EXPERIMENTAL COMPARISONS OF ENTITY-RELATIONSHIP AND OBJECT-ORIENTED DATA MODELS

Peretz Shoval

Information Systems Program, Department of Industrial Engineering and Management

Ben-Gurion University of the Negev, Beer-Sheva 84105 ISRAEL

(e-mail: shoval@bgumail.bgu.ac.il; fax: 972-7-6472958)

ABSTRACT

The extended entity-relationship (EER) model is being "threatened" by the object-oriented (OO) approach, which penetrates into the areas of system analysis and data modeling. The issue of which of the two data models is better for data modeling is still an open question. We address the question by conducting experimental comparisons between the models. The results of our experiments reveal that: a) schema comprehension: ternary relationships are significantly easier to comprehend in the EER model than in the OO model; b) the EER model supasses the OO model for designing unary and ternary relationships; c) time: it takes less time to design EER schemas; d) preferences: the EER model is preferred by designers. We conclude that even if the objective is to implement an OO database schema, the following procedure is still recommended: 1) create an EER conceptual schema, 2) map it to an OO schema, and c) augment the OO schema with behavioral constructs that are unique to the OO approach.

INTRODUCTION

The object oriented (OO) approach has spread into various areas of computing that include not only programming, but also systems analysis and design, database management, among others. There is no doubts about the advantages of the OO approach for programming, in that it supports software reuse and information hiding,. However, the superiority of OO approach in earlier stages of software development, i.e., system analysis and data modeling, has not as yet been proven.

I concentrate on data modeling (or conceptual design), an activity aimed at creating a conceptual schema. This schema is usually represented in a diagrammatic form, as it 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 DBMS used for implementation. This conversion is considered simple, being an algorithmic, automatic process. The major problem, however, is to create a good conceptual schema that is semantically correct and comprehensible.

For many years, Entity-Relationship, with its many extensions, generally termed EER, has been the most widely used model for conceptual design. Recently, the OO model has been threatening to replace the EER model. In addition to static (structural) constructs of the model whose representation of data structure is somewhat equivalent to EER representation (e.g., object classes considered equivalent to entity and relationship types) - the OO approach models system behavior through "methods" that are attached to the object classes, with messages passing among them. While this extra capability, unlike EER, provides more than just a data structure model, it does not necessarily mean that EER should be replaced by OO, since EER may indeed be "better" than OO for modeling the data structure. Taking into consideration that an EER schema can easily be mapped to an OO schema (Gogola et al 1993, Kornatzky & Shoval 1994, Nachouki et al 1992, Narasimhan et al 1993), it may be worthwhile to begin the process of data modeling with the design of an EER conceptual schema, then convert it to an OO schema (if the objective is to implement an OO DBMS), and finally add the behavioral constructs.

Such an "indirect" strategy for creating an OO schema could be considered when the EER model is regarded a "better" than the OO model. "Better" can be judged from various perspectives. I address this issue from four perspectives: a) comprehension: understandability of schemas; b) design quality: correctness of schemas; c) time to complete design tasks; and d) designers' preference of models. We have conducted two experimental comparisons which compare the two models on these dimensions. The next section discusses related studies, the following two sections overview the underlying models, then I describe the experiments and their results, and the final section summarizes the results and conclusions.

RELATED STUDIES

Various studies on the evaluation and comparison of data models have been conducted. The earlier studies mainly compared record-based models (hierarchical, network and relational). Later on, EER became the most frequently studied model. EER has been compared with relational and other record-based models, as well as with other semantic models (e.g. Batra et al 1990, Juhn & Naumann 1985). Some studies compared the data models from a designer perspective in an attempt to find out which yields more accurate and precise schemas. Other studies compared the query languages of the data models in order to determine the languages by which the designers compose queries with greater accuracy and speed. Yet, other studies took a user perspective, attempting to determine ease of

comprehension. A survey of earlier studies can be found in Batra & Srinivasan (1992). 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, a point that is stressed in our experiments

Recently we have begun to find studies which evaluate the OO model on the basis of controlled experiments. For example, Agrawal and Sinha (1992) examined the influence of designers' experience in functional analysis and task characteristics on quality of OO schema design. They found out that the OO model does not always lead to good results; the quality of result depends on task characteristics. Palvia et al. (1992) compared OO, ER and DSD models from an end-user perspective, to see which schema is more comprehensible. Their subjects were MIS students, and their database schemas were trivial in terms of size and complexity. Each subject received a version of a database corresponding to one data model, along with 30 questions that evaluate comprehension of the database. The total score on correct answers was used as a measure of comprehension. They concluded that users of the OO model showed a significantly better understanding of the data than users of the other two models. They admit, however, that their conclusion is weak, because comprehension was measured on overall terms only, and suggest that future research be directed to studying specific aspects of comprehension, e.g., different types of relationships.

