A Psychological Study on the Use of Relationship Concept -- Some Preliminary Findings

Keng Siau, Yair Wand, Izak Benbasat

Faculty of Commerce and Business Administration
University of British Columbia
2053 Main Mall, Vancouver BC
Canada V6T 1Z2

Abstract. Conceptual models are central to information systems analysis and design. One of the main constructs in information modelling is the concept of relationship which is popularized by the entity-relationship model. Despite the prominence of relationship concept in information modelling for the last two decades, little or no empirical research has been conducted to understand its usage by users. Several recent studies found that the relationship concept is problematic in information modelling, particularly the degree and connectivity (i.e., structural constraints) of the relationship. In our research, we investigate the importance of relationship concept for information modelling through a series of psychological experiments. This paper will present one of the studies that was conducted to investigate the use of relationship concept by expert users.

1 Introduction

Conceptual models can be considered as the cornerstone of information systems analysis and design. Not only do conceptual models provide the abstraction required for thinking about data-intensive applications [30], they also provide a formal basis for tools and techniques used in developing and using information systems [25]. In addition, conceptual models facilitate the communication between designers and end users [24] by providing them with a conceptual representation of the enterprise that does not include many of the details of how the data is physically stored.

One of the most popular and prominent conceptual models is the entity-relationship (ER) model introduced by Chen [8]. The ER model differs from the earlier models in that it adopts the view that the real world consists of entities and relationships, and that it makes a clear distinction between them. Brodie [5] states that *"the popularity of the ER model for high level database design is due to its economy of concepts and the widespread belief in entities and relationships as natural modelling concepts."* Chen [8] claims that the incorporation of relationships among entities makes the ER model easier to use than previously proposed models. The results of several empirical studies [1, 2, 7, 10, 18, 19, 27] generally indicate that ER model, compared to earlier models, is easier to understand by end users and that designers and end-users perform better when using the ER model.

During the last few years, the object-oriented (OO) approach has become increasingly important and popular; and it undoubtedly will be (if it is not already) the new

paradigm in information modelling. Kim [21] defines the core OO modelling concepts as consisting of object, object identifier, attributes, methods, encapsulation, message passing, class, class hierarchy, and inheritance. One of the main distinctions between OO and ER models, as noted by some researchers [11, 17], is that the abstraction of association (represented by relationship in ER models) is not directly supported in OO models but is instead achieved indirectly by allowing interobject references. Graham [17] indicates that *"(one) problem (with OO models) is the difficulty of modelling relationship types in an object-oriented database."* Elmasri and Navathe [11] conclude that *"this is an inherent weakness of the OO approach and is due to the fact that this approach treats each object as a self-contained unit of information."* Kim [21] also mentions that *"the core (OO) model, though powerful, simply does not capture some of the semantic-integrity constraints and semantic relationships that are important to many types of application."*

Even though the relationship concept is downplayed or ignored by some of the earlier OO models, it has resurfaced in some of the recently proposed OO models [4, 6, 13, 17, 20, 23, 26] -- sometimes under different names such as link or association. The reason is nicely summarized by Booch [4] -- *"an object by itself is intensely uninteresting. Objects contribute to the behavior of a system by collaborating with one another."* Similarly, Embley *et al.* [13] state that *"objects... are often meaningless unless we understand some relationship among them."* This leads to their proposal of object-relationship model. The OO models by Champeaux *et al.* [6] and Rumbaugh *et al.* [26] even include the notion of object relationships which is almost identical to the one proposed by Chen [8]. Martin and Odell [22] also introduce the concept of object-relationship diagram.

The revival of relationship concept in information modelling is important. A good conceptual model should allow one to build a description of the subject matter that is consistent to the way humans conceptualize the same subject matter [24]. If relationship is one of the main concepts that humans use to view the world, then the notion of relationship should exist in a conceptual model.

