1985

The Effectiveness of Data Representation Characteristics on User Validation

Sung H. Juhn
University of Minnesota

Justus D. Naumann
University of Minnesota

Follow this and additional works at: http://aisel.aisnet.org/icis1985

The Effectiveness of Data Representation Characteristics on User Validation

Sung H. Juhn
Justus D. Naumann

Department of Management Science
University of Minnesota

ABSTRACT

Many different data modelling or representation schemes have been used or proposed. One important use of such data representations is to communicate the data content of a proposed system design to users: the "user validation" task. The effects of the characteristics of four such data models on user comprehension were investigated in a controlled laboratory experiment. The results showed that the two primarily graphical representations were more understandable than two alternatives for most of a set of tasks designed to simulate user validation.

There were some preliminary indications that the graphical or "semantic" data models led to more systematic data modelling behavior. Relational models did out-perform graphical models with respect to relationship identifier recognition. Additional research is discussed that will more fully explore the role of data representations in systems development. The results of this experiment are also be applicable to "end-user computing."

Introduction

A data model is an intellectual tool that is used to model a portion of reality that is of interest to a person or an organization. As a modeling tool, a data model provides a set of constructs that can be used to specify the inherent structure of data in the reality, the operations that are permitted to be performed on the data, and the constraints that should be maintained for the data to be consistent with the reality [TSIC82]. As a representation tool, it communicates the modeler's view of data to users, analysts, and builders of information systems.

Many different data models are in use or have been proposed. The main focus in data model research to date has been on theoretical issues such as the mathematical foundations of data organzations, or formalization of the modeling constructs. Such theoretical orientation, however, has kept many users and practictioners in the field from understanding and using the models in information systems development processes. To be acceptable and useful to those who are less theoretically inclined, data models also must be researched from such human-factors perspectives as comprehensibility, usability, and usefulness.

This paper describes a study that tested four data models for their comprehensibility and usability to users. The four data models selected for the study were: the Entity-Relationship (E-R) model, the Relational Data Model (RDM), the Logical Data Structure (LDS) model, and the Data Access Diagram (DAD). A controlled laboratory experiment with student subjects was conducted to investigate the effect of representational characteristics of data models on user's understanding and use of a particular database.

In this section, the background and rationale for this research project is discussed and relevant prior research is reviewed. Our scientific and statistical research questions and hypotheses are then presented, along with the research instrument employed in the study. The next section describes the experiment and the details of the experimental results. The final section summarizes the results and suggests additional research.

REPRESENTATIONS

Models or *representations* are central to development and use of systems. Representations have been a major focus of attention in data modeling at least since Hollerith. Card, report, and record layout forms have been employed routinely to analyze, design, specify, and communicate information about data. Computer technological development has permitted manipulation of increasingly complex data structures. More complex—and more abstract—representations have emerged with technological capabilities.

Representations are communication media. In the simple case, a representation of data, a data model, communicates certain facts among individuals. In the process of developing information systems, however, representations do much more. Figure 1 is a model of the role of representations in the systems analysis and specification process. Representation is central to the process in at least the following ways:

1. The form of representation used by the analyst has the role of a *template*. Much of the analyst's task is discovering the set of application knowledge that completes a particular template. The Whorfian hypothesis, "language determines thought," suggests that knowledge not required by a particular representation will not be discovered—indeed, questions that might elicit such knowledge will not even be generated.

2. During the representation-building process, and especially at its completion, the representation must be *validated*. In information systems development, representation-based validation is a critical process. Only the representation of a system that is to be developed is available for validation until system construction and installation has been completed. The representation, then, must promote a clear, comprehensive, and accurate understanding of a system specification by its eventual users.

3. The end result of the systems analysis and design process is a *specification*. A specification is a representation of the system that can be used by system builders. The most critical aspect of a representation as a specification is that it must be *verifiable*. That is, in the subsequent stages of design, development, and testing, builders must be able to test their work against the standard provided by the representation-specification.

These three uses of representations are conflicting. Representations that drive discovery may not effectively communicate to users a clear and comprehensive under-

standing of the eventual system. Representations that are clear and understandable to users may be too imprecise and informal for verification by builders. The most easily understood representations lack sufficient precision and rigor to be useful for either reliable system construction or for verification of subsequent design steps.

Technological development increases our concern for data representations that are simultaneously rigorous and comprehensible. Many authors have predicted the arrival of the era of end-user computing where users will interact directly with database systems to satisfy their own information needs [BENJ82] [EDEL81] [MCLE79] [DICK82] [ROCK81]. Such recent advancements as fourth generation languages, integrated software packages, and applications prototyping bring these projections close to realization. A database is an essential component of user-driven information systems. A majority of end-user information processing activities centers on databases [DATE83]. End-user database representations that are simultaneously comprehensive and comprehensible are therefore in demand to provide usable information [SHNE78].

SPECIFIC DATA MODELS

The Entity-Relationship model (E-R) and the Logical Data Structure (LDS) are "semantic" models that focus on representing the meaning of data without considering implementation constraints. The Relational Data Model (RDM) and the Data Access Diagram (DAD), on the other hand, are "relational-based" models that may be more closely related to data structures visible to users.

Advocates of both semantic models and relational models claim ease of use. Claimed for the semantic models are the naturalness of their constructs "entity" and "relationship," their removal of physical implementation considerations from the data modeling process, and their use of graphics [CHEN76]. Claimed for the relational model is theoretical clarity, simplicity and naturalness of the constructs, plus non-procedural use [CODD82]. Empirical research to support these claims has not been reported. The study reported in this paper is an investigation of the efficacy of data models as representations to information system users.

