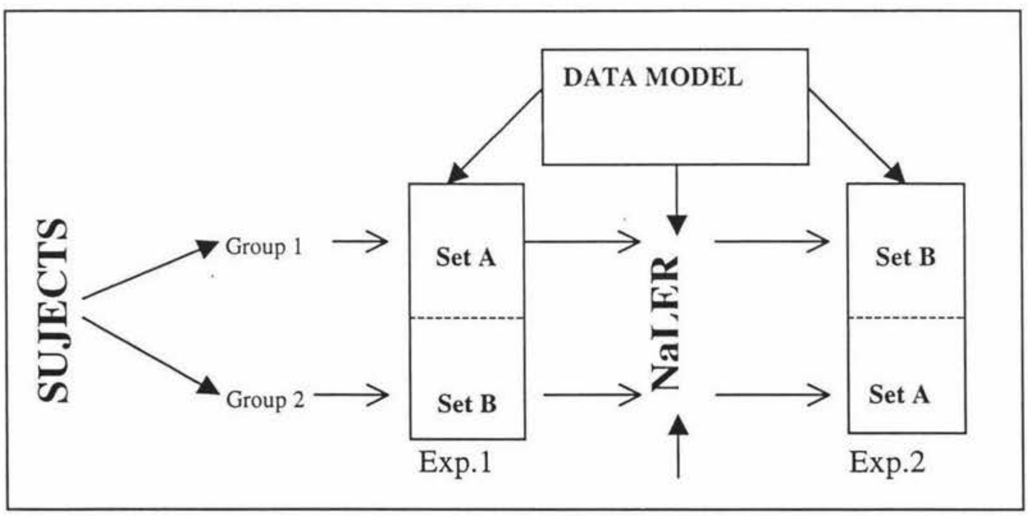


Figure 12 The procedure of two experiments

[4] Transcripts of these NaLER sentences can be found at Appendix 10

Both experiments were restricted to 30 minutes. This was based on the time required by subjects in the pilot study. However, this research did not measure the time to complete the task, but the limitation of the time frame was the same for both experiments.

4.2.5 Measurement

Measurement was based on two types of question; the 'true' and 'false' questions and the confidence questions. The purpose of the 'true' and 'false' questions was to find out not only the overall score of the correct answers but was also concerned with a difference in the specific categories of the understanding of the data model. The subjects were expected to review the data model and to mark each statement 'true' and 'false'. The level of understanding was measured by counting the number of the correct answers, which would be given 1 score for each correct answer. This was done per statement, per category and for all statements together.

For the confidence questions, the subjects were asked to indicate their level of confidence by using the five point likert scale; 1 = strongly agree, 2 = agree, 3 = neutral, 4 = disagree and 5 = strongly disagree. Confidence was measured in two ways; 'confidence correct' and 'confidence complete', both of which were measured as a mean score for each group and for each experiment.

4.2.6 Limitation

There are a number of the limitations in this research. The major limitation stems from the difficulties of creating a 'real world' environment in an experimental situation. As Galliers (1992) says the empirical studies of the experimental research type have inherent limitations, for example the isolation of such situations from most of the variables that are found in the real world. For example, due to the fact that students are not actually sophisticated users they have limited business knowledge. Moreover, there are two factors that make it difficult to equate student subjects with real-world users. They are the factors of motivation and experience. In terms of motivation, it is not possible to adequately simulate organisational consequences with student subjects. In addition, real world users are likely to have an experiential base of their organisation's problem or task domain that enable them to deal with more complex organisational situations. In a laboratory experiment with student subjects, it is difficult to approximate a similar experimental background (Jih *et al.*, 1989).

These students also represented a self-selected sample, because they had chosen both an Information System major, and then an advance Data Modelling course. The problem with a self-selected sample is that there may be many other variables, which differentiate between people who choose to do something and those who do not (Greene and D'Oliveira, 1982 p.149).

Another limitation was time, since the experiments were done by using the volunteer students as subjects, and both experiments took place in course tutorial time to avoid interrupting students' normal learning process. This research was unable to track individual student results because of the restriction of the ethics committee. It would be helpful to get more rigours use of criteria for determine the level of questions. In addition, the researcher herself had a limited time to complete her Master in Information Sciences. As a result there is no control group in this experiment.

Chapter five- Results and Discussion

5.1 Results

As described in the previous chapter, there were two groups of subjects and two sets of questionnaires. The questionnaire consists of two types of questions; the 'true' and 'false' statements and the confidence questions.

Both sets of questionnaires consisted of twenty 'true' and 'false' statements, which were categorised into three types. The three types of statements; simple, medium and difficult were designed to measure the degree of the understanding with which the subjects 'read' the data model. There were ten statements designed to test simple understanding, five statements to test medium understanding and five statements to test deep understanding. Each correct answer was given a score of 1.

The confidence of subjects' understanding of the data model was indicated by using a likert scale. The confidence was measured in two ways. The first measure, referred to as 'confidence correct', asked the subjects how **accurately** they felt they had understood the model. The second, referred to as 'confidence complete', asked the subjects to assess how **completely** they felt that they had understood the model. In both questions, confidence was assessed on the range 1-5 where; 1 = strongly agree, 2 = agree, 3 = neutral, 4 = disagree and 5 = strongly disagree.

The first section of this chapter shows the general results from both experiments and the overall results together. The next section shows the statistical analysis that was done using the statistics package Minitab for Windows.

5.1.1 Experiment 1

In the first experiment, the total numbers of subjects were thirty-five. Each subject was asked to read the E-R model and answer the questionnaire using the standard technique. Twenty-three subjects were assigned into group one and answered question set A, while twelve subjects were assigned into group two and answered question set B.

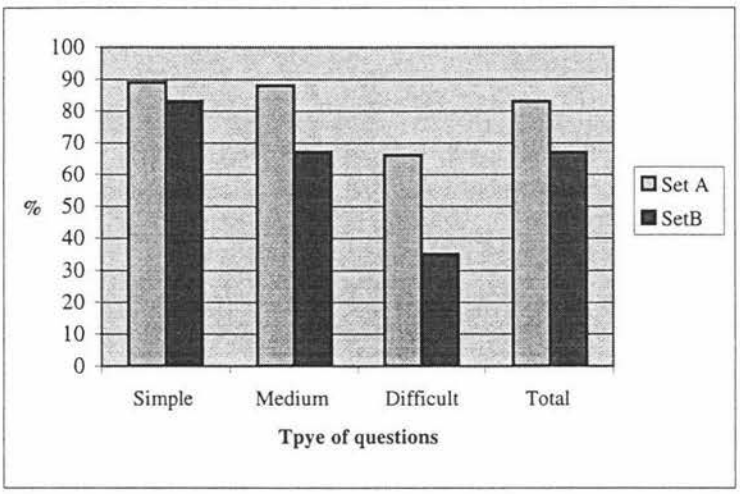


Figure 13 The correct scores (%) from experiment 1

Figure 13 shows the percentage of the correct answers from the first experiment. The result shows that subjects in group one, who did question set A, performed better than the subjects in group two, who did question set B, on every type of question; simple type 89% vs 83%, medium type 88% vs 67%, difficult type 66% vs 35% and total 83% vs 67%. The greatest difference is in the difficult questions. The more details of the first experiment are shown in Appendix 11.

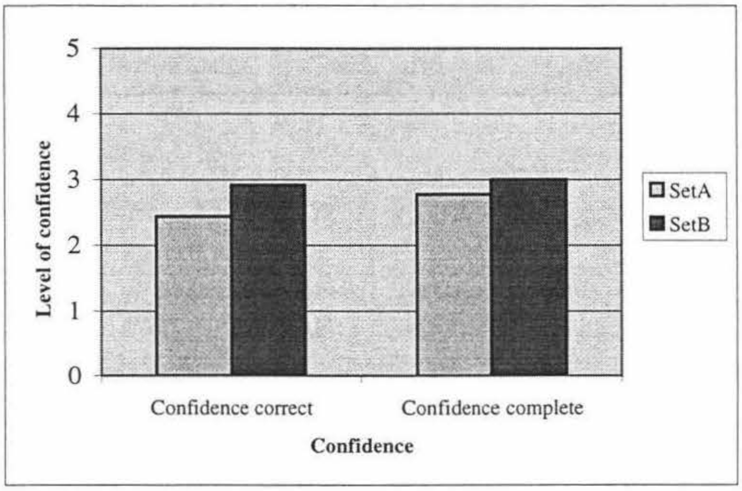


Figure 14 Confidence level using the standard technique

Figure 14 shows subjects' confidence on their understanding of the model when using the standard technique in experiment 1. The result shows that subjects in group one, who did question set A, felt more confident than the subjects in group two on both 'confidence correct' (mean values 2.44 vs 2.92) and 'confidence complete' (mean values 2.78 vs 3).

5.1.2 Experiment 2

In the second experiment, there were thirty-seven subjects. Each subject was asked to read the same E-R model as the first experiment by using the NaLER technique and answered the alternative set of questions. In the second experiment, seventeen subjects in group one answered question set B while twenty subjects in group two answered question set A.

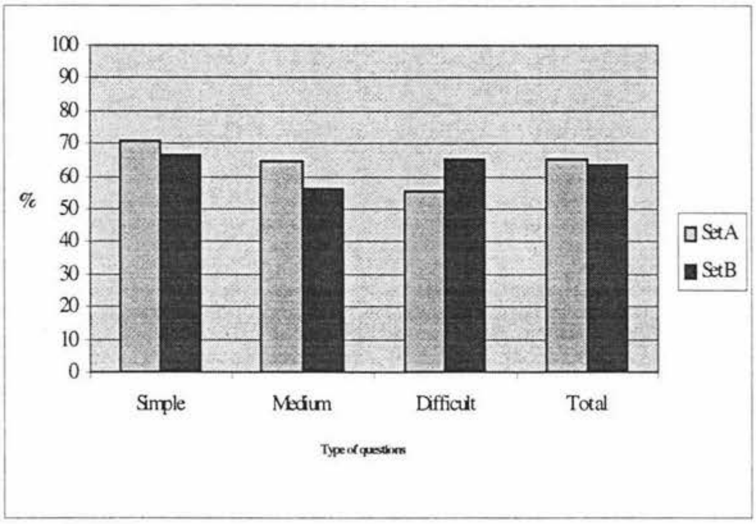


Figure 15 The correct scores (%) from experiment 2

Figure 15 shows the percentage of correct answers from the second experiment. The result shows some changes from experiment 1. Subjects in group two, who did question set A, performed better than the subjects in group one, who did question set B, on only some types of questions; simple type 71% vs 67%, medium type 65% vs 56% and total 65% vs 64%. However, for the difficult questions, subject group two who did question set A performed better with scores 65% vs 55%. The more details of the second experiment are shown in Appendix 12.

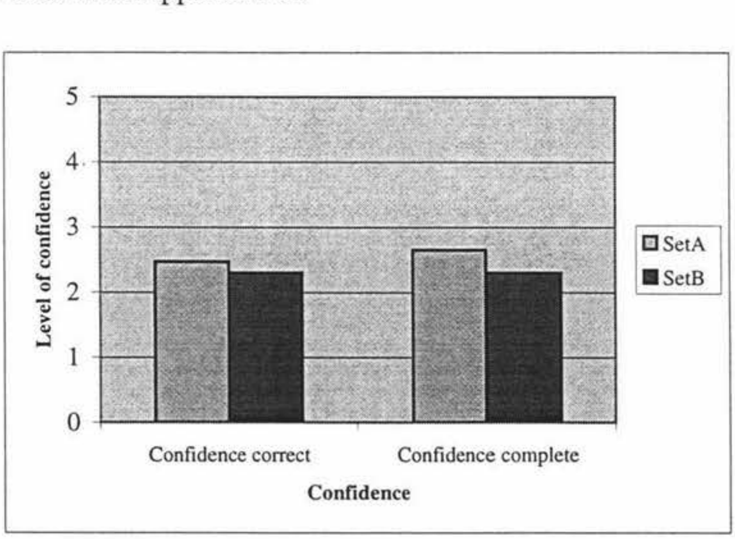


Figure 16 Confidence level using the NaLER technique

Figure 16 shows the subjects' confidence on their understanding of the model when using the NaLER technique in experiment 2. The result shows that subjects in group one, who did question set B, felt more confidence than the subjects in group two on both 'confidence correct' (mean values 2.30 vs 2.47) and 'confidence complete' (mean value 2.30 vs 2.65).

5.1.3 Overall

The percentage of correct answers from both experiments with all types of questions can be seen in Figure 17.

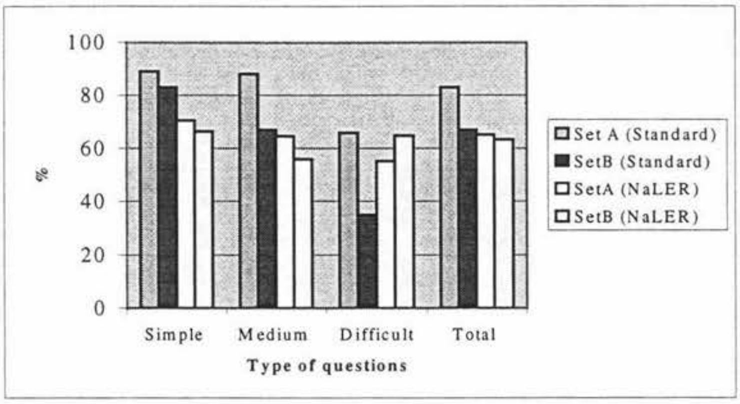


Figure 17 The correct scores (%) from experiment 1 & 2

The 'confidence correct' and 'confidence complete' when using difference techniques are shown in Figure 18.

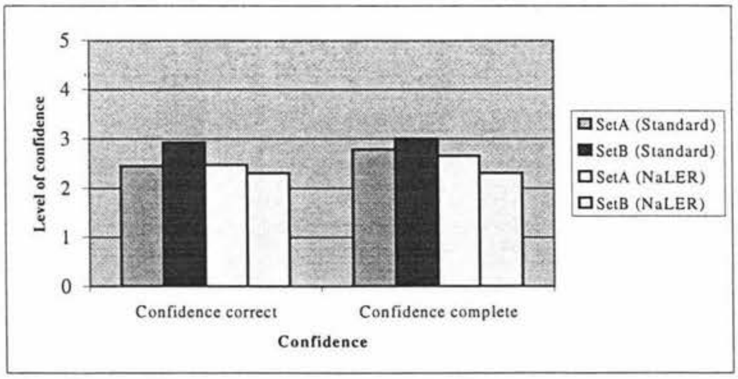


Figure 18 Confidence level of two techniques and two sets of questions

The following section shows the results of the number of correct answers (true and false) for the total scores for all question categories and within each of the difference categories (simple, medium and difficult) of the questions. The results shown are the mean scores and the significance of mean is at a level of confidence p = 0.05. The results are shown by technique, question and question* technique by using the GLM (General Linear Model) method from the statistics package Minitab for Windows (see Appendix 13).