Bock and Ryan (1993) reported on a comparison of OO and EER models from a designer perspective. They examined correctness of design of eight types of constructs, and measured correctness according to a grading scheme developed by Batra et al. (1990). Their results indicated that EER is better for representing attribute identifiers, unary 1:1 and binary m:n relationships, while there are no significant differences for other dimensions. Also, they found no difference in time to complete the tasks. Although these results are interesting, in our opinion they have limited external validity, mainly because the OO model they used (as in Kroenke 1992) represents relationships only by reference attributes in such a way that does not enable to distinguish between an object that represents a binary relationship from an object that represents a ternary relationship.
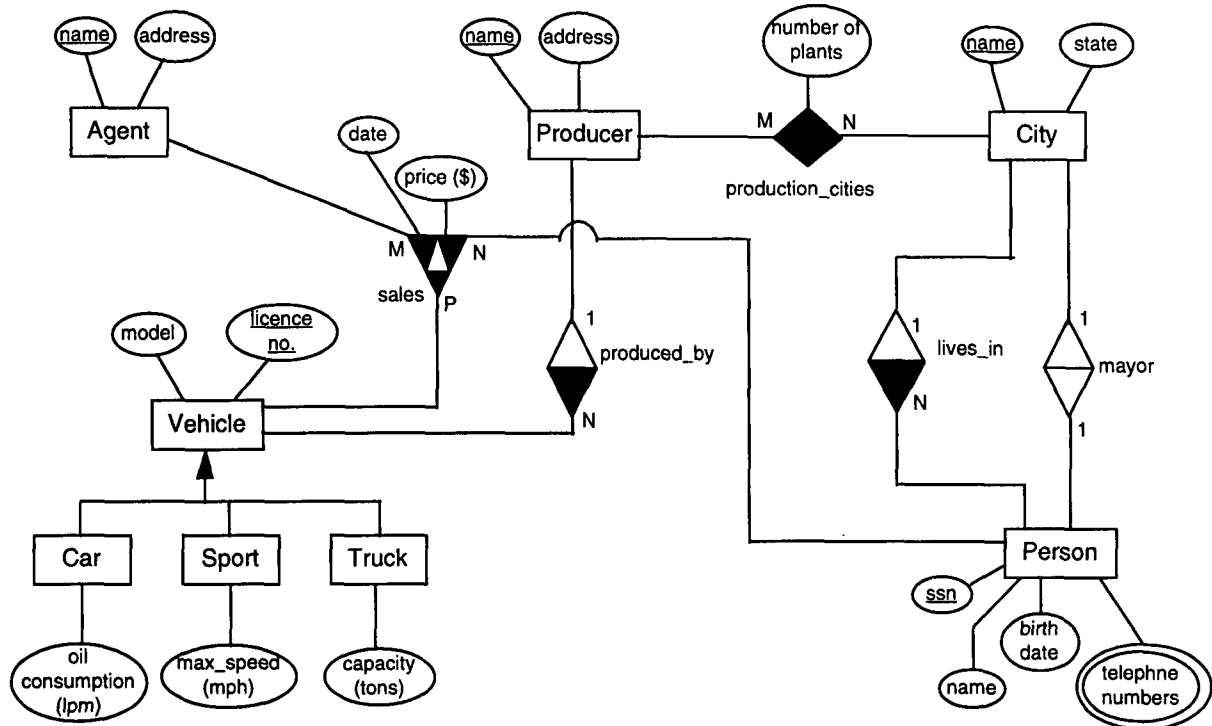
THE UNDERLYING EER AND OO MODELS

The objective of this section is to overview and exemplify the models we use for experimentation. Since each of the models has many variants, we confine ourselves to two specific types, that are sufficiently general to validate the results of the comparison.