PRIOR RESEARCH

Several research studies that concentrated on the human-factors aspects of query languages have been reported [REIS81] [EHRE81]. However, only a few human-factors studies have been reported in the data modeling area [BROS78] [SHNE78] [HOFF84]. Empirical, con-
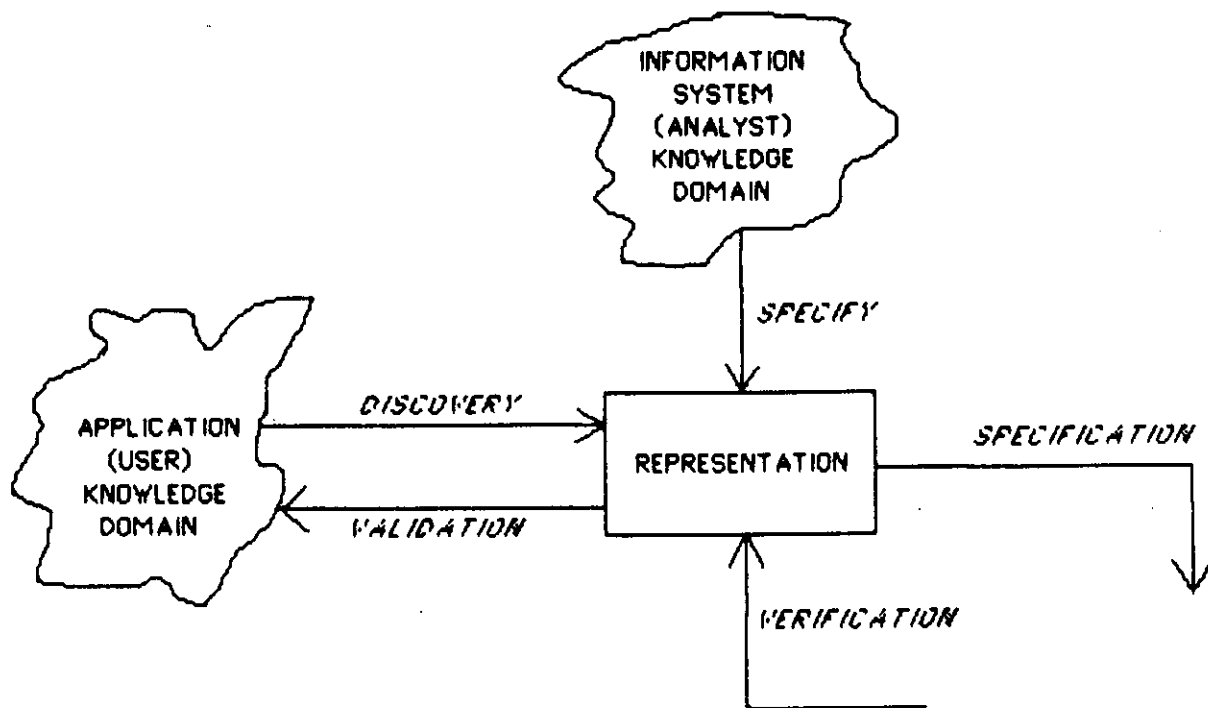
213

Figure 1

Representation Research Model

trolled investigations of the behavior of system users with respect to data models is therefore an important part of the study of representations [MORA81]. Although not an empirical study, the paper by McGee was one of the early works that emphasized the usage aspect of data models [MCGE76]. McGee suggested a set of user evaluation criteria for data models such as simplicity, elegance, picturability, modeling directness, and so forth. Brosey and Shneiderman report one of the first empirical studies of data model usage [BROS78]. They compared the relational model and the hierarchical model for "ease of use." They measured question comprehension, memorization, and behavior in a programming problem-solving task. Their results showed that the hierarchical model was easier to use, but only for those with less programming experience. The authors cautioned that the data modeled had a "natural tree structure," which may have biased the result.

Hoffer reports the results of an investigation of individual images of a database [HOFF84]. He found that subjects had individualized images of a database, and that a process-flow structure was the most frequently used image. He also reported that subjects omitted identification of database access keys from their images and were not able to clearly specify the nature of data relationships.

Durding, Becker, and Gould investigated the way people organize data [DURD77]. They found that people do have mental structures such as hierarchies, networks, lists, and tables available, and that they used them appropriately when organizing data.

Broadbent and Broadbent studied the database structures that subjects preferred [BROA78]. They concluded that individuals did not use a strictly hierarchical system, and that different educational background may influence preference of alternative representations.

The Research Question

The primary objective of this study was to identify the data model characteristics that best communicate to system users. In terms of the representation model, this is the "validation" question: what representation or which characteristics of a specific representation efficiently provide system users with a comprehensible model of data. The four data models (E-R, LDS, DAD, and RDM), each with a unique set of representational characteristics, were selected for comparison. Their representational differences are discussed below.

214

REPRESENTATIONAL DIFFERENCES

The data models selected differ in emphasis or focus on the underlying constructs they represent: entities, attributes, or relations. The primary difference among them, however, is the way they represent "relationships" among data items. The syntax the selected models use to represent relationships is shown in Figure 2.

There are three dimensions of difference in the representation of relationships: representation of relationship existence, representation of relationship cardinality constraints, and representation of relationship dependency constraints. In the E-R and LDS, relationships are specified by lines that connect entities participating in a relationship (Figure 2 a, b). In the E-R model, a relationship is a distinct representation construct (denoted by a diamond-shaped icon) distinct from and connected to participating entities. The LDS model, in contrast, treats a relationship as a "descriptor" of the participating entities, with no relationship icon other than a named connection. (For a many-to-many relation, the LDS requires definition of a new entity to represent the relationship.)

In the relational models, relationships are represented by the presence of common attributes in two or more relations (formally, attributes whose values come from the same domain). The Data Access Diagram (DAD) is the same as an RDM except that common attributes are explicitly and graphically interconnected to show the relationship (Figure 2, c, d).

