Human Factors Reseach 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 entities[1] 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[2] that have empirically evaluated some aspect of the usability of conceptual data modeling tools and methods[3]. 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.

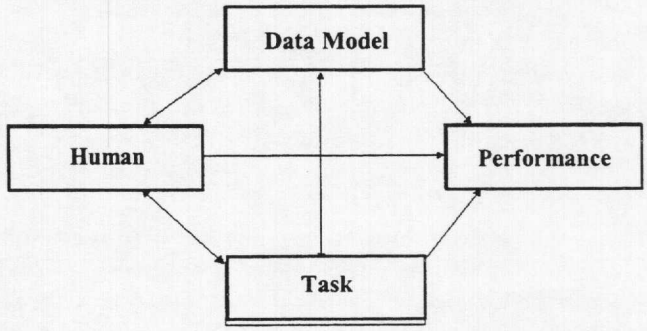
First, we will discuss the typical research variables used in these studies, and then, review the most important empirical findings.

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;

Figure 1: Widely used framework for human factors research on data modeling (see, for example, Batra et al., 1990)



Jarvenpaa & Machesky, 1989; Juhn & Naumann, 1985; Liao & Palvia, 2000; Sinha & Vessey, 1999) and/or two different semantic data models (Kim & March, 1995; Lee & Choi, 1998; Liao & Palvia, 2000; Nordbotten & Crosby, 1999; Palvia, Liao, & To, 1992). Several of the most recent studies have compared semantic data models to object-oriented data models (Bock & Ryan, 1993; Hardgrave & Dalal, 1995; Lee & Choi, 1998; Liao & Palvia, 2000; Palvia et al., 1992; Shoval & Frumermann, 1994; Shoval & Shiran, 1997; Sinha & Vessey, 1999).

The next category of independent variables consists of *user characteristics* (*Human* in the research framework in Figure 1). The most commonly used independent variable is experience: The level of *general* MIS or programming experience was used as an independent variable in studies by Brosey and Shneiderman (1978) and Hoffer (1982), whereas Batra and Davis (1992), Weber (1996), and Lee and Choi (1998) analyzed the differences between subjects with various levels of data modeling experience. Ramesh and Browne (1999) differentiated between "database-knowledgeable" and "database novice" based on the subjects' understanding of basic ER concepts. Agarwal, Sinha, and Tanniru (1996) investigated the impact of the type of design experience on modelers' ability to use different formalisms for different tasks. In addition to programming expertise, Hoffer (1982) studied the effects of cognitive style, another category of individual differences.

A set of *task characteristics* (*Task* in the research framework in Figure 1) has also been used as independent variables in the studies: Brosey and Shneiderman (1978) manipulated the *task type* (comprehension, problem solving, memorization), as did Batra and Antony (2001) (task's compatibility with a support tool). Hoffer (1982) varied the description of the situation on which the data model was based so that the situation was either related to a specific task or to the entire organization. *Task complexity* was used as an independent variable in Shoval and Even-Chaime (1987), Hardgrave and Dalal (1995), Weber (1996), and Liao and Palvia (2000). Jarvenpaa and Machesky (1989) investigated the effects of learning by using a within-subjects design and administering four data modeling tasks to each subject.

Dependent variables. The dependent variables can be divided into two broad categories: user performance and user attitudes. As seen earlier, the two main research questions of this area are related to modeling performance and user satisfaction, and therefore, the widespread use of these dependent variables is understandable.