1.1 Problems with Relationships

Although the relationship concept has been utilized heavily in information modelling, it is also one of the most problematic concepts. Goldstein and Storey [16] found that users of an automated database design tool had difficulty distinguishing between relationships and attributes. Elmasri and Navathe [11] mention that *"it is sometimes convenient to think of a relationship type in terms of attributes."* Similarly, Codd [9] states that *"one person's entity is another person's relationship."* A recent empirical study by Batra *et al.* [2], which compares the relational and ER representations, discovered that *"the most notable error found in the solutions prepared by the subjects was the incorrect representation of connectivity of relationships."*

The main reason for this unfortunate state is the lack of understanding on the use of relationship concept by users. Although many researchers have accepted, or once accepted, that relationship is natural in modelling the real world [8, 11, 29], this belief is based mainly on common sense and intuition rather than systematic and empirical

evaluation. However, our common sense can be false and radically misleading at times. As researchers, we need to approach the problem scientifically and provide, at least, an explanation of why certain concepts are important or why not. Common sense can never serve as a substitute for empirical research [15]. Our research, therefore, will attempt to fill this gap by empirically investigating the usage of relationship concept through a series of psychological experiments. This paper presents one of the studies that was conducted to understand the use of relationship concept by expert users.

2 Mandatory versus Optional Relationship

A structural constraint (sometimes known as cardinality constraint) can be considered as a constraint on the number of objects that must participate in a mapping. The structural constraint is usually specified using a pair of integer numbers (min, max). The minimum number indicates the least number of objects to which a given object must map. The maximum number indicates the greatest number of objects to which a given object must map. This gives rise to two types of relationship -- optional and mandatory.

For example, consider the following two diagrams. The first diagram indicates that a person entity can work for zero, one, or many employers. The second diagram, on the other hand, depicts that an employee must work for at least one employer.

Optional Relationship between Person and Employer

Mandatory Relationship between Employee and Employer

The optional relationship, thus, corresponds to the structural constraint $(0, *)$, which indicates that there may be zero, one or many interactions. The mandatory relationship, on the other hand, corresponds to the structural constraints $(x, *)$, $x \geq 1$, which indicates that there must be at least one interaction.

2.1 Importance of Distinction Between Mandatory and Optional Relationship

The proper use of mandatory and optional relationships in information modelling is important as it ensures that the semantics of the domain is correctly captured. Although the distinction between mandatory and optional relationships is rather

straight-forward and clear-cut, it is not difficult to discover that mandatory relationship is actually a "subset" of optional relationship. Rather than specifying a relationship as (1, *) (i.e., a mandatory relationship), the designers or users could easily specify it as (0, *). Though the second specification is "not wrong", it is not "correct" either. By changing the minimum participation from 1 to 0 (i.e., mandatory to optional), the underlying semantics of the relationship is changed. For example, by changing the minimum cardinality from 1 to 0 in the second diagram (as shown below), we have changed the meaning of the relationship. In this case, the structural constraint indicates that an employee need not work for any employer which is literally incorrect because a person is not an employee unless s/he works for an employer. Thus, there is a contradiction between the structural constraint and the underlying semantics in the diagram.

The incorrect use of structural constraint also directly affects the accuracy and integrity of the database. A mandatory relationship indicates that the existence of an entity depends on its being related to another entity via the relationship type. The use of an optional relationship when a mandatory relationship is warranted will result in an inconsistent database state. For example, the use of optional instead of mandatory relationship between employee and employer may produce employee records in the database that do not relate to any employer record. In short, if we want to capture the concept that "every employee instance must be related to an employer instance," we have to specify a mandatory relationship and not an optional relationship.

However, one wonders whether users really pay attention to the difference between mandatory and optional relationships? For example, on what basis does users decide whether it is a mandatory or an optional relationship? How would their decision change when they are faced with an information model for an unfamiliar domain?

These issues are addressed in this research using a laboratory experiment. To investigate whether users pay attention to the difference between mandatory and optional relationships, we asked the subjects to select the appropriate structural constraints (i.e., mandatory or optional) for a set of ER diagrams. To analyze how they will react when given models from unfamiliar domains, we provided them with two types of diagrams -- one type from familiar domains and the other from unfamiliar domains. Their selection of structural constraints for these two types of domains will also provide us some cues on their selection procedures.

We chose laboratory experiment in this study because of its superiority in controlling extraneous variables [3]. Another important factor is the ability to manipulate the independent variable [14]. Although field study will provide us with a more realistic background, it is very difficult, if not impossible, for us to manipulate the independent variable in this experiment using field study. The following section discusses the experimental design.