Figure 3 shows the way in which the cardinality and dependency properties of a relationship are represented in these four data models. These properties must be represented in narrative form in the RDM and DAD models, both graphically and as narrative in the LDS model, and both graphically and iconically in the E-R model.

The E-R model includes a specific icon (the double rectangle) for representing existence dependency constraints. No comparable symbol is available in either the RDM or the DAD. Specification of existence dependency constraints must therefore be made outside the models with prose statements. In the LDS, existence dependency constraints are partially represented by the graphical conventions prescribed. (The RM/T model [CODD79], though semantically more complete than RDM, was not included in this research because its highly theoretical nature was presumed to exclude it from consideration for user validation or system user communication.)

These representational differences are summarized in Figure 4. The following research questions are suggested by the above discussion and by the considerations depicted in Figure 4.

RESEARCH QUESTIONS

1. Do the representational differences outlined above affect the ability of system users to understand the underlying reality?

2. Do the representational differences cause different patterns of reading or understanding?

3. Do graphical models represent and communicate more of the semantics of data to users than the relational models?

4. Are graphical representations easier to understand than non-graphical ones?

5. Is one of the models under study easier to validate than others?

6. Are the concepts used in the models comprehensible to system users?

Development of a research instrument designed to provide insight into these and related questions is described in the next section.

RESEARCH INSTRUMENT

The research questions refer to the utility of alternative data model representations in the validation process. In validation a representation is evaluated with respect to reality by users who share a knowledge domain. In this experiment, given the intended use of student subjects as surrogates for systems users, the case setting had to be based on a knowledge domain common to the subjects. For that reason the "university setting" was chosen as a case.

Data model representation development

Representation of a typical university setting involving entities such as student, faculty, course, and so forth was developed in each of the four data models: E-R, LDS, DAD, and RDM. The entities, attributes, relationships, and integrity constraints—the semantic content—of each model was held constant with respect to the other models. For example, cardinality constraints were added in prose for the models (LDS, DAD, RDM) that did not provide for cardinality in the representation. The semantic content of each model was thus held approximately equivalent among the four experimental treatments.

The models each contain 10 entities and 14 relationships. The size was constrained so that each model could be represented on a single sheet of paper. A one-page de-

a) E-R

STUDENT

STUDENT
DEP

DEPENDENT

b) LDS

STUDENT

supports/is supported by

DEPENDENT

c) RDM

STUDENT

| student id | student name | ------- |

Common
Attribute

DEPENDENT

| dependent name | student id | ------- |

d) DAD

STUDENT

| student id | student name | ------- |

Common
Attribute

DEPENDENT

| dependent name | student id | ------- |
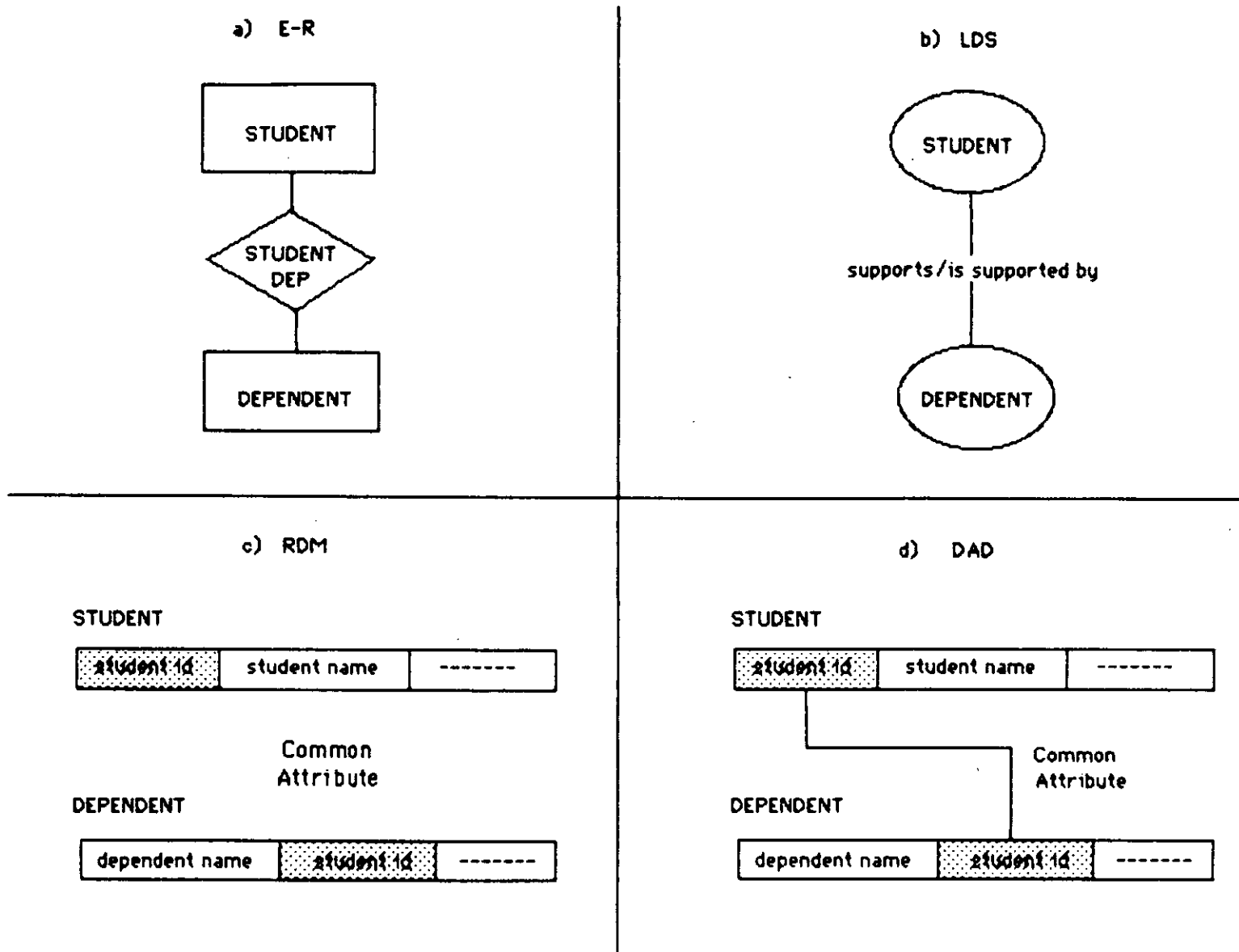
Figure 2