The EER model:

As EER is more "standard" than OO, we only present an example (Figure 1) using a diagrammatic notation as in Elmasri and Navathe (1994) and Teorey et al. (1986).

The OO model

The OO model is still evolving and as yet no standard has been defined. The model that we use is based on O_2 (Deux et al 1991) and ODE (Agrawal & Gehani 1989). By extending the model we are able to show all the details of the OO schema on diagram, as would be expected at the conceptual design stage. (Since we focus on modeling data structure and not system behavior, methods are not included.) An example of the OO schema is shown in Figure 2, representing the same reality as the EER diagram in Figure 1.

Figure 1: EER Diagram

An object has attributes. (For brevity, we omit "class" and simply say "object".) An attribute may be atomic or multi-valued, in which case the attribute name is preceded by "set". An identifier attribute of an object is underlined. An attribute may refer to another object, in which case the referenced object name in written in [ ] brackets next to the respective attribute name. This actually implies a bi-directional relationship between the objects, which is signified in two complementary ways: a) reference-attributes are included in each of the involved objects; b) a link connects the two objects, with an indication of the relationship cardinality. In the case of an 1:n relationship between two objects, the object in the "1" side has a set reference attribute to the other object, and in case of an m:n relationship both objects have set reference attributes. An attribute may be a tuple consisting of multiple (value or reference) attributes, and signified by { } brackets. For example, the set producers {[Producer], number of plants} of City.

An m:n relationship can also be represented as a separate object, with reference attributes to the related objects. A ternary or higher order relationships is represented as a separate object, linked to the other objects involved in the relationship. The reference attributes in all objects are included. In the example, we have store facts about agents selling vehicles to persons. Therefore, we used the objects Person, Agent and Vehicle, and for the relationship we created the object Sales. Each of the three original objects has a set reference attribute to Sales, while Sales has a tuple consisting of three reference attributes to those objects. The use of tuple of reference attributes in an object that represents a relationship enables to determine the degree of the relationship represented by that object (e.g., whether it is binary or ternary).

Finally, object hierarchies are signified by links from the subtypes to the supertypes, similar to the EER notation.

Producer

name
address
set where_producing {[City],
      number of plants}
set vehicles_produced [Vehicle]

Agent

name
address
set vehicles_sold [Sales]

City

name
state
set producers {[Producer],
      number of plants}
set inhabitants [Person]
mayor [Person]

Vehicle

licence no.
model
set sold_by/to [Sales]
produced_by [Producer]

Sales

sale {[Agent],[Person],
      [Vehicle]}
price ($)
date_of_sale

Person

SSN
name
birth_date
set telephone_numbers
lives_in [City]
mayor_of [City]
set vehicles_bought [Sales]

Car

oil_consumption (lpm)