Performance has been divided into three subcategories: *model correctness* (also referred to as *procedural or skill knowledge* of the user by Jarvenpaa and Machesky (1989) measured by the characteristics of the end result of the modeling process), *time* used to create the solution, and *declarative knowledge* (understanding of the notation (Jarvenpaa & Machesky, 1989)). In most cases, the correctness of the model has been measured with the degree to which it corresponds to a predefined "correct" solution. Batra et al. (1990) were the first to refine the concept of correctness by measuring the correctness of various *facets* or structural elements of the model (entities, identifiers, descriptors, categories, and five different types of relationships: unary, binary one-to-many (1:M), binary many-to-many (M:N), ternary one-to-many-to-many (1:M:N), and ternary many-to-many-to-many (M:N:O)). The same facet structure was used later by Bock and Ryan (1993), Shoval and Shiran (1997), Lee and Choi (1998), and Liao and Palvia (2000). Kim and March (1995) divided the analysis of model correctness into *syntactic* and *semantic* categories: Syntactic correctness refers to users' ability to understand and use the constructs of the modeling formalism, whereas semantic correctness is the extent to which the data model corresponds to the underlying semantics of the problem domain. Another widely used measure of performance has been the time it takes to finish a modeling or model comprehension task (Hardgrave & Dalal, 1995; Jarvenpaa & Machesky, 1989; Lee & Choi, 1998; Liao & Palvia, 2000; Palvia et al., 1992; Shoval & Even-Chaime, 1987; Shoval & Shiran, 1997).

The *user attitudes* measured within this area of research are confidence (Hoffer, 1982), preference to use a certain model (Shoval & Even-Chaime, 1987; Shoval & Shiran, 1997), perceived value of the modeling formalism (Kim & March, 1995), and perceived ease-of-use (Batra et al., 1990; Hardgrave & Dalal, 1995; Kim & March, 1995).

In a study in which the dependent variable does not belong to either one of the main categories, *data model characteristics* were the main point of interest for Hoffer (1982). His study focused on the nature of the data models which the subjects created when they were able to freely choose the way to describe a structure of a database. The two characteristics of the model in his study were "image architecture" and "image size", i.e., the modeling approach chosen and the number of entities.

Identified control variables. By investigating the nature of the explicitly identified control variables in previous research, it is possible to find potential independent variables of interest for future research, as well as summarize the variables that have to be controlled in future studies. *User characteristics* (*Human* in the framework in Figure 1) is the first category of specific control variables in the earlier studies. The most common individual variable in the user characteristics category is *experience*. The most common types of experience discussed in prior research are general work experience (Batra et al., 1990; Batra & Kirs, 1993; Jarvenpaa & Machesky, 1989; Juhn & Naumann, 1985; Liao & Palvia, 2000), general computer/IS experience (Batra et al., 1990; Batra & Kirs, 1993; Jarvenpaa & Machesky, 1989; Juhn & Naumann, 1985; Liao & Palvia, 2000), and database experience (Batra et al., 1990; Batra & Kirs, 1993; Jarvenpaa & Machesky, 1989; Juhn & Naumann, 1985; Liao & Palvia,

2000). Age (Liao & Palvia, 2000), education (Jarvenpaa & Machesky, 1989; Liao & Palvia, 2000), intellectual ability (Juhn & Naumann, 1985), and cognitive style measured with LSI (Jarvenpaa & Machesky, 1989) have been other types of individual differences which have been controlled. In most studies, user characteristics have been controlled by selecting subjects from a homogenous population and by random assignment to experimental conditions.

Controlling for *task characteristics* (*Task* in the framework in Figure 1) by keeping them the same across the treatments is a natural approach and not very interesting at the category level. Jarvenpaa and Machesky (1989) and Batra and Kirs (1993) both list specific characteristics of the task which were kept constant; these were complexity, structure, difficulty, and time, which are all related to a more general concept of difficulty. Kim and March (1995) specifically mentioned task complexity and time as task characteristics that were controlled. *Training* was also identified as a significant control variable by Batra and Kirs (1993) and Kim and March (1995); details controlled in these experiments include trainer characteristics and instructional examples. Table 1 summarizes the variables used in prior research.