Total scores for all question categories

Standard	NaLER	P_value
14.95	12.88	0.00

Table 3 Technique

The comparison of the total scores using the standard technique and after teaching the NaLER technique is shown in Table 3. The analysis shows that the total scores after teaching NaLER technique was significantly lower than when using the standard technique.

Set A	Set B	P_value
14.81	13.02	0.00

Table 4 Question

The comparison of the total scores using question set A and question set B is shown in Table 4. The analysis shows that the total score from question set A was significantly higher than from question set B.

Set A Standard	Set A NaLER	Set B Standard	Set B NaLER	P value
16.57	13.06	13.33	12.70	0.00

Table 5 Question*Technique

The comparison of the total scores of the interaction between question and technique is shown in Table 5. The analysis shows that there is a significant difference in the total scores. The analysis shows that the total scores from question set A and set B were higher when the standard technique was used.

<u>**Scores for simple questions**</u>

Standard	NaLER	P_value
8.56	6.85	0.00

Table 6 Technique

The comparison of the simple scores using the standard technique and after teaching the NaLER technique is shown in Table 6. The analysis shows that the simple score after teaching the NaLER technique was significantly lower than when using the standard technique.

Set A	Set B	P_value
7.96	7.45	0.00

Table 7 Question

The comparison of the simple scores using question set A and question set B is shown in Table 7. The analysis shows that the simple scores from question set A was significantly higher than from question set B.

Set A Standard	Set A NaLER	Set B Standard	Set B NaLER	P value
8.87	7.06	8.25	6.65	0.68

Table 8 Question*Technique

The comparison of the simple scores of the interaction between question and technique is shown in Table 8. The analysis shows that there was no significant difference in the simple scores. However, the result of the simple score from both questions set A and set B were higher when the standard technique was used.

Score for medium questions

Standard	NaLER	P_value
3.86	3.02	0.00

Table 9 Technique

The comparison of the medium scores using the standard technique and after teaching the NaLER technique is shown in Table 9. The analysis shows that the medium scores after teaching the NaLER technique was significantly lower than when using the standard technique.

Set A	Set B	P_value
3.81	3.07	0.00

Table 10 Question

The comparison of the medium scores using question set A and question set B is shown in Table 10. The analysis shows that the medium scores from question set A was significantly higher than from question set B.

Set A Standard	Set A NaLER	Set B Standard	Set B NaLER	P value
4.39	3.24	3.33	2.80	0.12

Table 11 Question*Technique

The comparison of the medium scores of the interaction between question and technique is shown in Table 11. The analysis shows that there was no significant difference in the medium scores. However, the result of the medium scores from both question set A and set B were higher when the standard technique was used.

Score for difficult questions

Standard	NaLER	P_value
2.53	3.01	0.17

Table 12 Technique

The comparison of difficult scores using the standard technique and after teaching the NaLER technique is shown in Table 12. The analysis shows that after teaching the NaLER technique, the difficult scores were higher than when using the standard technique. However, the difference was not significant.

Set A	Set B	P_value
3.04	2.50	0.12

Table 13 Question

The comparison of difficult scores using question set A and question set B is shown in Table 13. The analysis shows that difficult scores from question set A was higher than from question set B, but no significant difference was found.

Set A Standard	Set A NaLER	Set B Standard	Set B NaLER	P value
3.30	2.77	1.75	3.25	0.00

Table 14 Question*Technique

The comparison of difficult scores of the interaction between question and technique is shown in Table 14. The analysis shows that there was a significant difference in the difficult scores. The result from question set A when using the standard technique was higher than using the NaLER technique. However, the result from question set B when using the NaLER technique was higher than using the standard technique.

Confidence questions

The confidence was measured in two ways; confidence that their understanding was correct and the confidence that their understanding was complete. In the following discussion these two measures are referred to as 'confidence correct' and 'confidence complete' respectively. The mean values were then calculated and are used in the following analysis.

Confidence correct

	P value
Question	0.588
Technique	0.107
Question*Technique	0.037

Table 15 Confidence correct

As Table 15 shows, no significant difference of the 'confidence correct' was found when subjects answered different sets of questions. There was also no significant difference of the 'confidence correct' with the difference in techniques. However, when considering the interaction between question and technique, a significant difference was found in the 'confidence correct'.

Confident complete

	P value
Question	0.350
Technique	0.033
Question*Technique	0.115

Table 16 Confidence complete

As Table 16 shows, there was no significant difference of the 'confidence complete' when subjects answered different set of questions. There was also no significant difference of 'confidence complete' when considering the interaction between question and technique. However, a significant difference was found in the 'confidence complete' when the difference techniques were used.

The Comparison of confidence between question vs technique

Question	Technique	Mean
Set A	Standard	2.44
Set A	NaLER	2.47
Set B	Standard	2.92
Set B	NaLER	2.30

Table 17 Confidence correct

Table 17 shows that the 'confidence correct' for question set A when using the standard technique was higher than when using the NaLER technique. In contrast, the 'confidence correct' for question set B was higher when using the NaLER technique than when using the standard technique.

Question	Technique	Mean
Set A	Standard	2.78
Set A	NaLER	2.65
Set B	Standard	3.00
Set B	NaLER	2.30

Table 18 Confidence complete

As Table 18 shows, the 'confidence complete' of subjects using the NaLER technique was higher than when using the standard technique for both question set A and set B.

5.2 Discussion

The result of the comparison of the total scores of the two techniques shows that after subjects used the NaLER technique, there was a significant difference although not the one that had been expected. The total scores for the NaLER technique were lower than for the standard technique (Table 3). This may be because after learning the NaLER technique, subjects might have felt more aware of the complexity of the task and this could cause them to be less confident about their initial knowledge. Therefore, they may have tended to doubt their ideas and think more carefully than when they answered questions by using the standard technique. This could result in them doubting their initial understanding and consequently changing their previous (correct) answers.

The result of the questions from Table 4 shows that the total scores from question set A is significantly higher than from question set B. This may be because question set A was easier than question set B. In addition, the result shows that there was a significant difference between question set A and set B. However, the subjects had to answer both sets of questions in experiment 1 and 2, so the difference in difficulty of both sets of questions has no effect on the final results.

Similar results were found from the simple and the medium scores, which are shown in Table 6, Table 7, Table 9 and Table 10, respectively. However, the comparison of the simple and the medium scores of the interaction between question and technique was different from that found in the total scores. The analysis shows that there was no significant difference in the simple and the medium scores (Table 8 and Table 11).

On the other hand, comparison of the difficult scores of the interaction between question and technique was significantly different (Table 14). The number of the correct answers on the difficult questions when subjects using the NaLER technique was higher than using the standard technique, however the difference was not significant (Table 12). This may be because after learning the NaLER technique, subjects thought more than when using the standard technique and this improved their ability to answer more difficult questions although it seemed to hinder their ability to correctly answer the simpler ones.

The ‘confidence’ results when using difference techniques is also interesting. Subjects tended to give lower ‘confidence correct’ assessments when using the NaLER technique than when using the standard technique in question set A. However, there was higher ‘confidence correct’ assessments when using the NaLER technique than when using the standard technique in question set B (Table 17). This may support the previous suggestion that the subjects found the questions in set A easier than these in set B, as the NaLER technique which encourages a more precise and systematic approach, it may have been perceived as effective in dealing with more complex questions.

In addition, considering the mean values of the ‘confidence complete’, subjects had a higher confidence that their understanding of the model was complete after learning the NaLER technique (Table 18). Thus it appears that, after learning the NaLER technique, subjects’ confidence in their ability to understand a model was increased, even though they actually scored less well overall.

Another interesting point is that subjects tended to be able to get more correct answers on the difficult questions than the simple and medium questions when using the NaLER technique. This may be because the NaLER technique provides clear guidelines on how to ‘read’ a model which would encourage them to a deeper understanding. So, the results seem to suggest that the NaLER technique is useful for deep understanding but, in small amounts, adversely affects simple understanding or is even unsuitable. In addition, it seems reasonable to suggest that while a picture (or a diagram) can provide an understanding of the surface level of a model’s information, it is clear, unambiguous text which provides the detailed understanding required to answer ‘difficult’ questions.

The implications of these results are that the NaLER technique which uses natural language text can improve subjects’ ability to focus on what the model is actually recording in terms of information content. The need to investigate the semantics of the model in detail, appears to encourage them to gain a deeper understanding. However, there was no improvement on the subjects’ ability on the simple and medium understanding. This may be because subjects already visualised the meaning of the model from what they have seen. Looking in detail makes them question their previously correct answers. If we had given the subjects another few weeks to become more familiar with the NaLER technique, perhaps the results would be have been different.

5.3 Evaluation of results and hypotheses

Considering the hypotheses of this research,

Hypothesis 1: *E-R models read using the NaLER technique will yield a more accurate understanding than E-R model read using the standard technique.*

The result of the experiment was inconsistent with hypothesis 1, the results showed that subjects had more accurate understanding (measured by the number of correct answers given to a set of simple questions) when reading E-R models using the standard technique than reading E-R model when using the NaLER technique. Therefore, hypothesis 1 was not supported.

Hypothesis 2: *E-R models read using the NaLER technique will yield a more comprehensive understanding than E-R model read using the standard technique.*

The result was also inconsistent with hypothesis 2, when looking at a comprehensive understanding (measured by the number of correct answers given to sets of medium and difficult questions). The results were not in the same trend for medium and difficult questions, subjects had a higher understanding in the difficult questions but a lower understanding in the medium questions when reading E-R models using the NaLER technique. Therefore, hypothesis 2 was not supported.

Hypothesis 3: *Users using the NaLER technique will have a higher level of confidence in their understanding.*

The result of the experiment was consistent with the hypothesis. Subjects had more confidence when reading E-R model using the NaLER technique than when using the standard technique. As a result, hypothesis 3 was supported.

Chapter six – Conclusions and Further research

6.1 Conclusions

This research has focused on data models as tools of communication and concentrates on a tool to better 'read' data models. It was also interested in discovering how well people understand data models and whether their understanding can be improved.

In order to investigate the subjects' understanding of a data model, two experiments were conducted to test their understanding before and after learning the NaLER technique.

The results show that after learning the NaLER technique, subjects were more precise in what the model is actually recording in terms of information content. The results seem to indicate that the NaLER technique provides clear guidelines on how to 'read' a model which would encourage subjects to a deeper understanding. As March (1993) observes that people are very effective at intuitively accommodating ambiguity or confusion in a model by "...projecting meaning onto the data structures, rather than abstracting meaning out of them"(p.45). Using the NaLER technique thus forces them to abstract the meaning from the models rather than projecting meaning on to them.

The interesting point is that subjects tended to get more correct answers on the difficult questions than the simple and the medium questions. They also had higher confidence in their ability to understand the model when using the NaLER technique.

The results show that using the NaLER technique may improve subjects' deep understanding of a data model. However there appeared to be deterioration in their levels of simple and medium understanding. This anomalous result may have several explanations. It could be that subjects were relying on an imperfect memory of the previous experiment. As Kent (1978) notes that "...if a mind is committed to a certain model, then it will perform amazing feats of distortion to see things structured that way" (p.93). On the other hand, it could be due to a 'recency' effect, that is that the experiment was conducted too soon after the introduction of the new technique and before the subjects were fully familiar with it.

From the results, it can be noted that 'reading' E-R models using the standard technique provided the basic knowledge. Subjects used this to gain a surface understanding about the model. This finding is supported by a number of early studies which state that E-R models provide a clear and intuitive way of communicating with non-technical users (Konsynski, 1979; Brodie, *et al.*, 1984; Teorey *et al.*, 1986; Batra *et al.*, 1990). While not providing conclusive proof, the result of this research hint that subjects using NaLER may find it easier to develop a deeper understanding of a model's semantics.

6.2 Further research

This adverse effect might be solved by giving subjects more time to be familiar with the NaLER technique and to absorb their deep understanding. If a small amount of NaLER training can lead to deeper understanding, this may well support the desirability to conduct future experiments but with more NaLER training. Further research is required but with more time between learning NaLER and doing the second experiment. Perhaps, it might be valuable to use a 'think aloud' process tracing methodology (Srinivasan and Te'eni, 1995) to follow each individual's attempt to read the model.

When subjects have thoroughly learnt the technique, they might have a clear picture about the model in their mind and this would help them being able to correctly answer any type of questions. So one possibility would be to re-run the experiment with a control group who would not be given training in NaLER. However, this would create some potential ethical problems to those who do not receive any NaLER training. It would be also worth conducting future experiments with a larger number of subjects.

It might be useful to conduct the future experiment but allow tracing of individual subject results. Another interesting issue would be to conduct future research with the real users instead of students.

As was noted at the beginning of this research, data models are important in information system development, particularly as tools for expressing and communicating business requirement. Thus, the ability to understand the information content of data models is fundamentally important. This research has investigated subjects' understanding of a data model before and after learning the NaLER technique. The results of this research suggest that using the NaLER technique may improve subjects' deep understanding of a data model. This technique may also provide a pragmatic and easily learned technique for 'reading' a model for those who have no requirement in learning to construct one.

In conclusion, this research has highlighted the need for continued research in the communicative aspects of data modelling.

References

1. Amer, T.S. (1993). Entity-Relationship and Relational Database Modeling Representations for the Audit Review of Accounting Applications: An Experimental Examination of Effectiveness. *Journal of Information Systems,* 7(1), 1-15.
2. Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review,* 89, 369-406.
3. ANSI/X3/SPARC (1975). Study Group on Data Base Management Systems: Interim Report 75 02-08 *ACM SIGMOD* Newsletter, 7(2).
4. Atkins, C. F. (1996). Prescription or Description: Some Observations on the Conceptual Modelling Process, In PURVIS, M. (ed.), *Proceedings of Software Engineering: Education and Practical Conference*, University of Otago, Dunedin New Zealand, January, 34-41.
5. Atkins, C. F. & Patrick, J. D. (1998). NaLER: A Natural Language Method for Interpreting E-R Models. In PURVIS, M. (ed.), *Proceedings of Software Engineering: Education and Practical Conference*, University of Otago, Dunedin New Zealand, January, 2-9.
6. Atkins, C. F. (2000). INTECoM: An integrated conceptual modelling framework. *Unpublished PhD Thesis*, Massey University, New Zealand.
7. Avison, D. E. & Fitzgerald G. (1995). *Information Systems Development: methodologies, techniques and tools. 2nd edition.* Mcgraw-Hill International (UK) Ltd.
8. Avison, D. E. (1992). *Information Systems Development, A Database Approach. 2nd edition*. Blackwell Scientific Publications, Oxford.
9. Barker, R. (1990). *CASE*Method: Entity Relationship Modelling.* Addison Wesley.
10. Batini, C., Ceri, S. & Navathe, S. (1992). *Conceptual Database Design, an Entity Relationship Approach.* Benjamin Cummings, California.