Relationship Existence Representation
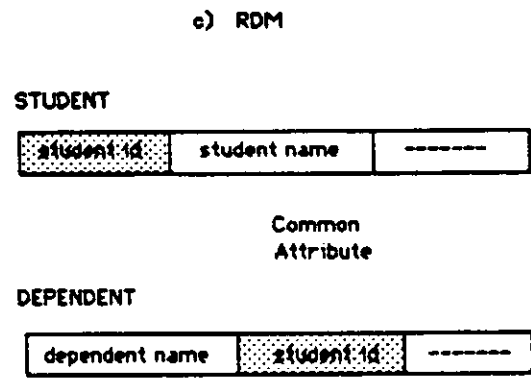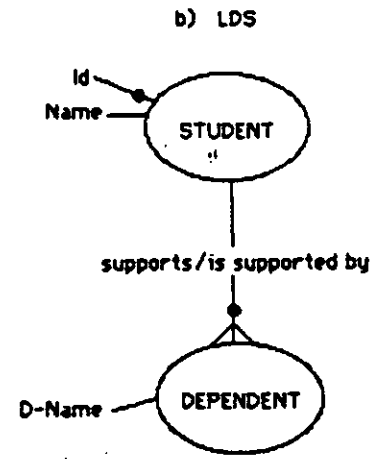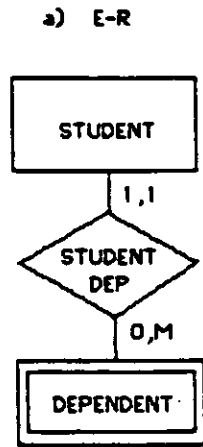
a) E-R

STUDENT

1,1

STUDENT
DEP

0,M

DEPENDENT

b) LDS

Id
Name
STUDENT

supports/is supported by

D-Name
DEPENDENT

c) RDM

STUDENT

| student id | student name | ------- |

Common
Attribute

DEPENDENT

| dependent name | student id | ------- |

Integrity Assertion 1 :
  A dependent cannot exist w/o a supporting student
  (Dependency)
Integrity Assertion 2 :
  A dependent has one and only one supporting
  student. (Cardinality)

d) DAD

STUDENT

| student id | student name | ------- |

Common
Attribute

DEPENDENT

| dependent name | student id | ------- |
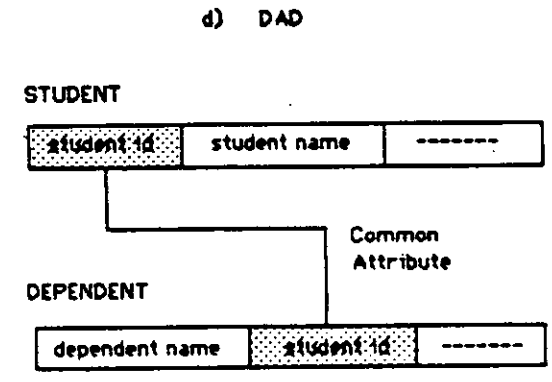
Integrity Assertion 1 :
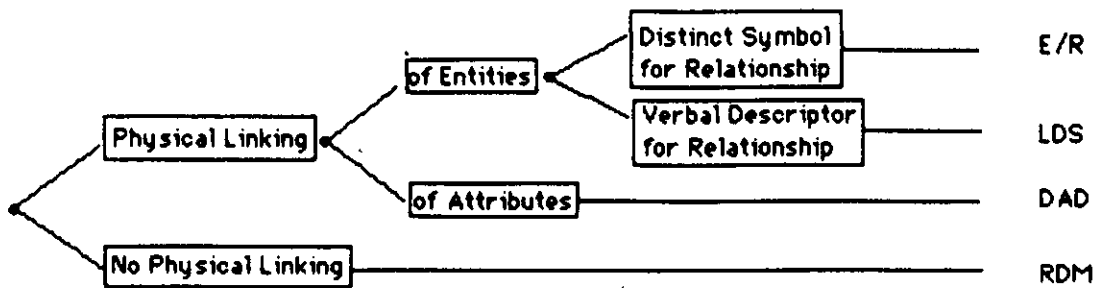 A dependent cannot exist w/o a supporting student
  (Dependency)
Integrity Assertion 2 :
  A dependent has one and only one supporting
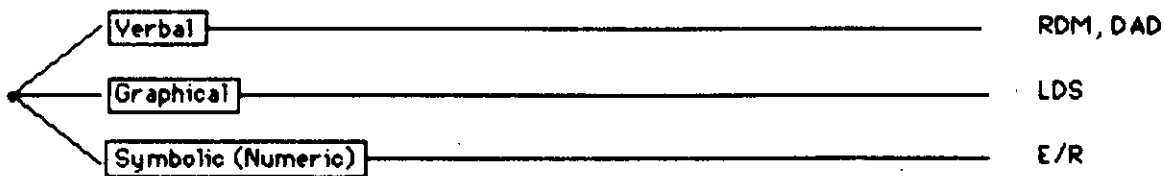 student. (Cardinality)