Table 1: Variables identified in human factors research on conceptual data modeling

Variable Type	Variable Category	Representative Examples
Independent variables	Data modeling formalism (Data Model)	Hierarchical vs. relational Relational vs. semantic Semantic vs. semantic Semantic vs. object-oriented
	User characteristics (Human)	General MIS experience Programming experience Data modeling experience Other modeling experience Cognitive style
	Task characteristics (Task)	Task type Task complexity
Dependent variables	Performance	Model correctness (facets, syntactic vs. semantic) Time Knowledge of the formalism
	User attitudes	Confidence Preference Perceived value Ease-of-use
Control variables	User characteristics (Human)	Work experience General IS/computer experience Database experience Age Education Intellectual ability Cognitive style
	Task characteristics	Complexity Time Structure Difficulty

Key Findings from Prior Studies

The results from the empirical studies reviewed can be categorized as follows: a) Effects of data modeling formalism on user performance and attitudes; b) Effects of user characteristics on user performance and attitudes; and c) Effects of task characteristics on user performance and attitudes. Most of the studies have focused on the first category. In addition to the associations between research variables, we will review the results for various task components (facets) and the main lessons from the studies with a process focus.

Effects of data modeling formalism on user performance and attitudes. The studies that have investigated the effects of the data modeling formalism on performance and attitudes can be divided into the following subcategories: a) those comparing a semantic model to the relational model; b) those comparing two semantic models to each other; and c) those comparing a semantic model with object-oriented models.

In the first subcategory, the seven studies (Amer, 1993; Batra & Antony, 1994; Batra et al., 1990; Jarvenpaa & Machesky, 1989; Juhn & Naumann, 1985; Liao & Palvia, 2000; Sinha & Vessey, 1999) that have investigated the differences between the ER/EER and relational modeling formalisms have all found support for the positive effect of the use of the ER/EER model on one or several aspects of modeling performance. The studies provide strongest support to ER/EER's advantage in modeling binary 1:M and binary M:N relationships; four of the studies (Amer, 1993; Batra et al., 1990; Liao & Palvia, 2000; Sinha & Vessey, 1999) support this finding, whereas the other findings–related to the identification of relationships and cardinalities, faster learning, understanding the notation, modeling ternary 1:M:N and unary relationships, and generalization modeling–are all based on only one of the studies. For the binary relationships, these results are in line with those of Cao, Nah, and Siau's (2000) meta-analysis, which included both modeling and query writing studies; our analysis did not find the strong support for ER/EER's advantage over relational model in modeling ternary 1:M:N relationships. The one study (Shoval

& Even-Chaime, 1987) that focused on the relationship between the relational model and a non-ER semantic model, NIAM, found the relational model to lead to better user performance and to require less time. As to the effects of the modeling formalism choice between semantic and relational models and the user attitudes, the results are scarce and inconclusive: Jarvenpaa and Machesky (1989) found that subjects perceived the ER/EER model to be easier to use than the relational model, but Shoval and Even-Chaime (1987) found that the subjects preferred the relational model over NIAM.

Research focusing on two semantic models has in most cases compared the ER/EER model with other semantic models. The two studies (Kim & March, 1995; Lee & Choi, 1998) that compared the ER/EER model with ORM/NIAM both found support to the claim that the use of ER/EER leads to better user performance in expressing the meaning of the problem domain. Weber's (1996) results regarding the natural human tendency to separate entities and attributes provide evidence to support this finding. In addition, in both studies the subjects found ER/EER to be easier to use than ORM/NIAM. Bock and Ryan (1993) and Lee and Choi (1998) compare the ER/EER formalism with another semantic model, Semantic Object Model (SOM), and both studies suggest that the use of ER/EER leads to better performance, although the results are not fully conclusive regarding the facets of modeling that ER/EER supports better. In Lee and Choi (1998), the novice subjects found ER/EER to be easier to use than SOM.