11. Batra, D & Davis, J.G. (1992). Conceptual Data Modelling in Database Design: similarities and differences between novice and expert designers, *International Journal of Man-Machine Studies,* 37, 82-101.

12. Batra, D. & Anthony S.R. (1994). Novice errors in conceptual database design. *European Journal of Information systems,* 3(1), 57-69.

13. Batra, D. & Marakas G. M. (1995). Conceptual Data Modelling in Theory and Practice, *European Journal of Information Systems*, 4, 185-193.

14. Batra, D. & Sein, Muang K. (1994). Improving conceptual database design through feedback. *International Journal of Human-Computer Studies*, 40, 653-676.

15. Batra, D. & Srinivasan, A. (1992). A review and analysis of the usaility of data management environments. *International Journal of Man-Machine Studies,* 36, 395-417.

16. Batra, D. & Zanakis, S.H. (1994). A conceptual database design approach based on rules and heuristics. *European Journal of Information Systems*, 3(3), 228-239.

17. Batra, D., Hoffer, J.A. & Bostrom, R.P. (1990). Comparing Representations with Relational and EER Models. *Communications of the ACM,* (Feb), 33(2), 126-139.

18. Benyon, D. (1997). *Information and Data Modelling*. Second Edition. McGraw-Hill, London.

19. Benyon-Davies, P. (1996). *Database Systems*. Macmillan Press, London.

20. Biller, H., & Neuhold, E. (1978). Semantics of Data Bases: the Semantics of Data Models. *Information Systems,* 3, 11-30.

21. Bittner, J. R. (1988). *Fundamentals of communication 2nd Edition* Prentice Hall.

22. Blaha, M.R., Premerlani, W.J. & Rumbaugh, J.E. (1988). Relational Database Design using an Object-Oriented Methodology. *Communications of the ACM,* 31(4), 414-427.

23. Blaxter, L., Hughes, C. & Tight, M. (1996). *How to research.* Open University Press.

24. Bock, Douglas, B. & Ryan, T (1993). Accuracy in Modeling with Extended Entity Relationship and Object Oriented Data Models. *Journal of Database Management,* 4(4), 30-39.

25. Borgida, A., Greenspan, S. & Mylopoulos, J. (1985). Knowledge representation as the basis for requirements specification. *Computer*, 18, 82-91.

26. Brodie, M. L., Mylopoulos, J. & Schmidt, J. W. (Eds.). (1984). *On Conceptual Modeling*. Springer-Verlag, New York.

27. Burch, John G. (1992). *Systems Analysis, Design and Implementation.* boyd & fraser publishing company, Boston.

28. Campbell, D. (1992). Entity-Relationship Modeling; One Style Suits All? *Database,* Summer, 12-18.

29. Chen, P.P. (1976). The Entity-Relationship Model- Toward a Unified View of Data. *ACM Transactions on Database Systems,* 1(1), 9-36.

30. Coad, P. & Yourdon, E. (1991). *Object Oriented Analysis.* Prentice Hall.

31. Connolly, T. M., Begg, C.E. & Strachan, A. D. (1996). *Database Systems*. Addison Wesley, Wokingham, England.

32. Davydov, M. M. (1994). From model to database. *Database Programming & design*, March, 46-52.

33. de Carteret, C. & Vidgen, R. (1995). *Data Modelling for Information Systems.* Pitman Publishing, London.

34. Elmasri, R., Weeldryer, J. & Hevner, A. (1985). The Category Concept: An extension to the Entity-relationship Model. *Data and Knowledge Engineering.* 1, 75-116.

35. Emmert, P., & Donaghy, W. C. (1981). *Human Communication: Elements and Contexts*. Random House, New York.

36. Ferguson, J. R. (1988). Data Modelling and Systems Analysis. *Paper Presented at 19th Colleges of Advanced Education Computing Conference, Australia* (cited in Firns, 1990).

37. Firns, Peter G. (1990). Determining a Useful Balance between Understandability and Rigour in data Modelling. *New Zealand Journal of Computing,* 2(1), 13-21.

38. Galliers, R. D. (1992). *Information Systems Research: Issues, Methods and Practical Guidelines.* Alfred Waller Ltd, Publishers.

39. Gibson, Michael L. & Hughes Cary T. (1994). *Systems Analysis and Design. A Comprehensive Methodology with CASE.* boyd & fraser publishing company, Massachusetts.

40. Goldstein, R. C., & Storey, V. C. (1990). Some findings of the intuitiveness of entity-relationship constructs. In *Entity-Relationship Approach to Database Design and Querying* (Lochovosky, F.H., Ed.) Elsevier Science Amsterdam, (quoted in Hitchman, 1995).

41. Graziano, A. M. & Raulin, M. L. (1993). *Research Methods: A Process of Inquiry. 2nd edition.* Harper Collins College Publishers.

42. Greene, J., & D'Oliveira, M. (1982). *Learning to use Statistical Tests in Psychology: A Student's Guide.* The Open University Press.

43. Griethuysen, J.J. (Ed.) (1983). *Concepts and Terminology for the Conceptual Schema. ISO/TC97/SC5-N695.* International Organization for Standardization, Geneva, Switzerland (quoted in Batra & Marakas, 1995).

44. Halpin, T.A. (1995). *Conceptual Schema and Relational database design 2nd edition*, Prentice Hall, Australia.

45. Hammer, M. & McLeod, D. (1981). Database Description with SDM: A Semantic Database Model. *ACM Transactions on Database Systems,* 693), 351-386.

46. Hawryszkiewycz, I. (1998). *Introduction to Systems Analysis and Design 4th Edition*, Prentice Hall.

47. Hirschheim, R., K., Heinz K. & Lyytinen, K. (1995). *Information Systems Development and data Modelling, Conceptual and Philosophical Foundations.* Cambridge University Press, Cambridge.

48. Hitchman, S. (1995). Practitioner perceptions on the use of some semantic concepts in the entity-relationship model. *European Journal of Information Systems,* 4, 31-40.

49. Hitchman, S. (1997). Using DEKAF to understand data modelling in the practitioner domain. *European Journal of Information Systems*, 6, 181-189.

50. Jarvenpaa, Sirkka L & Machesky, Jefry J. (1989). Data Analysis and Learning: an experimental study of data modelling tools. *International Journal of Man-Machine Studies,* 31, 367-391.

51. Jeffries, R., Turner, A. T., Polson, P. G. & Atwood, M. E. (1980). *The Processes Involved in designing Software.* Englewood, CO: Science Applications.

52. Jih, W. K., Bradbard, D. A., Snyder, C. A. & Thompson, N. G. A. (1989). The effects of relational and entity-relationship data models on query performance of end users. *International Journal of Man-Machine Studies,* 31, 257-267.

53. Juhn, S. & Naumann, J.D. (1985). The effectiveness of data representation characteristics on user validation. *Proceedings of the Sixth International Conference on Information systems* (Indianapolis 1985).

54. Kendall, K. E. & Kendall, J. E. (1999). *Systems analysis and design 4th Edition* Prentice Hall International, Inc.

55. Kent, W. (1978*). Data and Reality.* North-Holland, Amsterdam.

56. Kim, Young-Gul & March Salvatore T. (1995). Comparing Data Modeling Formalisms. *Communications of the ACM,* 38(6), 103-115.

57. Klein, H. & Hirschheim, R. A. (1987). A Comparative Framework of Data Modelling Paradigms and Approaches. *The Computer Journal,* 30(1), 8-15.

58. Konsynski, B. R. (1979). *Data Base Driven Systems.* University of Arizona. (quoted in Hitchman, 1995).

59. Kroenke, D. M. (1992). *Database Processing: Fundamentals, Design, Implementation. 4th Edition.* Maxwell Macmillan International.

60. Lindland, O.I., Sindre, G and Sølvberg, A. (1994). Understanding Quality in Conceptual Modelling, *IEEE Software,* March: 42-49.

61. MacEachren, Alan E. (1995). *How Maps Work.* The Guildford Press, New York.

62. Marche, S. (1993). Measuring the stability of data models. *European Journal of Information systems,* 2(1), 37-47.

63. Martin, J. (1987). *Recommended Diagramming standards for Analysts and Programmers.* Prentice-Hall, New Jersey.

64. McFadden, F. R. & Hoffer, J. A. (1994). *Database Management.* The Benjamin/Cummings Publishing Company Inc, California.

65. Moody, D. & Osianlis, A. (1996). Bringing Data Models to "Life": An Interactive Tool for Representing Entity Relationship Models. *Proceedings of the 7th Australian Conference on Information Systems (ACIS'96),* University of Tasmania Australia, 497-508.

66. Moore, N. (1983). *How to do research.* Library Association Publishing Limited.

67. Nijssen, G. M. & Halpin, T. A. (1989) *Conceptual Schema and Relational Database Design*, Prentice Hall.

68. Nordbotten, J. C. & Crosby, M. E. (1996). Reading strategies for graphic models from an experiment in data model perception, *Proceedings of 5th International Conference on user Modeling,* Hawaii, USA, January, 43-48.

69. Olle, T. William. (1993). Data Modelling and Conceptual Modelling: A Comparative analysis of functionality and roles. *Australian Journal of Information Systems*, 1(1), 46-57.

70. Pinner, D. & D. (1994). *Communication Skills 3rd Edition.* Longman Paul.

71. Ram, S. (1995). Deriving Functional Dependencies from the Entity-Relationship Model. *Communications of the ACM*, 58(9), 95-107.

72. Raymond, Darrell R., Canas, Alberto J., Tompa, Frank W. & Safayeni Frank R.(1989). Measuring the effectiveness of personal database structures. *International Journal of Man-Machine Studies*, 31, 237-256.

73. Ricardo, C. (1990). *Database Systems: Principles, Design & implementation.* Macmillan, New York.

74. Ruben, B. D., Josephine, R. H. & Jarice, K. H. (1982). *Communication Systems, Technology, and Culture. Communications and the Future: Prospects, Promises and problems.* edited by Howard, F. & Didsbury, Jr. World Future Society.

75. Sanders G. Lawrence (1995). *Data Modeling.* boyd & fraser, ITP Inc, Danvers, Massachusetts.

76. Schenck, Douglas A. & Wilson Peter R. (1994). *Information Modeling: The EXPRESS Way.* Oxford University Press, New York.

77. Shanks, G. (1997). Conceptual Data Modelling: An empirical study of expert and novice data modellers. *Australian Journal Information Systems*, 4(2), 63-73.

78. Shoval, P. (1985). Essential Information Structure Diagrams and database schema design, *Information systems*, 10(4), 417-423.

79. Shoval, P. & Even-Chaime, M. (1987). Database schema design: An experimental comparison between normalisation and information analysis. *Database,* 18(3), Spring, 30-39.

80. Shoval, P. & Frumermann, I. (1994). OO and EER Conceptual Schemas: A Comparison of User Comprehension. *Journal of Database Management,* 5(4), 28-38.

81. Shoval, P. (1997). Experimental Comparisons of Entity-Relationship and Object-Oriented Data Models. *Australian Journal of Information Systems,* 4(2), 74-81.

82. Siau, K., Wand, Y. & Benbasat I. (1995). A Psychological Study on the Use of Relationship Concept-Some Preliminary Findings. In Ilvari, J., Lyytinen, K., Rossi, M.(eds.),. *Lecture Notes in Computer Science,* 932, Springer, 341-354.

83. Simsion, G. (1994). *Data Modelling Essentials Analysis, Design and Innovation.* Van Nostrand Reinhold, Boston.

84. Smith, J. M. & Smith, D. C. P. (1997). Database Abstractions: Aggregation and Generalisation. *ACM Transactions on Database Systems,* 2(2), 105-133.

85. Smithson, S. (1984). *Business communication today: A guide to effective communication techniques.* ICSA Publishing. Cambridge.

86. Sproull, N. (1988). *Handbook of research Methods: a guide for practitioners and students in the social sciences.* The Scarecrow Press, Inc.

87. Srinivasan, A. & Te'eni, D. (1995). Modeling as Constrained Problem Solving: An Empirical Study of the Data Modeling Process. *Management Science,* 41(3), 419-434.

88. Teorey, Toby J., Yang, Dongqing & Fry, James P. (1986). A Logical Design Methodology for Relational Databases using the Extended Entity-Relationship Model. *Computing surveys,* 18(2), 197-222.

89. Tsichritzis, D.C. & Lochovsky, F.H. (1982). *Data Models.* Prentice Hall, Englewood Cliffs, New Jersey.

90. Urquhart, C., & Young, J. (1994). The role of communication is software project success. *Proceedings of the 4th Australian conference on information systems*, 521-534.

91. Veryard, R. (1984). *Pragmatic Data Analysis*, Blackwell Scientific Publication Oxford.

92. Veryard, R. (1992). *Information Modelling: Practical Guidance.* Prentice Hall, Englewood Cliffs, New Jersey.

93. Veryard, R. (1994). *Information co-ordination: The management of Information Models, Systems and Organizations.* Prentice Hall Englewood Cliffs, New Jersey.

94. Whitten, J., Bentley, L., and Barlow, V. (1998). *Systems analysis and design methods 4th Ed.* Richard D. Irwin, Inc., U.S.A.

Glossary

Analysis stage	In IS development, the stage where the user requirement is identified and specified, usually in a formal and precise record.
CASE tool	Computer Aided Systems (or Software) Engineering. A software package that provides computer support for key activities in the system development process.
Communication	An action or a process of communicating.
Conceptual data model	A model, or collection of models, that records the information requirements of a system with no consideration of the specific technology by which it will be implemented. Sometimes referred to as the 'logical model' of the system.
Data model	A representation of a particular set of data created according to the syntax of a specific data model. A data model will usually minimally include a graphical representation of the data and textual details. It may also be called an 'application model' or a 'user model'.
Data modelling approach	Usually used to describe a generic type rather than a specific data model, e.g. The E-R approach describes the general characteristics of the data model rather than those of a particular version.
Entity	In general terms, something that has a real or distinct existence. In data modelling terms, an abstracted view of a collection of properties (attributes) that described something of interest to the system under consideration.
E-R model	Entity Relationship model. A data model that uses three primitive elements, entities, attributes and relationships to represents data of interest to a system.
E-R/Relational hybrid model	An application data model that has a relational structure and is graphically represented with a form of e-r notation. The 'entities' in such a model are, in fact, relations and the 'relationships' are primary-foreign key links.
EER	Extended Entity Relationship model. The E-R model which has been extended to include the notion of categories. It's a logical design tool that can be used to conceptualise data requirements. The EER model can then be converted to a relational representation (or any other data model) for database implementation.
Implementation model	A type of data model which can be implemented by commercially available DBMSs, e.g. Relational Model.