3 Experiment Design

3.1 Independent Variables

The independent variable for this study is domain characteristics which consists of two levels -- familiar group and unfamiliar group. The familiar group consists of questions that are taken from common real-life examples that are well-understood by the subjects in our study -- such as university course selection and enrollment. For unfamiliar group, we used very specific concepts from areas such as neurocognition and psychophysics, which are alien to this group of subjects. With these unfamiliar cases, the subjects have to guess the structural constraints for the ER diagrams. For example, the first diagram below is from a familiar domain whereas the second is from an unfamiliar domain.

All the subjects are presented with the same set of questions from both familiar and unfamiliar domains. This repeated measure design is chosen so that each subject can act as his/her own control. The entire set of questions (i.e., familiar and unfamiliar groups) is randomly ordered in each questionnaire. In other words, every questionnaire is unique in order. This ensures that order effect is controlled by randomization. All the diagrams are presented without structural constraints.

The specific context of the diagrams are not given to the subjects because we wanted to analyze how the subjects will response with incomplete information. This is consistent with real-world scenarios as designers typically do not have a very detailed understanding of the application environment. Most of the time, designers rely on common sense and experience to help them construct and understand information models. Giving the subjects full knowledge of the contexts would also defeat the purpose of the study as the subjects would rely on the case descriptions rather than the information models to decide on the answers. Moreover, diagrams with unfamiliar domains would have become familiar with detailed descriptions.

3.2 Dependent Variables

The dependent variables in this study are the choice of interpretation (i.e., mandatory or optional relationship), and the confidence level of the interpretation. The choice of interpretation is a multiple-choice question. Subjects are presented with a choice of Must or May to describe the relationship. Must corresponds to mandatory relationship whereas May corresponds to optional relationship. We decided to use Must and May instead of (1, *) and (0,*) because the former is more intuitive and understandable. This ease of understanding is important because we intend to replicate this study using naive subjects.

To counter order bias, half of the questions present Must as the first choice and the other half list May as the first choice. This is to control for order effect in case the subjects simply circle the first choice for each question. The confidence of interpretation has values ranging from 1 to 7 (1 -- No confidence, 7 -- Absolute confidence).

In addition to these two dependent variables, we also captured the subject's perceived familiarity with the domain depicted by the ER diagram for each question. This serves as a manipulation check to ensure that the questions in familiar group are indeed familiar and questions in unfamiliar group are truly unfamiliar to the subjects. This domain familiarity variable has values ranging from 1 to 7 (1 -- Not familiar at all, 7 -- Very familiar). An example of the questionnaire format is shown in Appendix A.

3.3 Characteristics of Subjects

The subjects in this study are experts in the area of information modelling. In this paper, we will define experts as users who have substantial training and experience in data modelling. This group of subjects consists of MIS graduate students and professors from a well-known west coast university in North America. The experiment was conducted at the beginning of one of the weekly MIS workshops.

A demographic information sheet was completed by each subject prior to the experiment. A total of 24 subjects participated in this experiment. The subjects' expertise with the entity-relationship model and some of its concepts were also captured in the demographic information sheet and summarized in the table below. The expertise is measured using a scale of 1 to 5 (1 -- Totally unfamiliar, 5 -- Very familiar).

Variable	Mean	Std. Dev.
ER Model	4.17	0.92
Entity	4.33	0.70
Relationship	4.21	0.72
Cardinality	3.67	1.31

As can be seen from the table, the subjects were comfortable with the ER model and its constructs (the average is 4 on a 5-point scale). Cardinality, as expected, has the

lowest score and is the only one with a mean score of less than 4. This is consistent with the findings in other studies (e.g., the study by Batra *et al.* [2] which found that cardinality is one of the most difficult concepts to master).

3.4 Test Characteristics

Test characteristics refer to the method employed for data collection. A paper and pencil test was used in this study. For the choice of structural constraint, two choices were provided to the subjects -- Must or May (i.e., mandatory or optional). An open-ended interpretation test was not chosen because of the difficulties involved in interpreting the subjects' answers. A total of 12 questions consisting of 6 familiar and 6 unfamiliar domains were given to each subject. An instruction set that explains the basic modelling concepts in ER model (e.g., entity, relationship, structural constraints) and gives examples of different structural constraints was attached to the front of the questionnaire. No training or discussion was provided as the instruction set was designed to be self-contained and self-explanatory. Extra care was taken to ensure that the instruction set was not biased. The subjects could refer to the instruction set as and when necessary.