Figure 3

Relationship Dependency and Cardinality Representation

"RELATIONSHIP EXISTENCE" SPEC: FORMS:

```
                                        ┌─────────────────┐
                                    ┌──│ Distinct Symbol │─────────── E/R
                                   ╱   │ for Relationship│
                      ┌──────────┐╱    └─────────────────┘
                   ┌─│of Entities│◄
                  ╱  └──────────┘ ╲    ┌─────────────────┐
                 ╱                  ╲──│ Verbal Descriptor│────── LDS
  ┌──────────────┐                     │ for Relationship │
 ╱│Physical Linking│◄                  └─────────────────┘
╱ └──────────────┘  ╲   ┌────────────┐
◄                    ╲─│of Attributes│────────────────────── DAD
 ╲                     └────────────┘
  ╲ ┌──────────────────┐
   ╲│No Physical Linking│──────────────────────────────── RDM
    └──────────────────┘
```

"RELATIONSHIP CARDINALITY" SPEC:

```
     ┌──────┐
   ┌─│Verbal│────────────────────────────────────── RDM, DAD
  ╱  └──────┘
 ╱   ┌─────────┐
◄───│Graphical│──────────────────────────────────── LDS
 ╲   └─────────┘
  ╲ ┌────────────────┐
   ╲│Symbolic (Numeric)│──────────────────────────── E/R
    └────────────────┘
```

"RELATIONSHIP DEPENDENCY" SPEC:

```
     ┌──────┐
   ┌─│Verbal│────────────────────────────────────── RDM, DAD
  ╱  └──────┘
 ╱   ┌──────────┐
◄───│Identifier│───────────────────────────────────── LDS
 ╲   └──────────┘
  ╲ ┌─────────┐
   ╲│Graphical│──────────────────────────────────── E/R
    └─────────┘
```
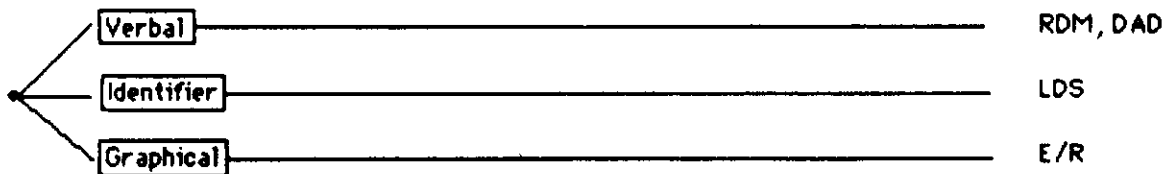
Figure 4

Dimensions of Representational Differences

scription of the constructs used in the model accompanied each model. The models and their descriptions were reviewed by doctoral students in Management Information Systems. Three iterations of review and revision were conducted. The final representations were judged to be correct, complete, and semantically identical by this panel.

Task Development

Five types of tasks were defined as characteristic of the validation process: relationship existence finding, relationship cardinality finding, identifier comprehension, database search, and data model development. Each of these experimental tasks were derived from the research questions defined in the preceding section.

Eighteen questions were designed to measure performance in each of these 5 validation tasks. Questions 1 through 3 asked some general questions about the university database represented in the model. Their purpose was to force the subjects to become familiar with the model. Questions 4 and 5, the relationship finding task, asked subjects to find as many semantic relationships as possible between two entities. For example, Q4 asked for a search for possible relationships between "student" and "faculty." The purpose of the task was to measure the amount of data semantics conveyed to the users (Research question 3).

218

The relationship cardinality finding task required subjects to find the cardinality either directly from the representation (Q6, 7, 8) or by inference from other information in the representation (Q9). Relationship cardinality is represented graphically in the E-R and LDS, and verbally in the DAD and RDM. The task therefore examined which mode of representation, graphical or verbal, was easier for the users to understand (Research question 4 in the preceding section). The identifier comprehension task (Q10, 11, 12) measured subjects' understanding of the nature and role of an identifier: uniqueness of its value (Q10), identifier as a search key (Q11), and update propagation in case of modification (Q12) (Research question 6).

The database search task (Q13, 14, 15, 16, 17) asked subjects to consider each entity or relation in the model as a physical "file" and to list in sequence the names of the files they would search to get the requested information. The purpose of the task was to see if subjects could generate an intuitive search strategy from a data model, and if different models resulted in different strategies (Research question 2). Each of the search questions required multiple file searches. Q13 and 16 required "serial" searches, where only the findings from the previous search process were to be used in each subsequent process. Q15 and 17 required the concept of "joining" the results of multiple searches to identify the desired data. Data Access Diagrams differ from the Relational Data Model only in terms of the explicit physical linking of common attributes in its representation of relationships (See Fig. 2 d). Hence any difference between subject groups using DAD and RDM might also have revealed whether or not such linking afffected the search process of the subjects. Another aspect we expected to determine from the task was whether or not the search sequence generated by the subjects was affected by the way questions were asked. For instance, in Q15, two search criteria were specified: find those who are advisee's of xxx who are enrolled in school yyy. The search sequence question was: did subjects start their search from the criteria that appeared first in the question (advisee's of xxx)?

Finally, the data model development task (Q18) asked subjects to draw a data model depicting the relationships between a specified set of data items (additions to the "university" setting), using the same representational constructs as in the model. By comparing the quality of the models produced by the subjects, this task was intended to provide some insight into which model was easier or harder to use, and which concepts in the model were (not) well understood by the subjects (Research question 5, 6).

Following is the set of specific hypotheses (stated in non-null forms) that relate to these tasks.

HYPOTHESES

Relationship finding task

H_1: There will be significant differences between groups in terms of the number of meaningful relationships found between data items.