Six studies (Hardgrave & Dalal, 1995; Lee & Choi, 1998; Liao & Palvia, 2000; Shoval & Frumermann, 1994; Shoval & Shiran, 1997; Sinha & Vessey, 1999) have investigated the effects of the choice between object-oriented models (although not consistently the same one) and ER/EER. The lack of consistency between the studies makes it difficult to draw any general conclusions, but the direction of the studies seems to suggest that using the ER/EER model leads to better performance in modeling tasks. The studies together indicate that the use of ER/EER has a positive effect on modeling performance in five of the modeling facets (unary 1:1, binary 1:1 and 1:M, and ternary 1:M:N, and M:N:O), but, unfortunately, the findings come from different studies that do not provide support for each other's findings. The only result related to user attitudes in these studies was made by Shoval and Shiran (1997), who found that ER/EER users' quality perceptions were higher than those of OO users.

Effects of user characteristics on performance and attitudes. Six empirical studies have significant results regarding the effects of user characteristics on performance and attitudes, and all of them have focused on some type of task-related experience. The results do not, unfortunately, build a highly consistent image because every study has investigated a different aspect of experience. Therefore, the studies will be discussed here in chronological order. Batra and Davis (1992) confirmed that well-known process differences between novices and experts can also be observed within this domain. According to Agarwal et al. (1996), subjects with experience in modeling with a process focus are able to utilize this experience when they are modeling behavior but not with data structures. Weber's (1996) results in his experiment using a recall task suggest that although NIAM experts' ability to recall model elements was slightly better than that of novices, their memory structures and recall strategies were the same. Lee and Choi's (1998) results regarding the differences between experienced ER modelers and novices are somewhat difficult to interpret, but it appears that in most respects ER experience led to higher performance with the other methods, too, although experienced modelers used more time. In all cases but one (ORM), experienced ER modelers perceived the methods to be easier to use than inexperienced modelers did. According to Ramesh and Browne (1999), "database-naive" subjects were better able to express causal relationships than "database-knowledgeable" subjects, and they attribute this to the inability of commonly used modeling formalisms to support the expression of causal relationships. Finally, Burton-Jones and Weber (1999) studied the effects of domain knowledge and ontological clarity of a representation on the subjects' ability to answer problem-solving questions. Their results provide limited support to the claim that ontological clarity is particularly important in cases when domain knowledge is low.

Effects of task characteristics on user performance and attitudes. None of the studies have directly focused on the effects of task characteristics on the main dependent variables, although four of them (Hardgrave & Dalal, 1995; Liao & Palvia, 2000; Shoval & Even-Chaime, 1987; Weber, 1996) used task complexity as an independent variable and all of them found a main effect for complexity on performance (in practice, this means that the experimental manipulation worked). This is understandable because in most cases the focus is on the moderating effects of task characteristics on the effects of other variables on performance, particularly the model formalism and user characteristics.

Differences between facets. As discussed above, most of the studies have used some version of the facet structure for analyzing user performance since Batra et al. (1990) originally presented it. Five of them have analyzed user performance in one or several of these facets with measures that are similar to each other and give us an opportunity to review users' relative performance with various facets. The performance data per facet from these studies is included in Table 2; no aggregate data is presented here because it is not in all cases clear whether or not the methods have been similar enough to justify the use of composite measures. This data does, however, lead to the following observations: 1) Identifying and modeling ternary relationships correctly is difficult for novice users, and even in the relatively simple experimental tasks users' average performance level is often below 50%.

Table 2: User modeling performance by Facet in empirical studies

	Batra et al., 1990		Batra & Kirs, 1993		Bock & Ryan, 1993		Shoval & Shiran, 1997		Palvia & Liao, 2000		
	Rel.	ER/EER	Rel.	ER/EER	ER/EER	OO	ER/EER	OO	Rel.	ER/EER	OO
Entity					98.0	96.0	99.0	99.0			
Identifier	72.4	73.9			96.0	80.0			62.8	69.7	77.3
Descriptor							95.0	94.0			
Category					92.0	82.0	99.0	99.0			
Unary	68.3	55.2			96.0	64.0	88.0	70.0	59.9	40.0	50.0
Binary 1:M	54.4	84.9	50.6	81.2	89.0	88.0	83.0	89.0	54.2	83.8	73.9
Binary M:N	57.1	92.9	67.5	92.5	100.0	63.0	81.0	79.0	41.2	74.4	65.3
Ternary 1:M:N	8.3	41.3	46.9	60.0	47.0	44.0	85.0	68.0			
Ternary M:N:O	33.3	45.2	40.6	45.6	79.0	72.0	94.0	76.0	35.4	57.5	47.7