Sport

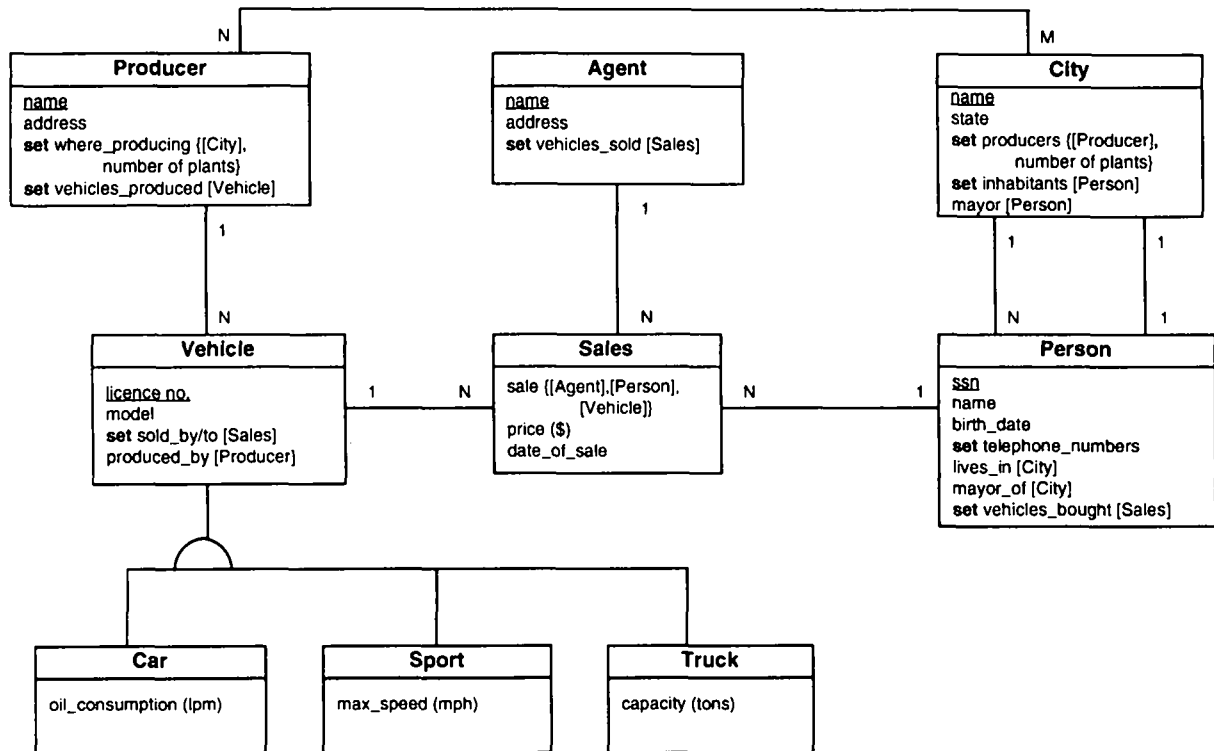max_speed (mph)

Truck

capacity (tons)

Figure 2: OO Diagram

THE USER COMPREHENSION EXPERIMENT AND RESULTS

This section provides a summary of the user-comprehension experiment and its results. For more details (see Shoval & Frumermann 1994).

The experiment

We examined user comprehension as follows: two groups of users, each studying a different model, were given a diagram of the model they studied, and a set of identical questions (statements) about facts in the schema. The users were students of behavioral science and management, having similar backgrounds and experience. They all took the same three computer related courses; therefore, they may be considered as "sophisticated users", i.e., users who can interact with professional analysts, express information needs and approve their products.

The subjects were randomly divided into two groups, each studying one model. The studies emphasized comprehension of the schema diagrams (rather than design of diagrams). The same instructor taught both models (to avoid bias in teaching quality), and the same amount of time - 1.5 hours - was allocated to teach each model. In order to motivate the students, they were told that their performance in the experiment would be considered as part of the final course grade.

We prepared a questionnaire that consisted of 48 "true" and "false" statements about facts in the conceptual schemas. The purpose was to find out not only if there is a difference in overall comprehension between the two schema diagrams, but also if there is a difference in the specific construct types. Therefore, we classified the statements according to different construct types: a) attributes of entities/objects , b) binary-relationships, c) two binary relationships, d) ternary-relationships, and e) other facts - with no direct relationships.

We prepared two equivalent schema diagrams, one for each model. The diagrams included exactly the same facts. The EER diagram included 12 entities, 7 binary-relationships, 2 ternary-relationships and various attributes. In the OO diagram there were 14 objects - some of which representing binary and ternary relationships - each with various types of attributes: atomic, multi-valued, reference-attributes and tuples.

Each participant was given a schema diagram, according to the model s(he) studied, and a questionnaire. Each user was expected to review the diagram and to mark each statement "true" or

"false". The level of comprehension was measured by counting the number of correct answers within each group for each of the categories of statements, and for all statements together.

Analysis of results

The results of the experiment are summarized in Table 1. For each category (including "All") and model it shows the mean grade in absolute values and in percents, and the standard deviation. The last two columns indicate the significance of difference between mean grades, determined according to two-tail Z tests at $\alpha=0.05$.

As it shows, we cannot reject the null hypothesis that overall, there is no difference of comprehension between the two models, although there is a slight difference in favor of the EER model. There is, however, a significant difference in favor of EER for "ternary-relationships": an average of 91.1% for EER vs. only 67.8% for OO. On the other hand, there is an advantage to OO for "other facts": 83.5% to OO vs. 74.1% to EER.

