The Quality of Data Representations Developed by NonExpert Designers: An Experimental Study

Dinesh Batra
Florida International University

Peeter J. Kirs
Florida International University

A laboratory study was conducted to compare the quality of conceptual data representations in the relational form developed by non-experts using the data aggregation (DA) approach and the logical relational design methodology (LRDM). While significant differences were not found between the quality of the relational solutions developed using the two techniques, differences were noted with respect to the intermediate diagrams produced. Furthermore, subjects using the LRDM developed satisfactory Entity-Relationship (ER) solutions, but were not able to appropriately translate the ER solutions into the relational form. Subjects using the DA approach developed poor quality representations both in the data aggregation and the relational form. The findings from this study suggest that conceptual data modeling performance can be improved by using design aids which can perform the translation of ER diagrams into relations.

Conceptual database models are higher level representations of data sets and the relationships between them, and are intended to be independent of any specific database management system (DBMS) (McFadden and Hoffer, 1991). A number of man-machine studies have been conducted to test the useability of data models in a conceptual data modeling task (Brosey and Shneiderman, 1978; Juhn and Naumann, 1985; Ridjanovic, 1986; see Batra and Srinivasan, 1992, for a detailed survey). While the target of these studies generally has been novice and nonexpert designers, the implications are also applicable to end users. Given the rapid increase in end user developed (EUD) applications (Munro et al., 1987-88; Benson, 1983; Panko, 1987; Rivard and Huff, 1984; Wetherbe and Leitheiser, 1985) and the subsequent concerns about their potential risks to organizations (Alavi and Weiss, 1986; Davis, 1989), the utility of conceptual database modeling approaches have taken on increased importance. There are implicit and explicit demands made upon end users to effectively use database development tools, especially for applications requiring a fair degree of developer sophistication.

To facilitate the conceptual data modeling task, data models, which effectively capture application semantics, have been proposed. The most commonly applied data model, the entity relationship (ER) model (Chen, 1975), has been empirically shown to be an effective tool for conceptual data modeling (Batra, Hoffer and Bostrom, 1990). As an extension of the model, the Logical Relational Design Methodology (LRDM), suggested by Teorey, Yang and Fry (1986), involves developing an ER representation which can then be translated to a relational representation.

An alternative data modeling approach, data aggregation (DA) has been proposed by Smith and Smith (1977a and 1977b). The DA model relies on the concept of aggregate objects, which themselves may be composites of attributes, to develop a logical design of relational databases. While the DA approach has not received the same attention as LRDM, and has essentially remained untested, the concept of aggregation is frequently cited in the object oriented database literature (Bertino and Martino, 1991).

This paper reports a study which compared the effi-
cacy of the ER and DA models as structures to guide translation into relational database structures. In accordance with the emphasis on end-user developed applications, the study used nonexperts engaged in a conceptual data modeling task.

**Problem Background**

With respect to database development, a few trends seem evident. Relational model based DBMS, especially on microcomputers and Local Area Networks (LANs), have become very popular. These DBMS (e.g., DBASE, FoxPro, Paradox, and Rbase) provide powerful database query languages (e.g., SQL and QBE) which allow non-procedural access for ad-hoc requests. However, while the data manipulation components of DBMS have been readily received by nonexpert users, the basic conceptual design principles of the relational model still appear to present considerable difficulties for this class of users (Batra, Hoffer and Bostrom, 1990).

Traditionally, relational database design has relied on low-level, bottom-up approaches intended to construct normalized relations using inter-data element dependencies (Codd, 1970; Smith, 1985; Maier, 1988; Date, 1990). However, several authors (Chen, 1976; Smith and Smith, 1977a; Teorey et al., 1986; Teorey et al., 1989) have observed that as the scale of the database expands, and the number and complexity of relationships increases, the overall structure can become obscured to even experienced analysts. In response, top-down conceptual data modeling has been suggested and examined as a means of increasing problem understanding, communication of requirements, and as a framework for transforming component elements into normalized relations (Beeri et al., 1978; Kent, 1981; Teorey et al., 1986; Blaha et al., 1988; Smith and Smith, 1977a). The implication is that top-down approaches will ease conceptual data modeling, especially for nonexpert designers. There is empirical evidence that some data models, specifically the ER model, do lead to better user performance in conceptual modeling (Batra et al., 1990). Usability studies contrasting such approaches thus merit further investigation.

Conceptual database methodologies are based on semantic data models (Hull and King, 1987; Peckham and Maryanski, 1988) that have been proposed for the explicit purpose of typifying structures and linkages in a simple and semantically appealing manner. Examples of such data models include the ER model as well as the extended entity relationship (EER) model (Chen, 1976; Elmasri et al., 1985; Yao, 1985; Teorey et al., 1986), the Nijssen Information Analysis Model (NIAM) (Verheijen and van Bekkum, 1982; Nijssen and Halpin, 1989), and the Semantic Data Model (Hammer and McLeod, 1981).

The LRDM approach is based on the ER model. In developing an ER representation, the analyst/designer identifies the relevant entities (persons, places, objects or concepts about which data is to be stored), determines the degree of each relationship (how many entities are involved) and the connectivity (how many instances of one entity are related to instance of another entity), and graphically maps the logical structures using prescribed symbols. Consider, for example, a simple hotel reservation system. A person reserves a room for certain occupancy dates. In the ER model, reservation may be expressed as the relationship between the entities person, room and occupancy_date. The connectivity of the associations, which in this case is one-to-many-to-many can be clearly mapped (see Figure 1.a). If the problem specification were to change (e.g., a person were allowed to reserve many rooms on a given day), the mapping could be readily modified to accurately represent the relationship.

Aggregation (Smith and Smith 1977a) facilitates abstraction by allowing a relationship between objects to be viewed as a higher level object. For example, a reservation may be depicted as an aggregate of the simple objects person, room and occupancy_date, or as an aggregate of person and assignment, where assignment is an aggregate of room and occupancy_date. Data aggregation also appears intuitively appealing since aggregation is a construct applied in everyday usage. In the above example, one can talk about a reservation (aggregate object) without mentioning the details (simple objects) that constitute it (see Figure 1.b.).

Note that the concepts inherent in the DA modeling approach (Smith and Smith, 1977a) are, in many respects, similar to the fundamental notion of the LRDM. Both methods offer diagraming conventions for modeling entities (or, in DA terminology, objects) and the relationships between them. The recent popularity of object oriented approaches, which are also based on the concept of data aggregation (Bertino and Martino, 1991), indicates that further investigation of the DA approach is warranted.

This study has two primary concerns: (1) which methodology (LRDM or DA approach) leads to better user performance in developing the relational representation of a given problem, and (2) to what extent errors are introduced...