The range of performance levels is, however, very large varying from 8.3% for 1:M:N relationships in Batra et al. (1990) to 94% for M:N:O relationships in Shoval and Shiran (1997). 2) Results are weak (below 70%) also for unary relationships, except with a semantic formalism (ER/EER) in Bock and Ryan (1993) and Shoval and Shiran (1997). The range is also large with this facet (from 40% to 96%). 3) With semantic and object-oriented modeling formalisms, users' average performance in modeling the binary relationships is consistently at a high level (above 80%), with the exception of binary M:N relationships in Liao and Palvia (2000). 4) Modeling identifiers, a seemingly simple task, appears to cause difficulties with all modeling formalisms with typical performance levels around 70%.

Other findings. Five of the studies included in this review analyzed some aspect of the process that subjects followed while creating a data model. As discussed earlier, Jarvenpaa and Machesky (1989) investigated whether the subjects chose a top-down or a bottom-up approach when constructing data models and whether the choice of the approach was dependent on the modeling formalism. They found that users of the ER based Logical Data Structure model were more likely to use a top-down approach than the user of the relational model. Batra and Davis (1992) studied the protocol differences between novices and experienced data modelers and found broad support for several findings from prior research regarding the differences between these two groups: experts had richer concept vocabulary and were better able to categorize constructs and automate processes, whereas novices were more likely to make a range of modeling errors. Batra and Sein (1994) analyzed at the individual level users' ability to improve the quality of their data modeling solutions based on feedback and found out that feedback can help users avoid errors in modeling ternary relationships. Srinivasan and Te'eni (1995) focused entirely on the results of the process analysis of a specific modeling behavior. Using verbalized protocols, they analyzed the use of several heuristics at various levels of abstraction to manage the complexity of the data modeling process. The most important results reported in Srinivasan and Te'eni (1995) were that efficient data modelers use specific heuristics to reduce the complexity of the problem, test models at regular intervals, and make orderly transitions from one level of abstraction of problem representation to another. In general, the study provides an important example of a research approach that makes it possible to evaluate data modeling at a detailed level as a problem solving process. Weber (1996) utilized a strong theoretical foundation in cognitive psychology and philosophy to evaluate whether or not humans tend to see entities and attributes as distinct constructs, and his conclusion based on a memory recall experiment is that these, indeed, are separate elements. Building on an important line of research, Batra and Antony (2001) investigated the effectiveness of a consulting system that is designed to reduce data modeling errors and found out that particularly individuals with a low initial knowledge level benefited from the consulting system.

Having reviewed the results of prior usability research on conceptual data modeling, we continue by evaluating the implications of these results and suggesting several new avenues for future research.

POTENTIAL FOR FUTURE RESEARCH

Given the maturity of data modeling in practice and the results summarized above, it would be easy to conclude that further human factors research related to conceptual data modeling may not add substantially to the existing body of knowledge. In the next section we hope, however, to demonstrate that because it has focused on a relatively narrow part of conceptual data modeling, prior research has left several potentially important questions still unanswered.

Most of the empirical studies reported above that have

investigated conceptual data modeling from the human factors perspective are based on the same relatively simple model: in a controlled laboratory study, subjects with relatively little experience complete one or several modeling tasks in which they create a graphical representation of an organizational situation based on a narrative using one or several conceptual data modeling formalisms. The results are typically evaluated by grading the models using a solution created by the researcher as a baseline; results achieved with different formalisms are then compared to each other with standard statistical techniques. This approach definitely has improved our understanding of the factors that affect subjects' ability to represent a case situation with graphical tools, and a controlled experiment is a perfectly valid methodology for investigating specific aspects of a cognitively complex task such as conceptual data modeling.