Table 1: Summary of the Results for Comprehension

Type of Construct	No of Questions	Model	Mean Score	Mean in %	STDS	Z-Value	which? (a=0.05)
Attributes	9	EER	7.83	87.00	1.34	-0.37	No
		OO	7.73	85.89	0.99		
Binary Relationships	11	EER	9.98	90.73	0.96	+0.32	No
		OO	10.05	91.36	0.94		
Two Relationships	9	EER	6.73	74.78	1.23	+0.83	No
		OO	6.95	77.22	1.08		
Ternary Relationships	9	EER	8.20	91.10	0.80	-8.55	EER
		OO	6.10	67.80	1.60		
Other Facts	10	EER	7.41	74.10	0.89	+3.47	OO
		OO	8.35	83.50	1.46		
All	48	EER	41.02	83.71	2.49	-1.4	No
		OO	40.03	81.69	3.64		

The insignificant differences between the models with respect to "attributes", "binary-relationships" and "two relationships" can be explained by the similarity in diagrammatic representation. As would be expected, in both models statements involving two binary-relationships are more difficult to comprehend than those involving just attributes or single relationships -- average scores were only 74.78% for EER and 77.22% for OO.

Since we found a significant difference between the models in the case of ternary-relationships, we further examined these results by distinguishing between m:n:1 relationships and m:n:n relationships. In both cases we found a significant difference in favor of the EER model. For m:n:1, the average score of EER was 93.33, compared to an average score of only 78.33 for OO. For m:n:n, the average scores were 90.67 and 62.17, respectively. The advantage of EER on ternary relationships can be explained by the lucidity of the diamond symbol which connects the three entities. It seems that this symbol is clearer and more visible than the "normal" object rectangle which is used in OO for any type of object. Moreover, in OO, a ternary-object includes sets of tuple attributes, which seem to be more complicated to comprehend compared to EER, where the diamond representing the ternary-relationship has only its own attributes.

With respect to "other facts", OO turned out to be superior. This is actually not a specific category as it includes different statements which deal with entities or objects that exist in the diagram but have no direct relationships. OO may be easier to comprehend for this category because here most of the information is encapsulated within the objects, as opposed to EER where it is "spread-out" among the entities and relationships.

THE DESIGN QUALITY EXPERIMENT

This section provides a summary of the design-quality experiment and its results. For more details)see Shoval & Shiran 1996).

The experiment

In this experiment we compared the models from a designer perspective, addressing three dimensions: a) design quality, i.e., correctness of schemas; b) time to complete design tasks, and c) designer

preferences of the models. With respect to design quality we defined null hypotheses that there is no difference in correctness of modeling these constructs types: 1) entities/objects, 2) attributes of entities/objects, 3) inheritance relationships, 4) unary 1:1 relationships, 5) binary 1:1 relationships, 6) binary 1:n relationships, 7) binary m:n relationships, 8) ternary m:n:1 relationships, and 9) ternary m:n:p relationships. With respect to time and preference, we hypotheses that: 10) there is no difference in time to complete design tasks, and 11) there is no difference in designer preferences of the models.

We defined two design tasks, both similar in size and complexity (involving the same types and numbers of various constructs). We created a narrative description for each task, on which basis each subject was asked to design a conceptual schema diagram using one of the two models.

The subjects included 44 students, all majoring in Information Systems, with a similar background and course of studies. All subjects were trained to use the two models. The same amount of time (six hours) was devoted to study each model, and the same instructor taught all the students. In order to motivate the subjects, they were told that their performance in the experiment would be part of the final grade. The subjects were randomly divided into two groups: subjects in one group designed the first task with OO and the second task with EER; subjects in the other group switched models and tasks. The starting and ending times of each task were recorded. At the end, each subject was given a short questionnaire, using 7-point scales, to express an opinion on the ability of each model to design a conceptual schema, and to indicate the relative preference of the models.

Correctness of each schema was measured according to the evaluation scheme described in Batra et al. (1990). Each construct type, in each model, was scored separately by subtracting a certain number of points for each error type (distinguishing between minor, medium and major errors). Then we averaged the scores of all subjects on each construct type for each of the two models.

Analysis of results

The results are summarized in Table 2 and discussed herein.

