Human Factors Research on Data Modeling: A Review of Prior Research, An Extended Framework and Future Research Directions

HEIKKI TOPI, Bentley College, USA
V. RAMESH, Indiana University, USA

This study reviews and synthesizes human factors research on conceptual data modeling. In addition to analyzing the variables used in earlier studies and summarizing the results of this stream of research, we propose a new framework to help with future efforts in this area. The study finds that prior research has focused on issues that are relevant when conceptual models are used for communication between systems analysts and developers (Analyst – Developer models) whereas the issues important for models that are used to facilitate communication between analysts and users (User – Analyst models) have received little attention and, hence, require a significantly stronger role in future research. In addition, we emphasize the importance of building a strong theoretical foundation and using it to guide future empirical work in this area.

INTRODUCTION
Conceptual data modeling continues to be an integral part of the foundation on which information systems are built. Depending on the development methodologies that are used for a particular project, the terms and methods used for conceptual data modeling vary, but in practice, a clear majority of methodologies used for systems development include a set of tools and methods for modeling data at the conceptual level. Therefore, it is not surprising that research in IS and its reference disciplines has shown a significant interest in various aspects of data modeling for the past 20 years. The focus of this paper is on research that examines the usability of various conceptual data modeling approaches, i.e., research that investigates human factors issues in conceptual data modeling. We review and analyze this literature and suggest several new directions for further research.

BACKGROUND
The concept of data modeling has been used with a variety of different meanings within various areas of study and practice. However, within the organizational context the core idea underlying all the definitions is the same: A data model is used for describing entities1 and their relationships within a real world domain. For example, McFadden, Hoffer, and Prescott (1999) define a data model as “an abstract representation of the data about entities, events, activities, and their associations within an organization.” A data model is an abstraction and a simplification of the domain it describes and thus, it always represents a limited part of reality.

The main focus of this paper, conceptual data modeling, requires further clarification. Based on the ANSI/SPARC definition, a conceptual data model is any model that is independent of the underlying hardware and software. This means that using this definition, models created using formalisms ranging from the relational model to the semantically rich variants (Teorey, Yang, & Fry, 1986) of Entity-Relationship modeling (Chen, 1976; Hull & King, 1987) can be considered to be at the conceptual level. A more restrictive definition of a conceptual model can be found in Batra and Davis (1992). They define a conceptual model as one that is capable of capturing the structure of the database along with the semantic constraints into a model that is easy to under-
stand, does not contain implementation details, and can be used to communicate with users. A key criteria in the above definition is the independence of modeling from the implementation technology. This means that in order to be categorized as a conceptual model, the representation must not be dependent on the characteristics of the database technologies available (e.g., relational, object-oriented, object-relational, network, or hierarchical).

We believe that both of the definitions presented above are, however, somewhat misleading because a true conceptual data model should capture the essential data characteristics of the domain of interest, and not necessarily the structure of the database. Thus, we define a conceptual data model as a set of constructs that can be used to create an abstraction of reality, i.e., a representation that is capable of capturing the data oriented (as opposed to process oriented) aspects of a domain of interest in a manner that is unambiguous and easy to understand for both designers and users alike. Note that this definition does not have any references to a database structure. This is because we believe that not everything captured in a representation created using a conceptual data model will (or needs to) be reflected in a database or the eventual system being developed.

Based on the above definition of conceptual data modeling, one can synthesize at least four different uses for a conceptual data model (Batra, Hoffer, & Bostrom, 1990; Cambell, 1992; Juhn & Naumann, 1985): 1) a communication tool between analysts and users for the discovery (elicitation and representation) and validation stages of the systems analysis process, 2) a formal conceptual foundation for organizational information systems at various levels (a common accepted model of reality and a communication tool between IS professionals, e.g., analysts and developers), 3) a foundation for applications developed by end users, and 4) an essential part of the system documentation for the maintenance of the system.

The main focus of this paper is to examine research on the human factors issues in data modeling, i.e., research that employs social science methods such as laboratory experiments to evaluate and improve the usability of the systems. Batra and Srinivasan define usability as “the ability of the user to represent a problem in a computing environment and effectively work with that representation” (1992, p. 395). Thus, two important research questions of human factors research on data modeling have traditionally been as follows: 1) how do the characteristics of the available tools affect users’ ability to succeed in their tasks (i.e., what is the level of usability of the tools)?, and 2) how satisfied are the users with the tools?

**REVIEW OF PRIOR RESEARCH**

In this section, we review the previous human factors research on data modeling. This review is based on a careful analysis of existing studies published in academic journals or in the Proceedings of the ICIS conference that have empirically evaluated some aspect of the usability of conceptual data modeling tools and methods. After a comprehensive search, we identified 27 articles published after (and including) Brosey & Shneiderman’s (1978) early work in 1978. A summary table of these studies is presented in Appendix A. The table includes a description of the independent variables (IV), dependent variables (DV), research tasks, and the most important results.

First, we will discuss the typical research variables used in these studies, and then, review the most important empirical findings.

**Variables of Interest in Empirical Studies**

**Research framework.** Figure 1 includes a schematic representation of the research framework that has been used either explicitly (as by Batra et al., 1990) or implicitly in many of the earlier studies. Human refers to the individual level factors related to the characteristics of the individuals who perform the data modeling tasks. Data Model is used in this context to describe the differences between the data modeling formalisms, and Task refers to the characteristics of the tasks of interest related to data models, such as model creation, comprehension, or validation. The model indicates a reciprocal relationship between Human, Data Model, and Task, which all, in turn, have an impact on the quality of the resulting data model, i.e., (human) Performance in the data modeling task. Variables in the Human, Data Model, and Task categories have been used in earlier studies as independent and control variables, as indicated in the discussion below, and Performance is a natural dependent variable in the studies.

**Independent variables.** The most frequently used independent variable in the earlier studies has been the data modeling approach or data model, as it is called by, for example, Batra and Davis (1992) and Navathe (1992) and in the research framework in Figure 1. In early research, Brosey and Shneiderman (1978) compared hierarchical and relational data models, whereas several later studies have compared different types of semantic and relational data models (Amer, 1993; Batra & Antony, 1994; Batra et al., 1990;