Logical data design	The stage of database design that is primarily concerned with the design of the data structures. In some approaches, it includes the analysis of the data. It is specifically separated from physical data design.
Logical model	The end result of logical data design.
NaLER	Natural Language for E-R. A new technique developed to use as an aid to the modeller within the design process and as a means of presenting the information content of the design model to users.
NIAM	Natural Language Information Analysis Methodology. A comprehensive and prescriptive method for the construction of relational databases, which uses a formal subset of natural language to identify the information requirements. Before 1994 it is known as the Nijssen Information Analysis Methodology.
ORM	Object-role model(ling). A data model which uses the primitive elements of objects and roles to represent data of interest to a system.
Physical model	A representation of a data model that has been tailored for implementation in a particular DBMS.
Problem domain	The area of an organisation that is under consideration as part of a database development activity.
Semantics	The information content of the data model.
Universe of discourse (UOD)	The preferred term in this study for the 'problem domain'.
User requirement	This is the users' perception of the system that the users want and includes the set of functional and information requirements that the user will demand of a system.
Way of communicating	This describes the form in which the models are to be communicated to human beings, usually in some form of graphical notation. Although models have the same way of modelling they can use different notation.

Appendices

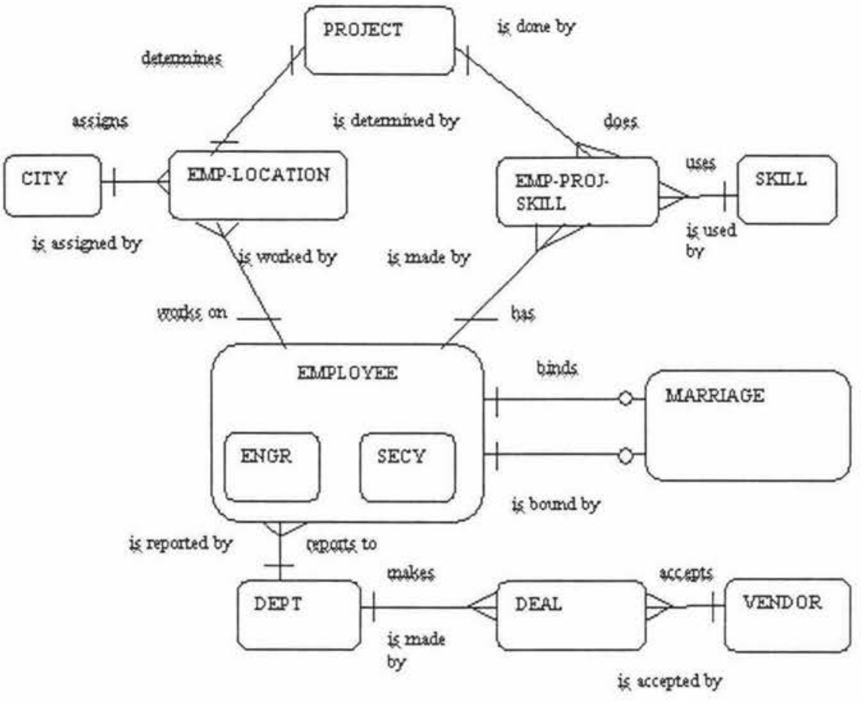
Appendix 1 Case Description (Batra et al., 1990)

Project Inc. is an engineering firm with approximately 500 employees. A database is required to keep track of all employees, their skills, projects assigned and departments worked in. Every employee has a unique number assigned by the firm. It is requires to store his/her name and date-of-birth. If an employee is currently married to another employee of Projects Inc., then it is required to store the date of marriage and who is married to whom. However, no record of marriage need be maintained if the spouse of an employee is not an employee of the firm. Each employee is given a job title (e.g. engineer, secretary, foreman, etc.) We are interested in collecting more data which is specific to the following types: engineer and secretary. The relevant data to be recorded for engineers is the type of degree (e.g. electrical, mechanical, civil, etc.) and for secretaries is their typing speeds. An employee does only one type of job at any given time and we need to retain information material for only the current job for an employee.

There are 11 different departments, and each has a unique name. An employee can report to only one department. Each department has a phone number. To procure various kinds of equipment, each department deals with many vendors. A vendor typically supplies equipment to many departments. It is required to store the name and address of each vendor, and the date of last meeting between a department and a vendor.

Many employees can work on a project. An employee can work in many projects (e.g. Southwest Refinery, California Petrochemical, etc.), but can be assigned to only one project in a given city. For each city, we are interested in its state and population. An employee can have many skills (e.g. preparing material requisitions, checking drawing, etc.), but he/she may use only a given set of skills on a particular project. (For example, an employee MURPHY may prepare requisitions as well as check drawings for California Petrochemicals.) An employee uses each skill that he/she possesses in at least one project. Each skill is assigned a number. A short description is required to be stored for each skill. Projects are distinguished by project numbers. It is required to store the estimated cost of each project.

Appendix 2 ER/R Model



Entities & Attributes

EMPLOYEE (EMP#, ENAME, DATE- BIRTH, DNAME, JOB_TITLE)

Categories of EMPLOYEE by JOB_TITLE:
EMP_ENGR (DEGREE)
EMP_SECY (TYPE_SPEED)

DEPT (DNAME, PHONE)

VENDOR (VNAME, ADDRESS)

DEAL (DNAME, VNAME, LAST_MEET)

SKILL (SKILL #, DESCRIPTION)

PROJECT (PROJ #, EST_COST)

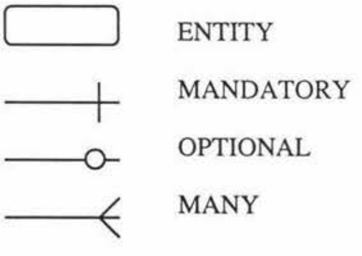
EMP_PROJ_SKILL (EMP #, SKILL #, PROJ #)

CITY (CITY NM, STATE, POPULATION)

EMP_LOCATION (EMP #, CITY_NM, PROJ #)

MARRIAGE (EMP #, DATE_MAR, SPOUSE -EMP#)

NOTATION



Appendix 3 Questionnaire (Common Set)

	Statement	Type	Set A#	Set B#
1.	One employee can use more than one skill on a project.(T)	medium	1	
2.	One employee can use more than one skill in one city. (T)	difficult	2	1
3.	One employee can have more than one skill. (T)	simple		2
4.	One employee can have more than one job title. (F)	simple	3	2
5.	We always know who an employee is married to. (F)	medium	4	
6.	One employee may report to more than one department. (F)	simple	5	3
7.	If an employee is married more than once, we can discover his/her previous spouse. (F)	medium		4
8.	We can check when one employee works on a particular project. (F)	medium	6	
9.	We can discover in which city an employee works on a particular project. (T)	medium	7	
10.	We always know the name of a vendor. (T)	simple	8	5
11.	One department can deal with many vendors. (T)	simple	9	6
12.	We know all the deals that have been made between one department and one vendor. (F)	difficult	10	
13.	We know the last deal made between each department and each vendor. (T)	medium	7	
14.	We always know which skills an employee has used on which project. (T)	medium	11	
15.	We know which project is in which city. (T)	difficult	12	
16.	We always know which employee works in which department. (T)	simple	13	8
17.	Each employee must have an employee #. (T)	simple	14	9
18.	We always know a department's name. (T)	simple	15	
19.	Many employees can work on one project. (T)	simple	16	

20.	One employee can work on many projects. (T)	simple		10
21.	One employee can work in many cities. (T)	medium		11
22.	One employee must use only one skill on one project. (F)	difficult	17	
23.	We always know which city an employee is working in. (T)	simple	18	12
24.	One project can be done in only one city. (T)	medium		13
25.	We always know a project name. (F)	medium		14
26.	Two departments can have the same name. (F)	simple		15
27.	Two departments can have the same phone-no. (T)	simple	19	
28.	One employee must have one skill. (F)	difficult	20	
29.	We know all the skills that an employee has. (F)	difficult		16
30.	We always know what degree an employee has. (F)	difficult		17
31.	An employee can work on two projects in the same city. (F)	difficult		18
32.	We need to know the location of a project to be able to estimate its cost. (F)	medium		
33.	We know how long an employee has been married. (F)	difficult		19
34.	An employee working in a particular city must be working on a project. (T)	difficult		20

Appendix 4 Questionnaire (Set A)

	Statement	True	False
1.	One employee can use more than one skill on a project.		
2.	One employee can use more than one skill in one city.		
3.	One employee can have more than one job title.		
4.	We always know who an employee is married to.		
5.	One employee may report to more than one department.		
6.	We can check how long one employee works on a particular project.		
7.	We can discover in which city an employee works on a particular project.		
8.	We always know the name of a vendor.		
9.	One department can deal with many vendors.		
10.	We know all the deals that have been made between one department and one vendor.		
11.	We always know which skills an employee has used on which project.		
12.	We know which project is in which city.		
13.	We always know which employee works in which department.		
14.	Each employee must have an employee #.		
15.	We always know a department's name.		
16.	Many employee can work on one project.		
17.	One employee must use only one skill on one project.		
18.	We always know which city an employee is working in.		
19.	Two departments can have the same phone-no.		
20.	One employee must have one skill.		

How confident are you that you have understood this model correctly? (Circle number)

1. strongly agree
2. agree
3. neutral
4. disagree
5. strongly disagree

How confident are you that you have understood this model completely? (Circle number)

1. strongly agree
2. agree
3. neutral
4. disagree
5. strongly disagree

Appendix 5 Questionnaire (Set B)

Statement	True	False
1. One employee can have more than one skill.	☐	☐
2. One employee can have more than one job title.	☐	☐
3. One employee may report to more than one department.	☐	☐
4. If an employee is married more than once, we can discover his/her previous spouse.	☐	☐
5. We always know the name of a vendor.	☐	☐
6. One department can deal with many vendors.	☐	☐
7. We know the last deal made between each department and each vendor.	☐	☐
8. We always know which employee works in which department.	☐	☐
9. Each employee must have an employee #.	☐	☐
10. One employee can work on many projects.	☐	☐
11. One employee can work in many cities.	☐	☐
12. We always know which city an employee is working in.	☐	☐
13. One project can be done in only one city.	☐	☐
14. We always know a project name.	☐	☐
15. Two departments can have the same name.	☐	☐
16. We know all the skills that an employee has.	☐	☐
17. We always know what degree an employee has.	☐	☐
18. An employee can work on two projects in the same city.	☐	☐
19. We always know how long an employee has been married.	☐	☐
20. An employee working in a particular city must be working on a project.	☐	☐

How confident are you that you have understood this model correctly? (Circle number)

1. strongly agree
2. agree
3. neutral
4. disagree
5. strongly disagree

How confident are you that you have understood this model completely? (Circle number)

1. strongly agree
2. agree
3. neutral
4. disagree
5. strongly disagree

Massey University Human Ethics Committee
c/- EXMSS House, Turitea

Telephone: 64 6 350 5799, extn 2360
Email: P.Dewe@massey.ac.nz
http://www.massey.ac.nz/~muhec

Massey University

Private Bag 11 222,
Palmerston North,
New Zealand
Telephone: 64 6 356 9099

Appendix 6 Ethics Committee Approval

6 September 1999

Ms Somrutai MALAIPONG
PG Student
Information Systems
TURITEA

Dear Somrutai

Re: Human Ethics Application – MUHEC 99/118
"Every Picture tells a Story: An investigation of Data Models as Tools of Communication"

Thank you for your letter of 26 August 1999.

The amendments you have made and the explanations you have given now meet the requirements of the Massey University Human Ethics Committee and the ethics of your application are approved.

Yours sincerely

pp Professor Philip J Dewe
Chairperson
Massey University Human Ethics Committee - Palmerston North

cc Ms Clare Atkins
Information Systems
TURITEA

MASSEY UNIVERSITY HUMAN ETHICS COMMITTEE

To: Ethics Secretary
Human Ethics Committee
AVC's Office (Research)
Turitea, Palmerston North Campus

APPLICATION FOR APPROVAL OF PROPOSED TEACHING/RESEARCH PROCEDURES INVOLVING HUMAN SUBJECTS

APPLICANT(S): **Name:** Ms.Somrutai Malaipong

Department: Information Systems

Contact Number: ext 7539

Status: Masterate student

Name of Employer: -

PROJECT: **Title:** "Every picture tells a story: An investigation of data models as tools of communication"

Status: Masterate

Funding Source: n/a

Clinical Trial Status: yes ❑ no ❑✔

ATTACHMENTS: Information Sheet, Consent Form

SUPERVISOR(S): **Name:** Ms. Clare Atkins

Department: Information Systems

SIGNATURE(S): **Applicant:**

Supervisor(s):

DATE:

OFFICE USE ONLY

Received: **Decision:**

1. DESCRIPTION

1.1 Justification

My research area is an investigation of data models as tools of communication. Data models are important in information system development, particularly as tools for expressing and communicating business requirements. As both business requirements and database technology, become more sophisticated, the need for understanding data models is increasing. The ability to understand the information content of data models is a fundamental skill by everyone involved with them. However, there has been little research on the use of data models as tools of communication.

The first step of database design is the development of the conceptual data model of the application. A conceptual data model is an abstraction of the real world (organisation) data pertinent to an enterprise. The process of deriving and analysing the data inherent in a business situation and of mapping objects of this understanding of reality in a conceptual model representation constitutes the discovery phase. The discovery phase consists of two parts. The first part involves the elicitation of the information requirements, which is traditionally a result of an interview between the user and the analyst. The second part, in the data modelling context, involves the representation of the information requirements into the form of a conceptual data model.

My research concentrates on the second part of this phase, that is, the translation of an understanding of the business situation into a representational form for a conceptual database. This schema is usually represented in a diagrammatic form, and serves as a communication tool between developers and users. Once approved by users, it is converted into a specific database schema, depending on the data model and the DBMS used for implementation. However, the diagram can be misunderstood or misread by the users and in order to explore this issue, I wish to conduct experiments. The purpose of this study is to compare the understandability of models by students (representing sophisticated users) before and after learning a technique to assist their 'reading' of ER (Entity Relationship) models.

The original ER (Entity Relationship) models has been in use since it's conception by Chen in 1976. An ER model is a graphical model that seeks to identify data relevant to an information system by categorizing such data into 'entities' and 'relationship' and representing these categories in a graphical form using standard notation. ER models are widely used as part of the information systems development process.

1.2 Objectives

The objectives of this research are based on the following questions:

1. How accurately are ER models understood using the standard technique?
2. How comprehensively are ER models understood using the standard technique?
3. How accurately are ER models understood using the NaLER technique?
4. How comprehensively are ER models understood using the NaLER technique?
5. Is there a higher level of user confidence in their understanding when using the NaLER technique?

Based on these questions, the three hypotheses defined for the experiment will be:

H1: *ER models read using the NaLER technique will yield a more accurate understanding than ER models read using the standard techniques.*

H2: *ER models read using the NaLER technique will yield a more comprehensive understanding than ER models read using the standard techniques.*

H3: *Users using the NaLER technique will have a higher level of confidence in their understanding.*

1.3 Procedures for recruiting Participants and Obtaining Informed Consent

Participants will be recruited from the 57.366 (Data Modelling Techniques) class, which has current enrolment 60 students.