H_1a: The subjects using semantic models (E-R, LDS) will find more meaningful relationships between data items than those using the relational models (DAD, RDM).

Relationship cardinality finding task

H_2: There will be significant differences between groups in terms of the accuracy of identifying the cardinality of relationships between data items.

H_2a: The subjects using the graphical representations (E-R, LDS) will more accurately identify the relationship cardinality than those using textual representations (DAD, RDM).

Identifier comprehension task

H_3: There will be significant performance differences between groups in finding and naming the identifiers of the entity/relation construct.

Database search task

H_4: There will be significant performance differences between groups in their database searches.

H_4a: There will be significant performance differences between the DAD group and the RDM group in their database searches.

H_5: Question formats will affect the search process of the subjects.

H_5a: Subjects will start searching from the first criteria mentioned in the question.

Data modeling task

H_6: There will be significant differences between groups in the quality of the models produced.

EXPERIMENT

The four data models and accompanying descriptions, together with instructions and a questionnaire representing the five research tasks, was administered to a selected group of research subjects. Data from completed questionnaires, including background data, were coded and analyzed.

SUBJECTS

Thirty second-year MBA students currently enrolled in a Systems Analysis and Design course participated in the experiment. The subjects had each taken two MIS courses prior to the experiment: an "Introduction to MIS" and a "Program Design and Programming" course. The courses contained no materials directly relevant to the data modeling concepts covered in the study. Subjects had an average of 2 to 3 years of work experience. About half of the subjects either owned a personal computer or had access to one. About one third had experience with some form of Database Management Systems (DBMS). The subject group was assumed to be representative of an end-user population with respect to their level of knowledge in the database area (Table 1). Eight Ph.D. students and two visiting scholars also participated in the experiment but were excluded from the analysis.

Each subject was randomly assigned to one of the four experimental treatments. The number of subjects in each group was controlled (LDS = 7, E-R = 8, DAD = 8, RDM = 7). The experiment was conducted during a class period in the Systems Analysis and Design course. No course credit was given for participating in the experiment, however, the potential future educational benefit of participation in the experiment was noted by the instructor. No negative reaction to the experiment was detected either during or after the experiment.

The behavior of MBA student subjects in this experiment is not the same behavior we would expect of system users. We believe, however, that the *differences* across experimental treatments would also be present in a field setting.

PROCEDURE

Each subject was given an identical set of questions plus one of the four data models. The model was accompanied by one page description of the representation constructs used in model.

The question set was divided into two sections. Thirty minutes were allowed for the first section and 15 minutes for the second. A brief statement about the general purpose of the experiment was made by the investigator at the beginning of the session. Section 1 contained 17 questions on the tasks of the relationship finding, relationship cardinality finding, identifier comprehension, and database search. After 30 minutes, time was called and subjects proceeded to the second section, the data model development task. Subjects were instructed not to go back to the questions in the previous section.

Since the experiment was conducted during a class, the amount of time available for the experiment was limited to 50 minutes. The time limits for the experiment had previously been judged to be reasonable based upon the results of the two pretests with Ph.D. students. After completing the modelling task, subjects completed a questionnaire on their perceptions of the level of difficulty of the data model, the questions, and the concepts used in the model.

RESULTS

The mean scores for the 4 task types are shown in Table 2. The scoring scheme used was: for the relationship finding task, the total count of meaningful relationships identified by the subjects in their response to the two relationship finding questions (Q4 and Q5). For the cardinality finding task, the correct answers to the three cardinality questions (Q6, Q7, and Q8) were graded 2 points each, while the question that required inference (Q10) was given 4 points, a total of 10 points. In the identifier comprehension task, 2, 4, and 2 points were assigned to Q10, Q11, and Q12, respectively. In the database search task, 1 point was awarded for each correct combination of file name and data item.

An Analysis of Variance showed statistically significant differences ($F < 0.05$) between groups in the relationship existence finding and the relationship cardinality finding tasks. Unequal variance was assumed for all the response variables in the analysis. (The alpha level for the test of significance was set at 0.05 level.) Four demographic variables: GPA, Computer Ownership, DBMS experience, and Work experience were included as covariates (simultaneous inclusion) in the Analysis of Covariance to see if they would further reduce the variance. No statistically significant variance reduction was observed. The results are discussed in greater detail in the next section.

Table 1

Subject Background

SEX:	Male (67 %) Female (33 %)
AGE:	Range: 21 - 45 Average: 28
Computer Ownership:	Own (57%) Not Own (43%)
Software Experience:	Database Management Systems: 36% Graphics: 33% Accounting/Finance: 33% Spreadsheet: 83%
Work Experience:	Range: 0 - 20 yrs Average: 2 - 3 yrs
GPA:	Range: 2.70 - 3.90 Average: 3.40

DISCUSSION

Relationship existence finding task

Analysis of variance showed that there was a significant difference (F < 0.05) between groups in terms of the number of semantic relationships found between two data items (Table 2). More specifically, subjects using the semantic models (E-R, LDS) could identify more relationships than those using the relational models (T < 0.01). The result therefore supported our hypothesis H1a, which suggested that the E-R and LDS models would convey more semantics than the relational models. The RDM group performed less well than the other three groups (T < 0.01). (If the above two statements are made simultaneously, the overall level of significance become 0.15 based upon the Bonferroni procedure.)

With the relational models, the relationships identified were limited to only a few more than those that were explicitly represented in the model. However, no statistical difference was found between the DAD and the RDM, which indicates that factors other than the explicitness of representation are at work.