Table 2: Summary of Results for design Quality, Time and Preference

No	Hypotheses	EER model		OO model		Significance		
		Average	STDV	Average	STDV	T-value	P-value	which?
1	Entity/Class	99.15	2.63	99.62	1.51	-1.22	.23	n
2	Attributes	95.46	3.37	94.23	4.29	1.95	.08	n
3	Inheritance	99.43	3.77	99.43	3.77	0	-	n
4	Unary 1:1	88.07	30.26	70.45	34.64	3.86	.0004	EER
5	Binary 1:1	94.32	22.14	94.32	20.01	0	-	n
6	Binary 1:n	82.58	27.13	88.64	18.86	-1.44	.157	n
7	Binary m:n	81.25	28.08	78.69	28.35	0.42	.68	n
8	Ternary m:n:1	85.23	31.59	67.61	36.80	3.18	.0028	EER
9	Ternary m:n:p	94.32	16.05	76.14	38.86	2.93	.0054	EER

| 10 | Time (minutes) | 53.50 | 16.71 | 68.00 | 15.44 | 5.16 | <.0001 | EER, |

| 11a | How good is model (1-7)? * | 5.77 | 1.05 | 4.70 | 2.82 | 3.45 | <.0001 | EER |

| 11b | Relative preference ** | OO[ ][ ][ ][ ][ ]| [ ][ ] EER   7   6   5   4   3 | 2   1   2.82 | | | | | EER |

* The is based on the Wilcoxon test; 3.45 is Z-value of the test.
** 7 = absolute preference of OO; 1 = absolute preference of EER

A. Correctness of design (hypotheses 1-9):

We used the T-test to measure the difference of mean scores for each of the nine types of constructs. The last columns indicate the significance of the results. The "which?" column indicates if the results are significant (at α=0.05) and which model is favored.

We found no significant difference for six of the constructs, namely entity types or object classes (1), attributes (2), inheritance, or supertype-subtype hierarchies (3), and the three types of binary relationships (5-7). These results are explained by the similarity of the models when dealing with these

constructs. The slight differences in the symbols that represent entities or objects, as well as inheritance links, turned out to be insignificant for design quality. This is evidenced by the high scores on those constructs - above or close to 90. Interestingly, in both models the scores for 1:1 relationships were higher than those for 1:n relationships, which, in turn were higher than those for m:n relationships. These consistent results strengthen their validity.

Contrary to the above, we found significant differences for unary (4) and ternary (8 and 9) relationships; all favored the EER model, as explained herein:

Unary 1:1 relationships: Representation of unary relationships may be complicated, especially in the OO model, because there are different options by which to represent them. For example, the 1:1 unary "marriage" relationship, which appeared in the experiment, can be represented in OO model as a reference attribute in Employee, or as a separate object (Marriage) related to Employee. As it turned out, in both models many errors occurred, but in OO we found more errors in cases where the designers chose to represent the relationship as a separate object.

Ternary m:n:p relationships: There were only a few errors in modeling m:n:p relationships with the EER model, but many errors with OO model. Typical errors in OO occurred when instead of a separate object, the relationship was represented as (complex) attributes of other objects. For example, to represent a ternary relationship between Worker, Project and Skill, some designers defined a tuple of reference attributes {[Skill], [Project], work_hours} in the object Worker.

Ternary m:n:1 relationships: Here too, the EER model was, indeed, better. The results indicate that in both models m:n:1 relationships are more difficult to model than m:n:p relationships, but OO designers made more errors. A typical error was representation of a ternary as two binary relationships (e.g., a ternary relationship between Worker, Project and City was represented by some designers as two binary relationships: one between Worker and Project, another between Project and City).

When we note the similarities and differences between our results and the results obtained in (Bock & Ryan 1993), we see that the OO model was not superior for any dimension. In both studies EER was better than OO for unary 1:1 relationships; but in (Bock & Ryan 1993) EER was better for binary m:n relationships, whereas in our study EER was better for ternary relationships.

B. Time for completing the design task (hypothesis 10):

As seen in Table 2, it takes significantly less time to design EER schemas. Although the time factor in itself is of minor importance, the fact that this result is consistent with the results on correctness of design, strengthens the validity of the experiment.