3.5 Hypotheses

The null hypotheses for the study is listed below.

Hypothesis H1: There is no difference in the choice of interpretation between familiar and unfamiliar groups.

Hypothesis H2: There is no difference in the confidence of interpretation between familiar and unfamiliar groups.

The alternative hypotheses to the null hypotheses are that the two groups will be different. A two-tailed test is used since we do not have strong theoretical prediction about the outcomes.

4 Experimental Results

The number of subjects for each group and the number of observations are summarized below. The number of observations for the unfamiliar group is less than 144 (i.e., 24 subjects X 6 questions) because some unfamiliar questions were left blank by the subjects.

Groups	No. of Subjects	No. of Questions Per Subject	No. of Observations
Familiar	24	6	144
Unfamiliar	24	6	131

The first hypothesis regarding the choice of structural constraints for the two groups is analyzed using the nonparametric χ^2 test. The results show $\chi^2 (1,275) = 0.018$ with $p < 0.893$. Thus, the null hypothesis is not rejected which means that there is no significant difference between the choices for the two groups. The following table summarizes the result of χ^2 test.

Frequency	Must Choice	May Choice	Total
Familiar Group	15	129	144
Unfamiliar Group	13	118	131
Total	28	247	275

Table of Group By Choice of Interpretation

As can be seen, the number of May is overwhelmingly greater than the number of Must for both groups. There is almost no difference between the two groups in the choice of structural constraints. Indeed, the percentage difference between the two groups is less than 1% for both choices, as shown in the diagram below.

The hypothesis regarding confidence level between the two groups is analyzed using General Linear Models. The results show that the confidence level of the familiar group is significantly higher than that of the unfamiliar group ($p < 0.0001$). Similarly, the perceived domain familiarity level for familiar group is significantly higher than that of the unfamiliar group ($p < 0.0001$).

Measure	DF	SS	MS	F	Pr > F
Confidence Level	1	346.90	346.90	148.06	0.0001
Domain Familiarity	1	750.75	750.75	340.00	0.0001

The means and standard deviations for the confidence level and perceived domain familiarity are depicted below. The average confidence level for the familiar group is almost twice that of unfamiliar group. As for domain familiarity, the score for familiar group is more than twice the score for unfamiliar group. This is important as it shows that the subjects perceived a great difference between the familiar and unfamiliar groups. In other words, the results of perceived domain familiarity show that the manipulation of familiarity and unfamiliarity domains has been very successful.

	Confidence Level Mean (Std. Dev.)	Domain Familiarity Mean (Std. Dev.)
Familiar Group	4.86 (2.10)	5.68 (1.49)
Unfamiliar Group	2.53 (1.88)	2.26 (1.71)

5 Discussion of Experimental Results

The aim of this experimental study is to investigate how the mandatory and optional relationships will be used by experts. The results show an overwhelming number of May versus Must, even for unfamiliar group. The popularity of May over Must is evident in all the questions (as shown below).

We believe the main reason for the popularity of May is that these subjects realize that May can be considered a 'superset' (as discussed in section 2.1). The use of Must precludes May whereas the reverse is not necessarily true. The choice of Must means that a relationship must exist between the two entities. However, May does not mean that the relationship does not exist. It simply means that the relationship can either exist or not exist. Hence, when faced with uncertainty, it is highly likely that the subjects simply chose the optional relationship -- just to be on the safe side.

	<i>Familiar Group</i>		
	No. of Must Choices	No. of May Choices	Confidence Level Mean (Std. Dev.)
Q1	1	23	4.71 (1.97)
Q2	2	22	4.29 (2.26)
Q3	6	18	5.00 (2.15)
Q4	2	22	4.25 (1.94)
Q5	4	20	5.42 (2.02)
Q6	0	24	5.50 (2.15)
	<i>Unfamiliar Group</i>		
Q7	1	21	3.14 (1.91)
Q8	4	18	2.05 (1.68)
Q9	4	18	2.55 (1.99)
Q10	0	22	1.91 (1.44)
Q11	4	18	3.45 (2.22)
Q12	0	21	2.10 (1.55)

So, what is the significance of this finding? We can assume that this group of expert subjects has the same characteristics as system analysts. In fact, some of the subjects were system analysts, project managers, or consultants before joining the university. As such, the tendency of the subjects in this study to go for optional relationship is expected to hold true for system analysts as well. End users, on the other hand, are not familiar with the intricacy of the relationship concept. It is likely, therefore, that they will spend more effort and time deciding whether to use mandatory or optional relationship. This concern of the novice users could explain the problems and difficulties face by novice users on relationship as reported in several studies [2, 16]. On the other hand, the experts resolve the uncertainty by simply using optional relationships. This difference in attitude might be a potential communication problem or conflict between end users and designers.