Relationship cardinality finding task

There was a significant difference (F < 0.05) between groups in finding the cardinality of relationships (H2). As

hypothesized (H2a), subjects using models with a graphical representation of cardinality (E-R, LDS) performed better than those with the textual representation (DAD, RDM) (T < 0.01). Again the RDM group performed significantly poorer than the other three (T < 0.05) (Overall alpha = 0.01).

Identifier comprehension task

Statistically significant differences were not found in the identifier comprehension task between models (F = 0.1107 in Table 2), hence, our H3 was not supported. However, the relational models (DAD, RDM) scored higher than the semantic models (E-R, LDS). Interestingly enough, the RDM group outperformed the other three in this task (T < 0.05). This result makes intuitive sense in that the relational model put more emphasis on attributes and identifiers than did the semantic models.

Database search task

No statistically significant difference was found between groups in this task (H4). The de-briefing questionnaires revealed that not enough time had been allowed for the first section. About half of the subjects were not able to complete the search task, the last part of section 1. As a consequence no conclusive statements can be made about the results of this task. There are, however, some indica-

tions that i) subjects had hard time grasping the concept of "join," and ii) no strong evidence was detected that the question format affected the search process (H5, H5a). There was no significant difference between the DAD group and the RDM group, although the former performed generally better (H4a).

Data modelling task

To avoid measurement errors, the evaluation criteria for this task was developed independently of the experimental models. This presented some difficulties to the investigators because each model is unique in its degree of specificity and because different criteria for "quality" exist for the different models. Two general criteria were initially adopted: the entity grouping and the relationship specification. Within the entity grouping, three aspects were examined:

1. Use of correct names for the entities or relations?

2. Attributes correctly grouped under the entities or relations?

3. Identifiers specified for the entities or relations?

The identifier specification was checked only for its presence, not for its accuracy, because subjects were not presumed to have knowledge of the functional dependency aspect of identifiers. For relationship specification, we examined whether entities were connected appropriately (for E-R and LDS) or whether relations were related using appropriate foreign keys (for DAD and RDM). Grading of the models based upon the criteria above revealed no statistically significant differences between groups, which suggests that the criteria may have been too superficial to detect any differences. When the criteria were adjusted for each individual model, we noticed an interesting gap in quality between the semantic models (E-R, LDS) and the relational models (DAD, RDM): most of the LDS and E-R models produced by the subjects were close to being correct and complete but many of the DAD and RDM models were inadequate. It appeared that subjects drawing DAD or RDM models did not follow a systematic modelling process such as i) identify entities, ii) identify attributes and identifiers of the entities, and iii) establish relationships between entities.

One indication of such an unsystematic approach to modelling was the location of the foreign keys within a relation. About half of the time, foreign keys appeared either between the attributes of a relation or at the beginning of a relation, which suggests that subjects' data item grouping process was mostly ad hoc. For about one third of the relationships shown, either relation-names or non-

id attribute names were used as foreign keys in relationship specifications (Table 3).

The LDS and E-R models produced by the subjects showed fewer signs of such problems. Since we had no reason to believe that the subjects in the semantic model groups were better modelers than those in the relational model groups, we suspect that some semantic models have the effect of hiding or making invisible the deficient modelling approaches of the subjects. The results also suggest that the relational models may not be adequate for the use as a first modelling tool.

A relationship can be specified in three ways in the relational model [ELMA80] (See Figure 5). Brosey and Shneiderman observed that people can better comprehend relationships if they are specified in a two-relation/ two-way fashion [BROS78]. We observed however that, as far as initial modelling was concerned, two-relation/ one-way was the dominant form (70%) of relationship specification.

Conclusions

In this experiment we have investigated contrasting representations of data in a narrowly defined systems analysis and design task. We have not addressed the value of specific data models in either determining or specifying system requirements but have restricted our investigation to understanding by potential system users. We found that the representation form does not affect understanding of the underlying reality, and that two different graphical models, E-R and LDS, promote model comprehension in the user knowledge domain.

Subjects were able to answer questions requiring comprehension of the representation of the important aspects of data models: entities, relationships, and attributes. In both the relationship and cardinality finding tasks, subjects presented with the semantic data models responded significantly more correctly than subjects using the relational models. This result indicates that a graphical network representation of entities and relationships is most appropriate for user validation.

The identifier comprehension task, in contrast, showed the relational data model significantly better than the other three. We think that the absence of explicit representation of relationships in the RDM facilitates recognition of identifier attributes. In contrast, results with the more graphical models imply that identifier attributes are harder to perceive when relationship information is explicitly represented.

Our experiment failed to adequately measure user ability to develop database search strategies, due to insufficient time allocated for this portion of the experiment. Addi-

Relationships between Employee (Employee Number as the identifier) and Department (Dept number as the identifier).

1) One-relation representation:

Dept-Employee

Dept Number	----	Employee Number	Employee Name

2) Two-relation/One-way representation:

Dept

Dept Number	---	Employee Number

One-way

Employee

Employee Number	Name	----

3) Two-relation/Two-way representation:

Dept

Dept Number	---	Employee Number

Two-way

Employee

Employee Number	----	Dept Number

4) 3-relation representation:

Dept

Dept Number	---

Dept-Employee

Dept Number	Employee Number

Employee

Employee Number	----

Figure 5

Relationship Representation in the Relational Model

Table 2

Mean Scores for Task Types

DATA MODEL / TASK TYPE	LDS	E/R	DAD	REL	ANOVA F
Relationship finding	4.43	5.38	3.25	2.71	0408
Relationship Cardinality finding	4.71	4.75	3.50	3.71	.0446
Identifier comprehension	2.39	2.63	3.25	3.71	.1107
Database Search	16.58	11.76	16.38	9.72	.2588

░░░ Lowest mean score for the task type.

tional research or replication of the experiment is indicated. We think the database search strategy task is an important one both for user validation and as an indication of the utility of alternative representations in more general uses of data models.