C. Preference of models by designers (hypothesis 11):
For this dimension as well, EER turned out to be significantly better. This is evidenced both by the scores given to each model independently (row 11a), and the relative preference of the two models (row 11b). The average of 2.82 on a 7-point scale indicated a significant preference of the EER model. Even if we agree that preference is not important as performance, the consistent results further validate the experiment.

SUMMARY AND CONCLUSIONS

In summary, we found that the EER model is superior to the OO model for the following reasons:

1. Comprehension of schemas: EER ternary relationships are easier to understand, whereas for OO, only non-specific ("other") facts are easier to understand, with no significant difference for the other constructs.
2. Design quality: Complex (unary and ternary) relationships are designed more correctly in EER, whereas there is no significant differences when designing other modeling constructs.
3. Time: it takes less time to design an EER schema.
4. Preferences: designers prefer to use the EER model.

Therefore, we conclude that even if the objective is to design and implement an OO schema, the following strategy is recommended: a) design an EER schema and validate it by users/ management; b)

map it to an equivalent OO schema (using an appropriate mapping algorithm or tool; and c) augment the OO schema with the behavioral constructs ("methods" and "messages").

REFERENCES

Agrawal, R. and Gehani, N. (1989) **ODE (object database and environment): The language and the data model**, ACM SIGMOD Intl. Conf. on Management of Data, pp. 36-45.

Agrawal, R. and Sinha, A. (1992) **An empirical study of the effects of experience and task characteristics on object-oriented modeling**, in: WITS'92 Workshop on Information Technology and Systems, pp. 225-234.

Batra, D. (1993) **A framework for studying human error behavior in conceptual database modeling**, Information & Management, Vol. 25 ,pp. 121-131

Batra, D., Hoffer, J. and Bostrom, R. (1990) **Comparing representations with relational and EER models**, Comm. of the ACM, Vol. 33 (2), pp. 126-139.

Batra, D. and Srinivasan, A. (1992) **A review and analysis of the usability of data management environments**, Int'l J. of Man-Machine Studies, Vol. 36, pp. 395-417.

Bock, D. and Ryan, T. (1993) **Accuracy in modeling with extended entity relationship and object oriented data models**, J. of Database Management, Vol. 4 (4), pp. 30-39.

Deux, O. et al. (1991) **The O2 system**, Communications of the ACM, 34 (10), pp. 34-48.

Elmasri, R. and Navathe, S. (1994) **Fundamentals of Database Systems**, 2nd Ed., Benjamin/ Cummings,.

Gogola, M., Herzig, R., Conrad, S., Denker, G. and Vlachantonis, N. (1993) **Integrating the ER approach in an OO environment**, Int'l Conference on ER Approach, pp. 376-389.

Juhn, S. and Naumann, J. (1985) **The effectiveness of data representation characteristics on user validation**, In: Proceedings of the Sixth Int. Conf. on Information Systems, pp. 212-226.

Kerlinger, F. (1973) **Foundations of Behavioral Research**, Holt, Rinhart and Winston.

Kornatzky, Y. and Shoval, P. (1994) **Conceptual design of object-oriented schemas using the binary-relationship model**, Data & Knowledge Engineering, Vol. 14, pp. 265-288.

Kroenke, D. (1992) **Database Processing**, 4th Ed., Macmillan Publishing.

Nachouki, L., Chastang, M. and Briand, H. (1992) **From entity-relationship diagram to an object-oriented database**, Int'l. Conference on Entity-Relationship Approach, pp. 459-481.

Narasimhan, B., Navathe, S. and Jayaraman, S. (1993) **On mapping ER and relational models onto OO schemas**, Int'l Conference on ER Approach, pp. 402-413.

Palvia, P., Lio, C. and To, P. (1992) **The impact of conceptual data models on end-user performance**, J. of Database Management, Vol. 3 (4), pp. 4-15.

Shoval, p. and Frumermann, I (1994) **OO and EER conceptual schemas: a comparison of user comprehension**, J. of Database management, Vol. 5 (4), pp. 28-38.

Shoval, P. and Shiran, S. (1996) **Entity-Relationship and Object-Oriented data modeling - an experimental comparison of design quality**, Data & Knowledge Engineering, forthcoming.

Teorey, T., Yang, D. and Fry, J. (1986) **A logical design methodology for relational database using extended entity-relationship model**, ACM Computing Surveys, Vol. 18 , pp. 197-222.