Informed consent will be obtained by providing participants with an information sheet about the study (attached) and by answering any queries they might have about the research. In addition, participants will be asked to complete a consent form (attached).

1.4 Procedure in which Research Participants will be involved

The purposed research will involved the experimental comparison of two techniques to reading an E-R model, the standard technique and the NaLER technique.

Under normal circumstances, students undertaking this paper would be taught both these techniques. In week 6, students, who agree, will be randomly assigned to two groups (group 1 and group 2). Each student will be given a copy of a copy of the data model and will be asked to write a short description of what it represents. They will also be given one of two sets of questions (set A or set B) to answer. These questions will ask them to indicate whether a statement about the model is true or false. Finally, the students will be asked how confident they are in their understanding of the model. All students will be given the same instructions.

By week 10, all students will have been taught another technique (NaLER). After completing this session, they will be given the same data model and the alternative set of questions, which they will complete as before. Again, all students will be given the same instruction.

Both experiments will take place in 57.366 tutorial time, and will take about 30 minutes to complete. Material will be distributed at the beginning by the researcher. The researcher will then correct the material as participants complete the tasks. The material would be direct benefit to the students, as it will support their learning process.

1.5 Procedures for handling information and material produced in the course of the research including raw data and final research report(s)

The researcher will be in sole charge of material, and of its analysis. There will be no need to associate student names with responses, also no demographic information is required.

The final research report will be the thesis, and possibly a paper submitted to an academic journal. The results of the data analysis will only be presented in aggregated form, and there will be no opportunity to identify individual students.

2. ETHICAL CONCERNS

2.1 Access to Participants

We will seek to enlist the help of volunteers for this project. All of the participants will be enrolled in Information Systems paper 57.366. This paper, entitled "Data Modelling Techniques", is designed for students who have a background in database concepts. Agreement has been gained from the paper controller and the head of department to proceed with the experiment.

2.2 Informed Consent

The participants' informed consent will be gained through a participant information sheet and consent form (both attached). This form advises the participants that they may ask questions at any point during the study, and may refuse further participation at any stage.

2.3 Anonymity and Confidentiality

All participants will be informed that only the researchers will see their individual data. The results of the questionnaire will not need to be identified with their names, they will just have to be identified by the group of which they are a part. To enhance confidentiality, results will only be presented in aggregated form.

2.4 Potential Harm to Participants

All students will be required to read the data model, and complete the questionnaire to check their understanding of the model. The final grade awarded to each student for the paper 57.366 will not be influenced in any way by participation (or non-participation) in this experiment.

2.5 Potential Harm to Researcher(s)

There is no chance of harm to the researcher(s).

2.6 Potential Harm to the University

There is no change of harm to the University as a result of this experiment.

2.7 Participant's Right to Decline to Take Part

The participants are invited to take part, and quite at liberty to decline the invitation. Further, they are free to withdraw at any time.

2.8 Uses of the Information

All participants will be informed that their individual results will remain totally confidential. The information gained as a result of this experiment will be used as the basis for a Masterate thesis. It might be used as the basis for a paper or papers to be submitted to information systems academic journals or conferences. Although the study will be published in these forms, it will be presented as numerical data only; there will be no reference to individual participants.

2.9 Conflict of Interest/Conflict of Roles

While the researcher is a Masterate student of the Information Systems department, she has no part to play in the paper 57.366. As a result the students will not see her until the actual experiment takes place. So there is no conflict of interest or conflict of roles.

2.10 Other Ethical Concerns

None

3. LEGAL CONCERNS

3.1 Legislation

3.1.1 Intellectual Property legislation

No concerns.

3.1.2 Human Rights Act 1993

No concerns.

3.1.3 Privacy Act 1993

No concerns.

3.1.4 Health and Safety in Employment Act 1992

No concerns.

3.1.5 Accident Rehabilitation Compensation Insurance Act 1992

No concerns.

3.1.6 Employment Contracts Act 1991

No concerns.

3.2 Other Legal Issues

None.

4. CULTURAL CONCERN

None.

5. OTHER ETHICAL BODIES RELEVANT TO THIS RESEARCH

5.1 Ethics Committees

None.

5.2 Professional Codes

None.

6. OTHER RELEVANT ISSUE

None.

Appendix 7 Information Sheet

RESEARCH PROJECT ON AN INVESTIGATION OF DATA MODELS AS TOOLS OF COMMUNICATION

INFORMATION SHEET

Purpose

I would like to invite you to take part in my research project which is investigating data models as tools of communication. Data models are important in information system development, particularly as tools for expressing and communicating business requirement. The ability to understand the information content of data models is a fundamental skill required by everyone involved with them. This experiment is investigating "How good are data models as communication tools?" and "How well are data models understood by the user?"

Description

The purposed research will involved the experimental comparison of two techniques to reading an ER (Entity Relationship) model, the standard technique and the NaLER (Natural Language for ER) technique.

Under normal circumstances, you will be taught these techniques. The first experimental task will take place during weeks 6 or 7. You will be randomly assigned to two groups (group 1 and group 2). You will be given a copy of the data model and will be asked to write a short description of what it represents. You will also be given one of two sets of questions (set A or set B) to answer. These questions will ask you to indicate whether a statement about the model is true or false. Finally, you will be asked how confident you are in your understanding of the model.

By week 10, you will have been taught another technique (NaLER). After completing this session, you will be given the same data model and the alternative set of questions.

I hope you will enjoy this exercise and complete it. However, you have the right to:

Refuse to answer any particular question

Withdraw from the experiment at any time

Ask questions about the experiment at any time

Provide information on the understanding that your name will not be used

Receive a summary of the findings from the experiment when it is completed

Please note that your participation (or non-participation) in this experiment will not affect your grade in 57.366 in any way.

This research project has been considered and approved by the Massey University Human Ethics Committee. Nevertheless, if you have any questions about the experiment, would like further information about the experiment please contact the people whose names appear below, all at Massey University, Turitea Campus.

Researcher: Somrutai Malaipong (Nok)
Department of Information Systems
Tel: ext 7539
Email: S.Malaipong@massey.ac.nz

Supervisor: Ms. Clare Atkins
Department of Information Systems
Tel: ext 4206
Email: C.Atkins@massey.ac.nz

Appendix 8 Consent Form

RESEARCH PROJECT ON AN INVESTIGATION OF DATA MODELS AS TOOLS OF COMMUNICATION

CONSENT FORM

I have read the Information sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I understand I have the right to withdraw from the study at any time and to decline to answer any particular questions.

I agree to provide information to the researchers on the understanding that my name will not be used without my permission. This information will be used only for this research and publications arising from this research project.

I understand that my grade for 57.366 will not be affected in anyway by my participation, or non-participation in, or withdrawal from this experiment.

I agree to participate in this study under the conditions set out in the Information Sheet.

Signed: ....

Name: ....

Date: ....

Appendix 9 Transcripts of the descriptions from Experiment 1

1. Depending on what 'degree' the employee obtains; his/her degree determines what job title they would be categorised as. A factor that may be of a bonus to the employee is their type speed capability, which may be of some advantage in giving them a higher preference over other employees. The 'skill' that the employee has with his/her degree is applied to use, and therefore would be considered an 'employee project skill'. 'City' tells us where the employee's 'employment location' is or who's his assigned and works. 'Emp-Location' and 'Emp-Proj-Skill' sums up the basis of the 'Project'. There may be recursive relationships between "'Employee' and 'Marriage' which may be possible if say a couple work for the same company, but sometimes it is not the case. 'Employee' reports everything about the projects progress to the 'department'. The department makes deals with the 'Employee'. At the end of this all, a 'vendor' is accepted and completes the model.

2. An employee's work placement is determined by the project. A record of whether the employee is married or not determines if the chances of an employee can relocate or not. The employee is accountable to a dept. All dealings with vendors are made through the department. A project determines the type of skills, location and employees for the job. The department is locally based to the employee's work placement.

3. The model represents a Project, which involves employee skill, emp-location. Employee is s supertype, within that is the ENGR and SECT (subtype). There is also information about employee's marriage, which is optional. Every employee belongs or reports to a department, then it makes Deal & than accepts the vendor. In the Emp-location, it is assigned by city.

4. There is one project in the city. The project to do it by the different location and each employee has more than one project-skill. Each dept reports to different employee and they can makes different (or more than one) deal and each vendor accepts more than one deal. For the employee they may or may not marriage. There are many different skills used for this project.

5. In describing this, the model is broken into 2 parts

   1) The first involves all of the entities & relationships above employees.

   2) The second involves all of the entities & relationship is under employees.

   1) For an employee to work on a project, he/she has to have the necessary project skills within the pool of skills that he/she has. The employee could be working in different location of the city regardless of whether they are working in the project or not.

   2) An employee working in a department can be married but not necessary. Within a department, the department can make many deals with vendors.

   Other thing that should be included in section 1 is that

      1. Can an employee not work for a project but still have skills

      2. Can an employee work on a project but have no location at all.

   For the second part

      1. Could you consider unarranged a skill

      2. What about employee working on a number of department i.e. part time

6. The E-R model represents a city with in a state is certain populations, which must assign many employment- locations and each location, must determine the project done by many skills. The emp-project-skill uses many skills. On the other hand each emp-location is worked by many employees who is made by the emp-skill. The employee has 2 categories of the job title namely the Emp-Enrg who reports to the department. This dept makes many deals and these deals are accepted by the vendor. The Emp-Secy may or may not be bound by a marriage as an employee (secretary) within that organisation.

7. This model shows a decipher of a few things though I am not sure if they are all relevant to each other. There are two types of employees; secretary and engineers who may or may not necessarily be related through marriage. Each employee is assigned to one department within the organisation for which they work. Each employee is assigned to one or more (engineering?) projects on which they work, and an engineering project has at least one but usually more employees. The employee's assignation to a project seems dependent on their skills, used to ensure all relevant skills for each project to be completed are given. A project (for argument? - engineering project) is located in one place only. The employees can move from one location to another to verity projects or work on them as necessary. –

Suggesting a lot of travelling representatives or something? The department that an employee works for seems to be the contract vendor for supplies –presumably for the projects employees' work on but possibly for department materials-it is ambiguous.

I think this scenario could be of a contracting engineering firm- a small company that has marriage inside departments. Employees travel to project location to work on the projects (building/designing bridges) and departments make deals with vendors (presuming vendors people who sell in this case) to the department, for departmental use. Not Project otherwise it would be all project).

8. This data model is about an organisation make contract with vendor, and assigning the human resources to do each project.

   Employee can separate into two major categories. Degree and type speed. Each employee might have marriage information. One or more employee reports to a department. Each department makes one or more deals with vendor. One or more deal is accepted by one vendor. Each employee has one or more project skills. Each employee can work on one or more places (location). Each project is done in only one location. One or more employee location is assigned by the city.

9. Department make many deal with vendor. Vendor makes many deals with department. Department have many employees. Employees have different degree and type-speed. Employee has many EM_PROJ_SKILL. Employee may in many EMP-LOCATION. Project determines EMP_LOCATION. City has many EMP_LOCATION. Employee may marriage or not marriage.

10. Each state has one or more cities, which have one project at each location. The employees who work for the city are employed to work on a project location. An employee may work on one or more locations and use skills on that project. Each project requires one or more skills. Employees may be married to other employees. Each department has a number of employees who are engineers. The department deals with vendors. One department may have many dealing with one or more vendors. The employees consist of two types of employees' engineers and secretaries.

11. The model stores information about a business concerning deals made with vendors and information about employees such as their marital status, the skills that they posses, the location in which they are working and the actual project that they are working on. The model also shows that employees can work in more than one location, can posses many skills, must belong to only one department and does not have to be married. It also shows that there must be a vendor to a deal. The model also shows that there are two different categories of employees for which different information is required.

12. This model describe how many employees is assigned to a specific project by a city, which is done by a skilled employee and their employees will have categories like type speed or degree. The employee could be married or not which shown in a marriage entity. These employees should be reported to a department and this department makes many deals, which could be accepted by one vendor.

13. This E-R model shows information about company's employee, it shows information such as whether they are married or not. Their work skill, where they live, what kinds of project skills they have, which department they belong to. It also shows information about a vendor and deal made with the vendor. It too shows whether employees have recursived skill to meet a vendor needs.

14. In this model, Project determines where the employee is to be located(ex City). The Project is done by the Skill level of the Employee. An Employee reports to the Department regarding the project details. And Department makes a deal with the vendor if he accepts the project.

15. This ER model describes employees- Modelling information to do with jobtypes, Skills, Current project and location. The Employee is part of a department, which deals with Vendor. The model describes marriages between two employees.

    Details:

    - The model also brakes down the generalised view of an employee job into an engineer (defined by degree (education)) and a secretary (defined by typing speed).
    - The location of the Employee is closely related to what project they participate in. The project also depends on employee skill base.

16. The diagram shows relationships and business rules between an "Employee" and several other things that it interacts with. The scope appears very large.

    It looks a bit like a data model for the way that a firm separates out information so that it is useful for assigning tasks. City, Deal, Vendor and marriage look a little more 'external' in nature compared to the rest of the diagram that appears to focus on information about employees and the work that they are either doing or could do.

17. The Employee can be ether ENGR or SECY. Employee that is married to other employee's is recorded. All employee report to departments. Departments deal with Vendors to get Projects.

    The skill of the EMP determines if the EMP works on a project. The Project determines where the EMP's location will be the location will be a city.

18. The information in this model is conveying an organisation of some kind where employee details are kept, the employees' skills are recorded and what projects they are worked on. The organisation also keeps record of the deals they make on or contracts. Departments tender for the vendors' contract. The organisation also record whether the employee is married or not and record spouses Emp# and date married.

19. Employee has a sub type of which is a job title. The employee has project skills and works on a location. The employees project skill uses skills, which are then used to do a project. The employee works on a location in a city, which determines the project, which gets done.

    The employee can be married to another employee. The employee reports to a department who then makes a deal, which is given to a vendor, who then may accept the deal. There is one project for one location, there is one city for emp-location but city can have many emp-location. Employee can have many emp-location, many project skills but reports to only one dept. Emp-project-skill has one project but one project can have many emp-project-skill etc.

20. The focus of this e-R model is the employee entity. It is divided into two subtypes- engineers and secretaries. It represents how employees are linked to projects primarily. Marriage entity is clearly shown as an optional link to employee- it doesn't affect employees' relationship with project or vendor. Employees' relationship with project is spit into employee location and employee-project-skill.

21. In this model the department makes a deal with the vendor in which the employee report to the department if the vendor accepted and the city determines the location on which the project is to take place at and the employees work there using these skills.

    The employees are broken up into two and they keep a record of a marriage of the employees. Which to me seems to have nothing related to the workers and the project.

22. Represents some sort of project showing which employee works on what project and where. The employees have two sub groupings of Engineers & Secretaries. The model also shows any links of marriage to other employees. Employees are selected from a project for their skills.