We present three key ideas that can help with future research efforts:

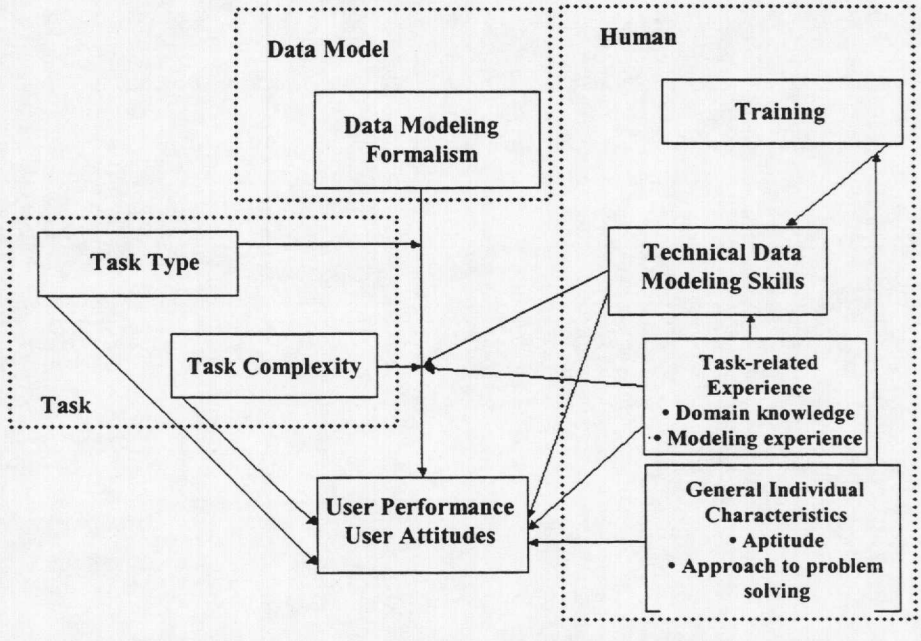
a) First, we note that because almost all of the research to date has focused on the technical characteristics of the modeling formalisms, we know very little about the effects of users' individual characteristics, task characteristics, or the interaction between the modeling formalism, user, and task. Below we discuss a new framework that we hope will provide additional clarity to future research efforts.
b) Second, we demonstrate how we can open new directions for the research in this area by investigating additional uses for conceptual data modeling.
c) Finally, we observe that we do not yet have a good understanding of why certain formalisms work well in some situations and not in others; the mechanisms mediating the relationships between the main research variables are not clear. We provide several suggestions for research that can be used to strengthen our understanding in this area.

An Expanded Framework for Human Factors Research in Data Modeling

Our review of prior literature and additional conceptual analysis of this stream of research leads us to believe that the traditional framework that has been used to guide human factors research on data modeling (see Figure 1 above) can be improved and clarified. In this section, we present and justify the suggested changes, which have been incorporated into a new framework presented in Figure 2. We supplement the framework in Figure 2 by using the findings from the studies reviewed earlier as well as our theoretical understanding of the domain. However, it is worth noting that the theoretical basis for this expanded framework as well as the Batra et al. (1990) framework lies in the classical general MIS task – technology – human research framework, which, in turn, is a derivation of Leavitt's (1965) organizational system model.