Another observation is that the highest confidence level (i.e., 3.45 for Q11) in unfamiliar group is still lower than the lowest confidence level in familiar group (i.e., 4.25 for Q4). It appears that the perceived familiarity of the domain gives the subjects higher confidence in their results. One question that is of particular interest is Q11. For this particular question, the entities are Floor and Button, and the relationship connecting them is Associates. This is the only question in the unfamiliar group that uses common words such as Floor and Button, though the subjects might not know the domain is lift operation. The rest of the questions in that group uses terminology from neurocognition and psychophysics that are unfamiliar to the subjects. Although this question (i.e., Q11) has a substantially lower confidence level than those questions in the familiar group, it has the highest confidence level (3.45) in its group. It thus appears that the use of common words also has a positive effect on the confidence of the subjects.

6 Conclusion and Future Research

The aim of this research is to perform an in-depth analysis of the relationship concept through a series of experiments. This paper describes one of the studies conducted to investigate the use of mandatory and optional relationships by expert users. The results of the study indicate a predominant use of optional over mandatory relationships by experts subjects, even for domains that are totally unfamiliar to them. This indicates that expert users generally go for optional relationships when faced with uncertainty (irrespective of familiar or unfamiliar domain). The results also show that familiar domains cause subjects to feel more confident (more accurately -- overconfident) in their answers.

Another of our studies extends this study by deliberately creating a conflict between the information depicted by the structural constraints and the underlying semantics. The aim of the study is to investigate whether the subjects will focus on the structural constraints or the underlying semantics of the ER diagrams. The results show that expert subjects exhibit *semantic negligence* and *attentional bias*. In other words, expert subjects tend to focus on the information represented by the structural constraints and ignore the underlying semantics depicted by the ER diagrams [28].

Our subsequent studies will investigate the difference between experts and novices on the use of relationship concept. This will allow us to gain a better understanding of the communication problem between end users and designers. As for the contribution, this research is likely to be the first that rigorously analyze a modelling construct using psychological experiments. Thus, this research promises to make significant contribution to information systems modelling. In addition, the methodology and technique used in this research can serve as a framework for future research endeavor in this area.

Acknowledgments. The authors would like to thank the anonymous reviewers for their insightful and valuable comments. This research was supported in part by the Natural Sciences and Engineering, and Humanities and Social Sciences Research Councils of Canada.

References

1. Batra, D. and Srinivasan, A., A Review and Analysis of the Usability of Data Management Environments, *International Journal of Man-Machine Studies*, 36, 1992, pp. 395-417.
2. Batra, D., Hoffer J.A., and Bostrom, R.P., Comparing Representations with Relational and EER Models, *Communication of ACM*, Vol. 33, No. 2, February 1990, pp. 126-139.
3. Benbasat, I., Laboratory Experiments in Information Systems Studies with a Focus on Individuals: A Critical Appraisal, in: *The Information Systems Research Challenge: Experimental Research Methods (Volume 2)*, Benbasat, I. (Ed.), Harvard Business School, 1989.