23. The model represents generally the details of a business employee. It shows details on where a particular employee is located from which department and employee belongs within the organisation. It also shows which particular project the employees are currently or have been assigned to. The assignment of employee to project is based on the employee skills and location.

    Furthermore, other extra information is also being recorded, such as marital status of an employee and the organisational departments dealing with outside vendors that influence a particular project.

24. It is about an employee who hold the title he/she had in a dept and whether he/she is married and the type of skill that employee had and the job that the employee is doing now, which the department had the deal with the vendor and the location of the project is hold. That mean that the E-R model is brief describes about the detail of the employee and the job he/she is assigned by the dept.

25. A project is undertaken by an employee/employees, the employees are divided into two categories, these categories include employees who have a degree and employees speed type. The organisation keeps details of employees spouses if they are married for some reason. Projects are undertaken in various locations in various

states/cities by one or many departments. The department involved in the project negotiate a deal with the client (vendor) and once the deal has been accepted by both parties (vendor & department) the process by which the project has been agreed upon is complete.

26. Employee is the central entity. An employee can have a degree or a type speed. An employee reports to (works in?) a department or on a location. An employee location determines a project. An employee may or may not be married and details need to be kept about marriage. A project uses skill (emp-proj-skill is the intermediate table for this). The deal is made by the dept the employee reports to, and the deal is accepted by the vendor. Each entity needs to have information kept about it and has a unique primary key.

27. This model shows an organisation and its association with vendor(s). The model shows information about employees, whether engineer or secretary, any association by marriage, what projects they work on and what department they report to and also what city the projects and employees are at.

28. The model provides information based around the employee attribute. Each entity or the model has a part that leads to the employee. The employee has 2 subscribes either secretary or engineer. The entities provide into the relationship that the employee has to them. The model represents how the employee is related to a particular ''entity e.g. Employee works on location or the skills needed to work on a project.

29. Showing about one instance of an employee the information represented is centered around employee, personal details (information) i.e. partner employee # of them who work at the same company. It also relates the employees assigned project, the location of the employee and the skill level of the employee. These determine the project that the employee should be assigned to i.e. this is a multi-national organisation with branches all around the word, if you are given a project in NZ, it should be assigned to an employee that also live in NZ. Also the employee should have the relevant skills to complete the project.

    It also shows the agreement details between the client and the organisation. Assuming that this is the project they are talking about and are assigned to the employee. It also shows three different types- info about the project, employee personal details and department dealings.

30. Vendor makes a Deal with the Dept of the company, for the designing/ analysis of a particular project. (Either making for the vendor or having the vendor helping with project by supplying s/w products).

    Employee of the Dept are chosen with respect to their Skills for the project (Emp-Pro_Skill) The Employee's location (Emp_Location) is also taken into consideration. Then these staff members take part in the project until completed or scrapped.

    Skill is connected to Emp-Pro-Skill because each project has a need for a person for a particular skill of which a Staff (Employee) might be suited.

    City is connected to Emp-Location for selection at team members.

Marriage details at Employee spouse.

31. One vendor can accept many Deals made by the Department. The department can make many deals and has many employees reporting to it. Employee is made up of two sub classes (Eng and Secy). An employee may be bound by marriage but it isn't necessary, but if there is a marriage recorded t must be to an employee.

    Each employee can have many project skills and can work on many locations. An employee can use any skill from their range of project skills and can use many project skills on one project. Each project determines the employee location, which is assigned to a city. One city may have many employee locations.

32. A Company has a number of projects. Each project needs a number of skills. If the employee has the skill(s), they are assigned to a project. However an employee works at a location, and the location determines the project or vice versa. The employee has personal details. An employee is categorised by their job title. A secretary is married. If the employee has a degree, they report to a dept, which makes a deal that the vendor accepts.

33. In a city are located many employees who are categorised by job title. A project they do depending on the employees project skill. The project the employee does will require many skills. It is optional for employees to be married. Many employees report to one department, which makes many deals with a vendor. A vendor makes many deals.

    This data model obviously represents the activities that occur within a single department in regards to postgrade student who are working for the department.

34. I think that data model is trying to depict the relationships between the employee working in a dept. and his/her job scope, and how a particular vendor is related or handled by the employee in the dept. I would also think that this model shows how a database would store the employee information in the dept and his/her relation to a particular vendor.

Appendix 10 Transcripts of the NaLER sentences from Experiment 2

1. City (City Nm) may assigns one or more Emp_location (Emp#, CityNm). Emp_location (Emp#, CityNm) may determine one project (Project#).

   Emp_location (Emp#, CityNm) may work by one or more employee (Emp#).

   Project (Project#) may be done by one or more Emp_Proj_Skill (Emp#,Skill#).

   Emp_Pro_Skill (Emp#,Skill#,Proj#) may use one or more Skill (Skill#).

   Emp_Proj_Skill (Emp#,Skill#,Proj#)may have one or more Employee (Emp#).

2. Project may done by one or more Employee, skill, project.

   City may assign one or more employee, city number.

   Skill may need by one or more employee, skill, project.

   Employee may has one or more Emp-Pro-Skill (Emp#, Skill#, Proj#).

   Employee(Emp#) work on one or more Emp-location (Emp#, City#).

   Dept (Dname) is reports by one or more employee(Empl#).

   Dept (Dname) may makes one or more Deal(Dname, Vname).

   Vendor (Vname) may accepts one ore more Deal (Dname, Vname).

3. One Employee(Emp#) bound by one or more range(Empl#).

   One Marriage(Emp#) must bound one Employee(Empl#).

   One Dept(Dname) must be reported by one or many Employee(Empl#).

   One Employee(Emp#) reports to one Deal(Dname, Vname).

   One Deal(Dname, Vname) must be made by Dept(Dname).

   One Deal(Dname, Vname) must be accepted by Vendor(Vname).

4. City must assign one or more Emp_locations(Emp#, City#).

   Emp_location must determine one Project (Proj#).

   Project(Proj#) must be mused by or Emp_Project_Skill(Empl# Skill# Proj#).

   Skill (Skill#) must be used by or more Employ_Skill

   Employee(Emp#) work on Emp_locations(Emp#, City_nm).

5. Project (Proj#) must determine one Emp_location (Emp#, Cityname).

   Skill (Skill#) must be made by one or more Emp_Proj_Skill (Emp#, Skill#, Proj#).

   City (Cityname) must assign one or more Emp_location(Emp#, Cityname). Employee(Emp#) may binds one Marriage(Emp#).

   Dept (Dname) must is reported by one or more Employee (Emp#).

6. Each Vendor (Vname) may accept one or more Deals (Dname, Vname).

   Each Deal(Dname, Vname) must be accepted by one Vendor(Vname).

   Each Deal (Dname, Vaname) must be made by one Dept (Dname).

   Each Dept (Dname) may make one or more Deals(Dname, Vname).

   Each Dept (Dname) may be reported by one or more Employee (Emp#).

   Each Employee (Emp#) must report to one Dept (Dname).

   Each Employee (Emp#) may work or one or more Emp_locations (Emp#, Cityname).

   Each Emp_locations (Emp#, Cityname) must be worked by one Employee (Emp#).

7. Each Vendor (Vdame) may accept one or more Deals (Dname).

   Deals must be accepted by a Vendor (Vname).

   Each Deal (Dname) must be made by a Dept#.

   Each Department (Dname) may reported by one or more Employees (Emp#).

   Each Employee (Emp#) must report to a Department(Dname).

   Employee (Emp#) may be bound by a Marriage(Emp#).

   A Marriage(Emp#) must bind an Emp#.

   An Employee(Empt#) may work on one or more Emp_location(Emp#).

   Each Emp_location(emp#) must be on worked by an Employee (Emp#).

   Each City (City name) may be assigned by one or more Emp_location(Emp#).

   Each Emp _location (Emp#) must assigned a City.

8. Each Vendor identified by Vendor name may accept one or more Deals.

   Each Deal identified by Dealname must be accepted by a Vendor.

   Each Deal identified by Dealname, Vendor name must be made by a Dept.

   Each dept identified by Dname may make one or more Deals.

   Each Employee identified by Emp# must report to the Dept.

Each Dept identified by Dname may have one or more Employee.

Each Dept identified by Dname may have one or more Employee.

Each Employee identified by Emp# may be bound by a Marriage.

Each Marriage identified by Emp#, must bind an Employee.

Each Employee identified by Emp# may have one or more Emp_Proj_Skill.

Each Emp_Proj_Skill identified by Emp#, Skill#, Pro#.

9. Project (Proj#) is detemined by one Emp_location (Emp#).
   Project (Proj#) is done by many Emp_Proj_Skill(Emp#).
   Emp_location(Emp#) determines one Project (Proj#).
   Emp-location(Emp#) is marked by one Employee(Emp#).
   Emp_Proj_Skill(Emp#) does one Project(Proj#).
   Emp_Proj_Skill(Emp#) uses one Skill(Skill#).
   Emp-Proj_Skill(Emp#) has one Employee(Emp#).
   Skill(Skill#) is used by many Emp_Proj_Skill(Emp#).
   Employee(Emp#) marks on many Emp_location (Emp#).
   Employee(Emp#) is made by many Emp_Pro_Skill(Emp#).
   Employee(Emp#) optionally binds Marriage(Emp#).
   Employee(Emp#) report to one Dept(Dname).
   Marriage(Emp#) is bound by one Employee(Emp#).
   Dept (Dname) is reported by many Employee(Emp#).
   Dept(Dname) makes many Deal(Dname).
   Deal(Dname) is made by one Dept(Dname).
   Deal(Dname) is accepted by one Vendor(Vname).
   Vendor(Vname) accepts many Deals(Dname).

10. Skill is used by many Emp_Proj_Skill.
    Project is done by many Emp_Proj_Skill.
    One Emp_location is determined by one Proj.
    A City is assinged many Emp_location.
    Employee works on many Emp_location.
    Emplyee has many Emp_ Proj_Skill.
    Marriage must bind in Employee.

Dept is reported by many Emps.

Dept makes many deals.

Vendor accepts many Deals.

Deal must be accepted by one Vendor.

Deal must be made by one Dept.

Emp may be bound by Marriage.

Emp must report to one Department.

11. One Project is determine to one Emp_location and one Employee works on many Emp_location.

    One Employee has many Emp_Project_Skill.

    One Project is done by many Emp_Proj_Skill.

    One Vendor accepts many Deals.

    One Department makes many Deals.

    One Skill is used by many Emp-Proj-Skill.

    One City assigns to many Emp-location.

    One Dept is reported by many Employees.

12. Employee(Emp#) must work on one or many Emp-location(Emp#, Cityname). Employee(Emp#) must have one or many Emp_Proj_Skill (Emp#, Skill#, Proj#). Employee(Emp#) many binds one Marriage(Emp#).

    Dept(Dname) must have one or many Employee(Emp#).

    Dept(Dname) must make one or many Deal(Dname, Vname).

    Vendor(Vname) must makes one or many Deal(Dname, Vname).

    Skill(Skill#) must use one or many Emp-Proj-Skill(Emp#, Skill#, Proj#).

    Project(Proj#) must be done by one or many Emp-Proj-Skill(Emp#,Skill#, Proj#). Project(Proj#) must determines one Emp-location(Emp#, Cityname).

    City(Cityname) must assigned by one or many Emp-location(Emp#, Cityname).

13. Deal (Dname, Vname) must be made by Dept (Deptname).

    Project (Proj#) must be done by Emp-Proj-Skill.

    Skill (Skill#) must be used by Emp_ Proj Skill (Emp#, Skill#, Proj#).

    Emp-engr (degree) must have Emp-Proj-Skill (Emp#, Skill#,Proj#).

    Project (Proj#) is determined by Emp-location (Emp#, Cityname).

Emp_location (Emp#,Cityname) must be assigned by City (City name).

14. Dept = each Dept is uniquely identify by Dname one dept identify by Dname must have one phone.

    Employee = each Employee is uniquely identify by Emp# are employee identify by Emp# must have one date-birth. One Employee identify by Emp# must have one Dname. One Employee identify by Emp# must have one job_title.

    Vendor = each Vendor is uniquely identify by Vname one Vendor identify by Vname must have one address.

    Deal = each deal is uniquely identify by Dname, Vname. One deal identify by Dname, Vname must have one last meet.

    Skill = each Skill is uniquely identify by Skill# one Skill, identify by Skill# must have one description.

    Project = each Project is uniquely identify by Proj# one Project identify by Proj# must have one Est_cost.

    Emp_Pro_skill = each Emp_Proj_ Skill is uniquely identify by Emp#, Skill# Proj#.

    City = each City is uniquely identify by Cityname. One City identify by Citynm, must have one state. One City identify by Citynm must have one population.

    Emp_location = each Emp_location is uniquely identify by Emp#, Cityname one Emp-location identify by Emp#, Cityname, must have one Proj#.

15. One City (CityNm) assigned one or more Emp-location (Emp#), (CityNm).

    One Emp_location (Emp#,CityNm) determines one Project (Pro#).

    Many Emp_location is assigned by one City (CityNm).

    One Project (Pro#) is determined by Emp_location (Emp#).

    One Project is done by one or more Emp_Proj_Skill (Emp#, Proj#,Skill#).

    Many Emp_Proj_kil does one Project (Pro#).

    Many Emp_Proj_Skill uses one Skill (Skill#).

    One skill (Skill#) is used by one or more Emp_Proj_Skill (Emp#,Proj#,Skill#).

    Many Emp_Location is worked by one employee (Emp#).

    One Employee works on many Emp_location (Emp#,CityNm).

    Many Emp_Proj_Skill (Emp#,Proj#,skill#) is made by Employee(Emp#).

    One Employee (Emp#) has one or more Emp_Proj_Skill(Emp#,Proj#,Skill#).

    One Employee may be bound by one marriage (Emp#).

One Marriage binds one Employee (Emp#).

Many Employee (Emp#) reports to one Dept (Dname).

One Dept (Dname) is reported by one or more Employee (Emp#).

One Dept (Dname) makes one or more Deal (Dname,Vname).

Many Deals are made by one Dept.

Many Deal is accepted by one Vendor.

One Vendor accepts many Deals.

16. A Project is done by an Emp_Proj_Skill.

An Emp_Pro_Skill has skill.

A project is determined by Emp_Pro_Skill.

An Emp_Location is assigned by City.

An Employee works on an Emp_Location.

An Employee has Emp_Proj_skill.

An Employee is bound by Marriage.

A married binds an Employee.

Marriage is linked to an Employee.

An Employee has personal detail.

An Employee has a job title.

An Employee is categorised by their job title.

A job title is Engr and Secy.

An Employee reports to a Dept.

A Dept makes many Deals.

A Deal is accepted by a Vendor.

A Vendor accepts a Deal.

A Vendor can accept many Deals.

A Deal can only made with one Dept.

A Dept has many Employees.

An Employee can be an Engr & Secy.