The final task in the experiment, data modelling, provided some interesting preliminary indications of subject comprehension, although no conclusive findings are reported here. Subjects using the semantical models E-R and LDS appeared to take a systematic modelling approach. This is in contrast to the rather haphazard approach indicated by the results from subjects using the RDM and DAD. We think this indication has important implications for data modelling beyond the user validation task. In particular, selection of appropriate data model representations for user-developed systems and for distributed microcomputer systems may strongly influence understanding and therefore correct use. Additional research is clearly called for in this area.

Finally, we note that the utility of a particular representation in the user validation task is expected to conflict with its utility in the discovery and specification tasks that are critical to successful systems development. Much more research is needed in these areas to identify the characteristics of data representations that will improve the science of systems development.

Table 3

Data Modelling Task

a) Entity Grouping:

		LDS	E/R	DAD	RDM
Use of an appropriate entity/relation name?	yes	94%	80%	85%	69%
	no	6%	20%	15%	31%
Attributes correctly placed?	yes	89%	80%	87%	83%
	no	11%	20%	13%	17%
Identifiers correctly specified?	yes	86%	54%	50%	49%
	no	14%	46%	50%	51%

Lowest performance for the task type

b) Relationship specification in DAD and RDM:

Foreign Key specification using;

Identifier	65%
Non-identifier	35%

Location of foreign key within relations;

At End	51%
Between Attributes	49%

Types of relationship specification used;

3-relation	2-relation 1-way	2-relation 2-way	1-relation
17%	70%	11%	2%

REFERENCES

[BENJ82] Benjamin, Robert, "Information Technology in the 1990's: A Long Range Planning Scenario," *MIS Quarterly*, June 1982.

[BROA78] Broadbent, D. E. and Broadbent, M. H. P., "The Allocation of Descriptor Terms by Individuals in a Simulated Retrieval System," *Ergonomics*, 21, 1978, pp. 343-354.

[BROS78] Brosey, Margaret, and Shneiderman, Ben, "Two Experimental Comparisons of Relational and Hierarchical Database Models," *International Journal of Man-Machine Studies*, Vol. 10, 1978.

[CHEN76] Chen, Peter P., "The Entity-Relationship Model—Toward a Unified View of Data," *ACM Transactions on Database Systems*, Volume 1, Number 1, March, 1976, pp. 9-36.

[CODD70] Codd, Edgar F., "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, Volume 13, Number 6, June, 1970, pp. 377-387.

[CODD79] Codd, Edgar F., "Extending the Database Relational Model to Capture More Meaning," *ACM Transactions on Database Systems*, Vol. 4, No. 4, December 1979, pp. 397-434.

[CODD82] Codd, Edgar F., "Relational Databases: A Practical Foundation for Productivity," *Communications of the ACM*, Vol. 25, No. 2, February 1982, pp. 109-117.

[DATE83] Date, Christopher J., *An Introduction to Database Systems*, Vol.II, Addison-Wesley, 1983.

[DICK82] Dickson, Gary W., "Requisite Functions for a Management Support Facility," in Sol (ed), .i "Processes and Tools for Decision Support," North-Holland, 1982.

[DURD77] Durding, Bruce M., Becker, Curtis A., and Gould, John D., "Data Organization," *Human Factors*, Vol. 19, No. 1, 1977.

[EDEL81] Edelman, Franz, "The Management of Information Resources—A Challenge for American Business" *MIS Quarterly*, Volume 5, Number 1, March, 1981, pp. 17-27.

[EHRE81] Ehrenreich, S. L. , "Query Languages: Design Recommendations Derived from the Human Factors Literature," *Human Factors*, Vol. 23, No. 6, 1981.

[ELMA80] El-Masri, Ramez, and Wiederhold, Gio, "Properties of Relationships and their Representation," *Proceedings of National Computer Conference*, 1980.

[HOFF84] Hoffer, Jeffrey A., "An Empirical Investigation into Individual Differences in Database Models," Working Paper, Dept. of Operations and Systems Management, School of Business, Indiana University, Bloomington, Indiana.

[KENT81] Kent, W., "Data Model Theory Meets a Practical Application," *Seventh International Conference on Very Large Databases*, Cannes, France, September 9-11, 1981, pp. 13-22.

[KVIN83] Kvinge, Heidi, and Carlis, John, "Introduction to Logical Data Structure," Working manuscript, Computer Science Dept., University of Minnesota, September, 1983.

[MCGE76] McGee, William C., "On User Criteria for Data Model Evaluation," *ACM Transactions on Database Systems*, Volume 1, Number 4, December 1976, pp. 370-387.

[MCLE79] McLean, Ephraim R., "End Users as Application Developers," *MIS Quarterly*, December 1979, pp. 37-46.

[MORA81] Moran, Thomas P., "An Applied Psychology of the User," *Computer Surveys*, Volume 13, Number 1, March 1981.

[POWE84] Powers, Michael J., Adams, David R., and Mills, Harlan D., *Computer Information Systems Development: Analysis and Design*, South-Western, 1984, pp. 426-456,

[REIS81] Reisner, Phyllis, "Human Factors Studies of Database Query Languages: A survey and Assessment," *Computing Surveys*, Volume 13, Number 1, March 1981.

[ROCK81] Rockart, John F., and Flannery, L. S., "The Management of End User Computing," *Proceedings of Second International Conference on Information Systems*, December 1981, pp. 351-363.

[SHNE78] Shneiderman, B., "Improving the Human Factors Aspect of Database Interactions," *ACM Transactions on Database Systems*, Volume 3, Number 4, December 1978, pp. 417-439.

[TSIC82] Tsichritzis, Dionysios C., and Lochovsky, Frederick H., *Data Models*, Prentice-Hall, 1982.