4. Booch, G., Object Oriented Design with Applications, The Benjamin/Cummings Publishing Company, 1991.
5. Brodie, M.L., On the Development of Data Models, in : On Conceptual Modelling, Brodie, M.L., Mylopoulos, J. and Schmidt, J.W. (eds.), Springer-Verlag, 1984.
6. Champeaux, D.D., Lea D., Faure, P., Object-Oriented System Development, Addison-Wesley Publishing Company, 1993.
7. Chan, H.C., Wei, K.K. and Siau, K.L., User-Database Interface: The Effect of Abstraction Levels on Query Performance, Management Information Systems Quarterly (MISQ), Vol. 17, No. 4, December 1993, pp. 441-464.
8. Chen, P.P., The Entity-Relationship Model: Toward a Unified View of Data, ACM Transactions on Database Systems Vol. 1, No. 1, 1976, pp. 166-192.
9. Codd, E.F., The Relational Model for Database Management : Version 2, Addison-Wesley, 1990.
10. Davis, J.S., Experimental Investigation of the Utility of Data Structure and ER Diagrams in Database Query, International Journal of Man-Machine Studies, 1990, 32, pp. 449-459.
11. Elmasri, R. and Navathe, S.B., Fundamentals of Database Systems, Addison Wesley, 1989.
12. Elmasri, R., Hevner, A. and Weeldreyer, J., The Category Concept: An Extension to the Entity-Relationship Model, Data Knowledge Engineering, Vol. 1, No. 1, 1985, pp. 75-116.
13. Embley, D.W., Kurtz, and Woodfield, S.N., Object-Oriented System Analysis -- A Model-Driven Approach, Yourdon Press, Englewood Cliffs, New Jersey, 1992.
14. Emory, C.W., and Cooper, D.R., Business Research Methods, Fourth Edition, Irwin, 1991.
15. Flanagan, M., and Dipboye, R., Research Settings in Industrial and Organizational Psychology: Facts, Fallacies and The Future, Personnel Psychology, 34, 1981, pp. 37-47.
16. Goldstein, R.C. and Storey, V., Some Findings on the Intuitiveness of Entity-Relationship Constructs, Entity-Relationship Approach to Database Design and Querying, Lochovsky F H (ed.), 1989, pp. 9-23.
17. Graham, I., Object Oriented Methods, Addison-Wesley, 1991.
18. Jarvenpaa, S.L., and Machesky, J.J., Data Analysis and Learning: An Experimental Study of Data Modeling Tools, International Journal of Man-Machine Studies, 1989, 31, pp. 367-391.

19. Jih, K. W.J., Bradbard, D.A., Snyder, C.A., and Thompson G.A., The Effects of Relational and Entity-Relationship Data Models on Query Performance of End Users, *International Journal of Man-Machine Studies*, Vol. 31, 1989, pp. 257-267.
20. Khoshafian, S. *Object-Oriented Databases*, John Wiley & Sons, 1993.
21. Kim, Won, *Introduction to Object-Oriented Databases*, The MIT Press, 1990.
22. Martin , J., and Odell, J.J., *Object-Oriented Analysis and Design*, Prentice Hall, 1992.
23. McGregor J.D., and Sykes, D.A., *Object-Oriented Software Development: Engineering Software for Reuse*, Van Nostrand Reinhold, 1992.
24. Mylopoulos, J., *Conceptual Modeling and Telos*, In: *Conceptual Modeling, Databases and CASE*, John Wiley & Sons, New York, 1992.
25. Rolland, C., and Cauvet, C., *Trends and Perspectives in Conceptual Modeling*, In: *Conceptual Modeling, Databases, and CASE*, John Wiley & Sons, New York, 1992.
26. Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W., *Object-Oriented Modeling and Design*, Prentice Hall, Englewood Cliffs, NJ, 1991.
27. Siau, K.L., Chan, H.C. and Wei, K.K., *User-Database Interaction : Experimental Comparisons of Conceptual and Logical Abstraction Levels*, working paper, National University of Singapore, 1993.
28. Siau, K.L., Wand, Y., and Benbasat, I., *The Use of Relationship Concept in Information Modelling*, University of British Columbia, Working paper, 1995.
29. Teorey, T.J., *Database Modeling and Design -- The Entity-Relationship Approach*, Morgan Kaufmann Publishers, 1990.
30. Willumsen, G., *Conceptual Modeling in IS Engineering*, In: *Executable Conceptual Models in Information Systems Engineering*, Trondheim, Nov. 1993, pp. 11-21.

Appendix A

Circle one of the following two options that more correctly reflects the participation of the Employee entity type in the Assigns relationship type:

1. must assign
2. may assign

What is the confidence level of your above choice?

1	2	3	4	5	6	7
No confidence					Absolute confidence	

What is your familiarity with the domain depicted in the diagram?

1	2	3	4	5	6	7
Not familiar at all					Very familiar	