17. City has a location.

An Employee works on one Emp_Location.

One Employee binds one or many Marriage.

One Dept makes one or many Deal.

Many Deals are accepted by one Vendor.

18. Employee (Emp#) is worked by one or more Emp-Location (Emp#).
Employee (Emp#) is made by one or more Emp_Proj_skill (Emp#,Proj#,Skill#).
Employee (Emp#) may be bound by Marriage (Emp#).
Employee (Emp#) reports to one Depts (Dname).
Vendor (Vname) accepts one to many Deals (Dname,Vname).
Dept (Dname) reports to one or more Employees (Emp#).
Dept (Dname) makes one or more Deals (Dname,Vname).
Deal (Dname,Vname) is accepted by one Vendor (Vname).
Deal (Dname,Vname) is made by one Dept (Dname).
Skill (Skill#) is used by many Emp_Proj_Skill (Emp#,Proj#,skill#).
Project (Proj#) does many Emp_Proj_Skill (Emp#,Proj#,skill#).
Emp_Proj_Skill (Emp#,Proj#,skill#) is done by one Project (Proj#).
Emp_Proj_Skill (Emp#,Proj#,skill#) uses one Skill (Skill#).
Emp_Proj_Skill (Emp#,Proj#,skill#) has one Employee (Emp#).
City (CityNm) assigns many Emp_Locations (Emp#,CityNm).
Emp_Location (Emp#,CityNm) determines one Project (Proj#).
Emp_Location (Emp#,CityNm) is worked on by one Emp (Emp#).
Marriage (Emp#) binds one Employee (Emp#).

19. Vendor (Vname) may have one or more Deals (Dname).
Deal (Dname) must have one Vendor (Vname).
Deal (Dname) must have one Dept (Dname).
Dept (Dname) may have one or more Deals (Dname).
Dept (Dname0 may have one or more Employee (Emp#).
Employee (Emp#) must have one Dept (Dept#).
City (CityNm) may have one or more Emp_Location (Emp#, CityNm).
Emp_Location (Emp#, CityNm) must have one City (CityNm).
City (CityNm) may have one or more Emp_Location (Emp#, CityNm).
Emp_Location (Emp#, CityNm) must have one City (CityNm).
Emp_Location (Emp#, CityNm) must have one Project (Proj#).

Project (Proj#) must have one Proj_Location (Emp#, CityNm).
Project (Proj#) may have one or more Emp_Proj_Skill (Emp#,Proj#, Skill#).
Emp_Proj_Skill (Emp#,Proj#, Skill#) must have one Project (Proj#).
Emp_Proj_Skill (Emp#,Proj#, Skill#) must have one Skill (Skill#).
Skill (Skill#) may have one or more Emp_Proj_Skill (Emp#,Proj#, Skill#).

20. Emp_Proj_Skill (Emp#,Proj#, Skill#) does one Project (Proj#).
Project (Proj#) is done by one or many Emp_Proj_Skill (Emp#,Proj#, Skill#).
Vendor (Vname) accepts one or many Deals (Dname, Vname).
Deal (Dname, Vname) is made by one Dept (Dname).
Dept (Dname) makes Deal one or many Deals (Dname).
City (CityNm) is assigned by one or many Emp_Location (Emp#, CityNm).
Emp_Location (Emp#, CityNm) assigns one City (CityNm).
Skill (Skill#) is used by one or more Emp_Proj_Skill (Emp#,Proj#, Skill#).
Emp_Proj_Skill (Emp#,Proj#, Skill#) uses one Skill (Skill#).
Project (Proj#) determines one Emp_Location (Emp#, CityNm).
Emp_Location (Emp#, CityNm) is determined by one Project (Proj#).

21. Deal could be with only one Vendor.
Dept can have many Deals.
Employee reports to many Dept.
An Employee can have o or 1 marriage.
Employees have many Locations.
A Location should be in one City.
Employees have many Emp_Proj_Skill.
A Project uses many Skills.
Employee must have Skill.
A Vendor can get many Deals.
A Dept have one or more Employee working for it.

22. City (CityNm) may have one or more Emp_Location (Emp#,CityNm).
Employee (Emp#,CityNm) may have one or more Employee (Emp#).
Emp_location (Emp#,CityNm) has one Project (Propj#).

Project (Propj#) has one or more Emp_Proj_Skill (Emp#,Proj#, Skill#).
Emp_Proj_Skill (Emp#,Proj#, Skill#) may have one or more Skill (Skill#).
Emp_Proj_Skill (Emp#,Proj#, Skill#) may have one Employee (Emp#).

23. Each City is uniquely identified by CityNm.
Each City is identified by CityNm must have one state.
Each City is identified by CityNm must have population.

24. Each City is uniquely identified by CityNm.
City (CityNm) must have many Emp_Location (Emp#,CityNm).
Each Project is uniquely identified by Proj#.
Emp_Location (Emp#,CityNm) must have one Project (Proj#).
Each Skill is uniquely identified by Skill (Skill#).

25. City (CityNm) may assign one or more Emp-location (Emp#, CityNm).
Emp-location (Emp#, CityNm) may only determine one project (Proj#).
One Emp-location (Emp#, CityNm) is worked by one Employee (Emp#).
One project (Proj#) is done by one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
One Emp-Proj-Skill is made by one Employee (Emp#).
Emp-Proj-Skill (Emp#, Skill#, Proj#) uses one Skill (Skill#).
Each Skill (Skill#) is used by one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
Each Employee (Emp#) has one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
Each Employee (Emp#) works on one or more Emp-Location (Emp#, CityNm).

26. One City (City Nm) must/may assigns many Emp-Location (Emp#).
One Emp-Location (Emp#) must/may be assigned by one City (City Nm).
One Emp-Location (Emp#) may determines one project (Project#).
One Project (Project#) may be determined by one Emp-Location (Emp#).
One Emp-Location (Emp#) may be worked by one Employee (Emp#).
One Employee (Emp#) may works on many Emp-Location (Emp#).
One Employee (Emp#) may has many Emp-Proj-Skill (Emp#, Skill#, Proj#).
One Emp-Proj-Skill (Emp#, Skill#, Proj#) may be made by one Employee (Emp#).

27. A City (City Nm) must has many Emp-Location (Emp#, City Nm).
    A Project (Proj#) must determines one Emp-Location.
    An Employee (Emp#) must works on many Emp-Location.
    An Employee (Emp#) must have one or many Emp-Proj-Skill.
    An Emp-Proj-Skill uses one Skill.
    An Employee may binds Marriage.
    An Employee reports to one Dept.
    A Dept is reported by many Employee.
    A Dept makes many Deals.
    A Deal is accepted by a Vendor.
    A vendor accepts many Deal.
    A Deal is made by a Dept.
    A Project is done by many Emp-Proj-Skill.

28. Dept (Dname) makes one or more Deal (Dname, Vname).
    Vendor (Vname) accepts one or more Deal (Dname, Vname).
    Dept (Dname) is reported be one or more Employee (Emp#).
    Employee (Emp#) may be bound by one Marriage (Emp#).
    Employee (Emp#) has one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
    Skill (Skill#) uses one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
    Project (Proj#) is done by one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).
    Project (Proj#) determines one Emp-Location (Emp#, City Nm).
    City (City Nm) assigns one or more Emp-Location (Emp#, City Nm).
    Employee (Emp#) works on one or more Emp-Location (Emp#, City Nm).
    Deal (Dname, Vname) is accepted by one Vendor (Vname).
    Deal (Dname, Vname) is made by one Dept (Dname).

29. Each City is uniquely id. By City Name.
    Each Dept is uniquely id. By Dept Name.
    Each Vendor is uniquely id. By VName.
    Each Deal is uniquely id. By Dname, Vname.
    Each Skill is uniquely id. By Skill#.
    Each Project is uniquely id. By Proj#.

Each Emp-Proj-Skill is uniquely id. By Emp#, Skill#, Proj#.

Each Emp-Location is uniquely id. By Emp#, CityNm, Proj#.

Each Marriage is uniquely id. By Emp#.

One Dept must have one phone.

One Vendor must have one Address.

One Deal must have one last meet.

One Skill must have one description.

One Project must have one Est-Cost.

One City must have one State & Population.

One Emp-Location must have one Proj#.

One Marriage must have one Date of Marriage, Spouse emp#.

One City, project#, emp# is assigned to Emp/Location.

One Deal must have a Deptname, Vendor.

30. A City (City Nm) has many Employment Locations (Emp#).

    An Employment Location (Emp#) is determined by the Project (Proj#).

    A Project (Proj#) is done by many Employed Project Skills (Emp#, Skill#, Proj#).

    A Skill (Skill#) is used by one or many Employed Project Skill (Emp#, Skill#, Proj#).

    An Employee has one or many Employed Project Skills.

    An Employee is bound by one or no Marriage (Emp#, Spouse-Emp#).

    An Employee works on one or many Employment Location (Emp#, Proj#).

    An Employee reports to one or many Department (Dname).

    A Department (Dname) makes one or many Deal (Dname, Vname).

    A Deal (Dname, Vname) is accepted by a Vendor (Vname).

    A Vendor (Vname) accepts one or many Deal (Dname, Vname).

31. Each Employee is identified Emp#.

    One Employee (Emp#) must work on one location.

    One employee (Emp#) must have one employee name.

    One employee (Emp#) must have one Emp-Proj-Skill.

    One Employee (Emp#) must report to a Dept.

Each Emp-Proj-Skill is uniquely identified by Emp#.

Each Emp-Project-Skill must use Skill.

Each Skill is uniquely identified by Skill#.

One Skill must have one skill description.

Each Dept is identified by Dept name.

One dept has one phone.

Each project is uniquely identified by Proj#.

One project (Project#) is determined by Proj_location.

32. Project (Proj#) may be done by many Emp-Project-Skill (Emp#, Skill#, Proj#).

Emp-Location (Emp#, City Nm) must determine the project.

Skill (Skill#) may be Emp-Project skill.

Emp-Project-Skill (Emp#, Skill#, Proj#) must has Employee (Emp#).

Emp-Location (Emp#, City Nm) must be work by Employee (emp#).

City (City Nm) may be assigned by many Emp-location.

Emp-Location (Emp#, City Nm) must assigns City (City name).

Employee (Emp#) may binds Marriage (Emp#).

Employee (Emp#) must reports to Dept (Dname).

Employee (Emp#) may works on Emp-Location (Emp#, City Nm).

Employee (Emp#) may has Emp-Project Skill (Emp#).

33. Project (Proj#) may have one or more Emp-Proj-Skil (Emp#, Skill#, Proj#).

Project (Proj#) may have one Emp-Location (Emp#, Proj#).

City (City Nm) may have one or more Emp-Location (City Nm, Emp#).

Skill (Skill#) may have one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).

Emp-Proj-Skill (Emp#, Skill#, Proj#) may have one Skill (Skill#).

Emp-Proj-Skill (Emp#, Skill#, Proj#) may have one Employee (Emp#).

Employee (Emp#) may have one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).

Employee (Emp#) may have one or more Emp-Location (Emp#, Proj#).

Employee (Emp#) may have one Dept (Dname).

Dept (Dname) may have one or more Employee (Emp#).

Dept (Dname) may have one or more Deal (Dname, Vname).

Deal (Dname, Vname) may have one Dept (Dname).

Deal (Dname, Vname) may have one Vendor (Vname).

Vendor (Vname) may have one or more Deal (Dname, Vname).

Employee (Emp#) may have one Marriage (Emp#).

Marriage (Emp#) may have one Employee (Emp#).

Emp-Location (Emp#) may have one Project (Proj#).

Emp-Location (Emp#) may have one City (City Nm).

34. City (City Nm) assigns one or more Emp-Location (Emp#, City Nm).

Emp-Location (Emp#, City Nm) is assigned by one City (City Nm).

Project (Project#) determines one Emp Location (Emp#, City Nm).

Emp Location (Emp#, City Nm) is determined by one Project (Project#).

Project (Project#) is done by one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).

Emp-Proj-Skill (Emp#, Skill#, Proj#) is done by one Project (Project#).

Emp-Proj-Skill (Emp#, Skill#, Proj#) uses one one Skill (Skill#).

Skill (Skill#) is used by one or many Emp-Proj-Skill (Emp#, Skill#, Proj#).

Emp_ Location (Emp#, City Nm) is worked by one Employee (Emp#).

Employee (Emp#) works on one or many Emp_Location (Emp#, City Nm).

Emp-Proj-Skill (Emp#, Skill#, Proj#) is made by one Employee (Emp#).

Employee (Emp#) has one or more Emp-Proj-Skill (Emp#, Skill#, Proj#).

Vendor (Vname) accepts one or more Deal (Dname, Vname).

Deal (Dname, Vname) is accepted by one Vendor (Vname).

Deal (Dname, Vname) is made by one Dept (Dname).

Dept (Dname) makes one or more Deal (Dname, Vname).

Dept (Dname) is reported by many Employee (Emp#).

Employee (Emp#) reports to one Dept (Dname).

Employee (Emp#) binds to zero or one marriage (Emp#).

Marriage (Emp#) binds one Employee (Emp#).

35. Dept (Dname) makes one more Deal (Dname, Vname).

Vendor (Vname) accepts one or more Deal (Dname, Vname).

Employee (Emp#) works on one or more Emp-Location (Emp#, City Nm).

Dept (Dname) reported by one or more Employee (Emp#).

Employee (Emp#) has one or many Emp-Proj-Skill (Emp#, Skill#, Proj#).

Skill (Skill#) is used by one or many Emp-Proj-Skill (Emp#, Skill#, Proj#).

Project (Proj#) done by one or many Emp-Proj-Skill (Emp#, Skill#, Proj#).

Marriage (Emp#) bind none or one Employee (Emp#).

Employee (Emp#) bound by 0 or one Marriage (Emp#).

Deal (Dname) accepted by many or one Vendor (Vname).

City (City Nm) assigned one or many Emp Location (Emp#, City Nm).

36. Project (Proj-id) must determine one Emp-Location identified by (Emp-id).

Emp-Location (Emp-id) must assigns one City identify by (City Nm).

Emp-Location (Emp-id) must works on one Employee identified by (Emp-id).

Employee (Emp-id) must reports to one Dept identify by (Dept-name).

Employee (Emp-id) must have one or more Emp-Proj-Skill identify by (Emp-id).

Employee (Emp-id) may binds one Marriage identify by (Emp-id).

Dept (Dname) must makes one or more Deal identify by (Dname).

Vendor (Vname) must accept one or more Deal.

Appendix 11 The data results from Experiment 1

Common set number	Question TYPE	Set A Q. no.	Set A (23 subjects)																							Total	%
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		
1	M	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	100
2	D	2	0	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	1	0	16	70
3	S.																										
4	S.	3	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	20	87
5	M	4	0	1	0	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	0	1	16	70
6	S.	5	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	21	91
7	M																										
8	M	6	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	20	87
9	M	7	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	20	87
10	S.	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	100
11	S.	9	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	19	83
12	D	10	0	1	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	6	26
13	M																										
14	M	11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	22	96
15	D	12	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	19	83
16	S.	13	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	21	91
17	S.	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	100
18	S.	15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	100
19	S.	16	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	83
20	S.																										
21	M																										
22	D	17	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	21	91
23	S.	18	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	21	91
24	M																										
25	M																										
26	S.																										
27	S.	19	0	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0	0	14	61
28	D	20	0	1	1	1	0	0	0	0	1	1	0	1	1	0	1	1	1	1	0	0	1	1	1	14	61
29	D																										
30	D																										
31	D																										
32	M																										
33	D																										
34	D																										
S Total			6	8	8	9	9	8	10	10	9	10	9	9	9	8	9	9	10	10	9	8	10	9	8	204	89
M Total			4	5	4	4	4	5	5	4	3	5	5	5	5	4	5	4	4	4	5	4	4	4	5	101	88
D Total			1	4	4	4	3	3	3	2	3	4	3	5	4	3	5	4	5	2	2	2	3	4	3	76	66
Total			11	17	16	17	16	16	18	16	15	19	17	19	18	15	19	17	19	16	16	14	17	17	16	381	83
																										MEAN	
Confident Correct			3	3	3	3	3	2	2	3	2	2	2	1	2	2	3	3	3	2	3	3	2	2	2	2.44	
Confident Complete			3	3	2	4	3	3	2	4	3	2	3	2	2	3	3	3	3	2	3	2	5	2	2	2.78	

S = Simple M = Medium D = Difficult

Common set number	Question TYPE	Set B Q. no.	Set B (12 subjects)												Total	%
			1	2	3	4	5	6	7	8	9	10	11	12		
1	M															
2	D															
3	S.	1	1	1	1	0	1	1	1	1	1	1	1	1	11	92
4	S.	2	0	0	1	0	0	1	1	0	1	0	1	0	5	42
5	M															
6	S.	3	1	1	0	1	1	1	1	1	0	1	1	1	10	83
7	M	4	1	1	0	1	1	0	1	1	1	1	0	1	9	75
8	M															
9	M															
10	S.	5	1	1	1	1	1	1	1	1	1	1	1	1	12	100
11	S.	6	0	1	1	1	1	1	1	1	1	1	1	1	11	92
12	D															
13	M	7	1	0	1	1	1	1	1	0	1	1	1	1	10	83
14	M															
15	D															
16	S.	8	1	1	1	1	1	1	1	0	1	0	1	1	10	83
17	S.	9	1	1	1	1	1	1	1	1	1	1	1	1	12	100
18	S.															
19	S.															
20	S.	10	1	0	1	1	0	1	1	1	0	1	1	0	8	67
21	M	11	1	0	0	0	0	1	1	1	0	1	1	0	6	50
22	D															
23	S.	12	0	1	1	1	1	1	1	0	0	0	1	1	8	67
24	M	13	1	1	0	1	1	1	0	0	0	0	0	0	5	42
25	M	14	1	1	0	1	1	1	1	1	1	1	1	0	10	83
26	S.	15	1	1	1	1	1	1	1	1	1	1	1	1	12	100
27	S.															
28	D															
29	D	16	0	0	0	0	0	1	0	0	0	1	0	0	2	17
30	D	17	0	0	0	0	0	1	0	0	1	1	0	0	3	25
31	D	18	1	1	0	0	1	0	0	0	1	0	0	0	4	33
32	M															
33	D	19	0	0	1	0	0	1	0	0	0	0	1	1	4	33
34	D	20	1	1	0	1	1	0	1	0	0	1	1	1	8	67
S Total			7	8	9	8	8	10	10	7	7	7	10	8	99	83
M Total			5	3	1	4	4	4	4	3	3	4	3	2	40	67
D Total			2	2	1	1	2	3	1	0	2	3	2	2	21	35
Total			14	13	11	13	14	17	15	10	12	14	15	12	160	67
															MEAN	
Confident Correct			3	3	3	3	3	3	2	3	2	3	3	4	2.92	
Confident Complete			3	4	3	3	2	2	3	3	3	3	3	4	3	

S = Simple M = Medium D = Difficult

Appendix 12 The data results from Experiment 2

Common set number	Question TYPE	Set B Q.no.	Set B (20 subjects)																				Total	%
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
1	M																							
2	D																							
3	S.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	100
4	S.	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	2	10
5	M																							
6	S.	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	3	15
7	M	4	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	4	20
8	M																							
9	M																							
10	S.	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	19	95
11	S.	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	18	90
12	D																							
13	M	7	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	0	15	75
14	M																							
15	D																							
16	S.	8	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	19	95
17	S.	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	100
18	S.																							
19	S.																							
20	S.	10	0	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1	13	65
21	M	11	1	1	1	1	1	0	1	0	1	1	1	1	0	0	1	0	0	1	0	1	13	65
22	D																							
23	S.	12	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	95
24	M	13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	100
25	M	14	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	5	25
26	S.	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	S.																							
28	D																							
29	D	16	1	0	1	1	1	0	1	1	0	1	1	1	1	0	1	0	0	1	0	0	12	60
30	D	17	1	1	1	1	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	0	15	75
31	D	18	0	0	1	1	1	1	0	1	1	1	0	0	1	1	1	0	1	0	0	0	11	55
32	M																							
33	D	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	16	80
34	D	20	1	0	1	1	1	0	1	0	0	0	0	1	1	0	0	1	0	1	1	1	11	55
S Total			6	7	7	5	7	7	7	6	6	7	7	7	6	8	8	5	8	6	6	7	133	67
M Total			4	3	3	3	3	2	3	3	3	3	4	4	1	2	3	1	2	3	3	3	56	56
D Total			4	2	5	5	5	3	4	3	2	4	3	3	5	3	2	2	2	4	3	1	65	65
Total			14	12	15	13	15	12	14	12	11	14	14	14	12	13	13	8	12	13	12	11	254	64
																							MEAN	
Confident Correct			3	3	2	2	2	1	3	2	1	2	3	4	2	3	2	2	2	2	2	3	2.30	
Confident Complete			1	3	2	2	2	1	3	2	2	3	3	2	2	3	2	2	2	3	2	4	2.30	

S = Simple M = Medium D = Difficult

Common set number	Question TYPE	Set A Q.no.	Set A (17 subjects)																	Total	%
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17		
1	M	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	0	1	1	12	71
2	D	2	0	0	0	1	1	1	1	1	1	1	1	0	1	1	0	1	0	11	65
3	S.																				
4	S.	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	2	12
5	M	4	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	7	41
6	S.	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	12
7	M																				
8	M	6	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	1	5	29
9	M	7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	100
10	S.	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	16	94
11	S.	9	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	15	88
12	D	10	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	13	76
13	M																				
14	M	11	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	14	82
15	D	12	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	15	88
16	S.	13	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16	94
17	S.	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	100
18	S.	15	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	15	88
19	S.	16	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	14	82
20	S.																				
21	M																				
22	D	17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6
23	S.	18	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	14	82
24	M																				
25	M																				
26	S.																				
27	S.	19	1	0	0	1	0	1	0	0	0	1	1	1	0	1	1	0	1	9	53
28	D	20	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1	0	1	7	41
29	D																				
30	D																				
31	D																				
32	M																				
33	D																				
34	D																				
S Total			7	7	5	8	7	8	9	7	6	8	8	8	4	6	8	7	7	120	71
M Total			3	3	2	4	3	3	2	4	3	3	3	4	4	5	2	3	4	55	65
D Total			2	3	4	3	4	3	3	1	3	3	3	2	2	4	3	3	1	47	55
Total			12	13	11	15	14	14	14	12	12	14	14	14	10	15	13	13	12	222	65
																				MEAN	
Confident Correct			3	2	3	2	2	2	2	2	3	2	2	2	3	3	2	4	3	2.47	
Confidént Complete			3	3	3	2	2	2	3	2	2	3	2	3	4	3	3	4	3	2.78	

S = Simple M = Medium D = Difficult

Appendix 13 Data analysis from Minitab

True and False Questions

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Technique	fixed	2	NaLER	Standard

Analysis of Variance for Total, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	81.225	54.717	81.225	28.26	0.000
Technique	1	88.159	72.728	88.159	30.67	0.000
Question*Technique	1	35.031	35.031	35.031	12.19	0.001
Error	68	195.460	195.460	2.874		
Total	71	399.875				

Unusual Observations for Total

Obs	Total	Fit	StDev Fit	Residual	St Resid
1	11.0000	16.5652	0.3535	-5.5652	-3.36R
29	17.0000	13.3333	0.4894	3.6667	2.26R
31	10.0000	13.3333	0.4894	-3.3333	-2.05R
69	8.0000	12.7000	0.3791	-4.7000	-2.84R

R denotes an observation with a large standardized residual.

Least Squares Means for Total

Technique		Mean	StDev
NaLER		12.88	0.2796
Standard		14.95	0.3019
Question			
A		14.81	0.2711
B		13.02	0.3095
Question* Technique			
A	NaLER	13.06	0.4112
A	Standard	16.57	0.3535
B	NaLER	12.70	0.3791
B	Standard	13.33	0.4894

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Technique	fixed	2	NaLER	Standard

Analysis of Variance for Simple, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	12.844	4.488	12.844	11.44	0.001
Technique	1	51.062	49.369	51.062	45.48	0.000
Question*Technique	1	0.188	0.188	0.188	0.17	0.683
Error	68	76.350	76.350	1.123		
Total	71	140.444				

Unusual Observations for Simple

Obs	Simple	Fit	StDev Fit	Residual	St Resid
1	6.0000	8.8696	0.2209	-2.8696	-2.77R
39	5.0000	7.0588	0.2570	-2.0588	-2.00R
49	4.0000	7.0588	0.2570	-3.0588	-2.98R

R denotes an observation with a large standardized residual.

Least Squares Means for Simple

Technique		Mean	StDev
NaLER		6.854	0.1748
Standard		8.560	0.1887
Question			
A		7.964	0.1695
B		7.450	0.1935
Question* Technique			
A	NaLER	7.059	0.2570
A	Standard	8.870	0.2209
B	NaLER	6.650	0.2369
B	Standard	8.250	0.3059

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Technique	fixed	2	NaLER	Standard

Analysis of Variance for Medium, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	14.4000	9.4631	14.4000	22.05	0.000
Technique	1	13.5508	12.1114	13.5508	20.75	0.000
Question*Technique	1	1.6455	1.6455	1.6455	2.52	0.117
Error	68	44.4038	44.4038	0.6530		
Total	71	74.0000				

Unusual Observations for Medium

Obs	Medium	Fit	StDev Fit	Residual	St Resid
24	5.00000	3.33333	0.23327	1.66667	2.15R
26	1.00000	3.33333	0.23327	-2.33333	-3.02R
50	5.00000	3.23529	0.19599	1.76471	2.25R
66	1.00000	2.80000	0.18069	-1.80000	-2.29R
69	1.00000	2.80000	0.18069	-1.80000	-2.29R

R denotes an observation with a large standardized residual.

Least Squares Means for Medium

Technique		Mean	StDev
NaLER		3.018	0.1333
Standard		3.862	0.1439
Question			
A		3.813	0.1292
B		3.067	0.1475
Question* Technique			
A	NaLER	3.235	0.1960
A	Standard	4.391	0.1685
B	NaLER	2.800	0.1807
B	Standard	3.333	0.2333

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Technique	fixed	2	NaLER	Standard

Analysis of Variance for Difficult, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	2.669	4.850	2.669	2.46	0.122
Technique	1	2.067	3.914	2.067	1.90	0.172
Question*Technique	1	17.655	17.655	17.655	16.24	0.000
Error	68	73.928	73.928	1.087		
Total	71	96.319				

Unusual Observations for Difficult

Obs	Difficult	Fit	StDev Fit	Residual	St Resid
1	1.00000	3.30435	0.21741	-2.30435	-2.26R
73	1.00000	3.25000	0.23315	-2.25000	-2.21R

R denotes an observation with a large standardized residual.

Least Squares Means for Difficult

Question		Mean	StDev
A		3.035	0.1667
B		2.500	0.1904
Technique			
NaLER		3.007	0.1720
Standard		2.527	0.1857
Question* Technique			
A	NaLER	2.765	0.2529
A	Standard	3.304	0.2174
B	NaLER	3.250	0.2332
B	Standard	1.750	0.3010

Appendix 14 Data analysis from Minitab

Confidence Questions

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Techniqu	fixed	2	NaLER	Standard

Analysis of Variance for confidence correct, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	0.1174	0.4113	0.1174	0.30	0.588
Techniqu	1	1.0579	1.4319	1.0579	2.66	0.107
Question*Technique	1	1.8067	1.8067	1.8067	4.55	0.037
Error	68	27.0041	27.0041	0.3971		
Total	71	29.9861				

Unusual Observations for confidence correct

Obs	confidence	Fit	StDev Fit	Residual	St Resid
12	1.00000	2.43478	0.13140	-1.43478	-2.33R
52	4.00000	2.47059	0.15284	1.52941	2.50R
59	1.00000	2.30000	0.14091	-1.30000	-2.12R
62	1.00000	2.30000	0.14091	-1.30000	-2.12R
65	4.00000	2.30000	0.14091	1.70000	2.77R

R denotes an observation with a large standardized residual.

Least Squares Means for confidence correct

Technique		Mean	StDev
NaLER		2.385	0.1039
Standard		2.676	0.1122
Question			
A		2.453	0.1008
B		2.608	0.1151
Question* Technique			
A	NaLER	2.471	0.1528
A	Standard	2.435	0.1314
B	NaLER	2.300	0.1409
B	Standard	2.917	0.1819

General Linear Model

Factor	Type	Levels	Values	
Question	fixed	2	A	B
Techniqu	fixed	2	NaLER	Standard

Analysis of Variance for confidence complete, using Sequential SS for Tests

Source	DF	Seq SS	Adj SS	Seq MS	F	P
Question	1	0.4694	0.0714	0.4694	0.89	0.350
Technique	1	2.5025	2.9628	2.5025	4.73	0.033
Question*Technique	1	1.3521	1.3521	1.3521	2.55	0.115
Error	68	35.9954	35.9954	0.5293		
Total	71	40.3194				

Unusual Observations for confidence complete

Obs	confidence	Fit	StDev Fit	Residual	St Resid
21	5.00000	2.78261	0.15171	2.21739	3.12R
73	4.00000	2.30000	0.16269	1.70000	2.40R

R denotes an observation with a large standardized residual.

Least Squares Means for confidence complete

Technique		Mean	StDev
NaLER		2.474	0.1200
Standard		2.891	0.1295
Question			
A		2.715	0.1164
B		2.650	0.1328
Question* Technique			
A	NaLER	2.647	0.1765
A	Standard	2.783	0.1517
B	NaLER	2.300	0.1627
B	Standard	3.000	0.2100