As we have seen in the review of prior literature and summary of the results above and will discuss below, many of the relevant relationships are between specific components of the framework elements (see also Table 1). Hence, it is important to elaborate on the broad construct categories Task, Data Model, Human, and Performance. *Task Complexity* and *Task Type* should be presented as separate concepts, because these dimensions of the task are largely independent and their effects should be investigated separately from each other. For example, it is understandably possible to have various levels of complexity for comprehension, validation, and modeling tasks and both could be used separately as independent variables in the same study at the same time. As to *Human*, we can differentiate between multiple categories of individual characteristics, which are independent from each other. Underlying all other aspects of an individual's performance are *general individual characteristics* such as intelligence, cognitive style, and problem-solving approach, which affect a particular individual's performance in all cognitive tasks. The only data modeling study so far that has explicitly used a variable from this category is Hoffer (1982). An individual also has *experience* in a variety of areas, many of which are potentially relevant to their performance in the task of interest (general problem-solving experience, programming experience, general modeling experience, modeling experience

Figure 2: Proposed framework for human factors research on data modeling



with specific formalism(s), etc.). This category of variables has been utilized widely in earlier research, as discussed in the review above (Agarwal et al., 1996; Batra & Srinivasan, 1992; Brosey & Shneiderman, 1978; Burton-Jones & Weber, 1999; Hoffer, 1982; Lee & Choi, 1998; Ramesh & Browne, 1999; Weber, 1996). Finally, an individual's *technical skills* in the use of a specific data modeling formalism should be conceptually separated as a factor affecting user's performance. One of the reasons why it is essential to differentiate technical skills from other aspects of individual differences is that this is the only category of these that can be affected by *training* (other factors that could be influenced by training include confidence, self-efficacy, task motivation, etc). Technical skills have been used as an independent variable in several studies (Batra & Antony, 2001; Weber, 1996). In general, the division of the framework elements into components forces us to specify the nature of the relationships of interest at a significantly more detailed level. This, in turn, will lead us closer to true theoretical models at least in part based on applicable theories from relevant reference disciplines, such as Anderson's ACT theory with its variants (Anderson, 1993), which was suggested as an important theoretical basis for research on information modeling (including conceptual data modeling) by Siau (1999).

Second, the framework should incorporate two different types of dependent variables to acknowledge the fact that we are not only interested in objective performance but also users' attitudes towards the tools, the tasks, and their own performance. The most often used non-performance dependent variables are ease-of-use perceptions (Batra et al., 1990; Hardgrave & Dalal, 1995; Kim & March, 1995; Lee & Choi, 1998) and modeling formalism preference (Batra & Sein, 1994; Kim & March, 1995; Shoval & Even-Chaime, 1987; Shoval & Shiran, 1997).

Third, the framework should acknowledge and explicitly incorporate the potentially complex moderating effects of other variables on the relationship between the data modeling formalism and user performance and attitudes. The direct effect of task complexity on the dependent variables, particularly performance, is seldom the main point of interest; in most cases, we are interested in the way different formalisms support users at various task complexity levels. The same is true with task type: a relevant research question is the suitability of various modeling formalisms for specific task types and thus, we should explicitly express in our research model that task type moderates the relationship between the data modeling formalism and the dependent variables. The best examples of this are the experiments by Kim and March (1995), who studied the use of two formalisms for user (validation) and analyst (modeling) tasks, and Lee and Choi (1998), who compared four different formalisms in two task types. The commonly used analysis of performance by facets (Batra et al., 1990; Bock & Ryan, 1993; Lee & Choi, 1998; Liao & Palvia, 2000; Shoval & Shiran, 1997) is, in fact, a form of analysis of the moderating effects of task type, because modeling a specific facet can be seen as a subtask. As discussed above in the summary of results, the facet being modeled often moderates the impact of a specific modeling formalism on performance.

Finally, the research framework should explicitly acknowledge that various individual characteristics have differential effects on user performance and attitudes and that many of the effects of individual differences moderate the relationship between the data modeling formalism and the dependent variables. In addition, some of the relationships between the categories of individual characteristics affect each other in a significant way: Task-related experience affects an individual's technical data modeling skills (in addition to training), and the general individual differences (such as intelligence) moderate the relationship between the training an individual receives and the individual's skills.

We believe that the use of the framework in Figure 2 and any extensions of it would provide future human factors research on conceptual data modeling with a stronger conceptual foundation and give the researchers an incentive to specify the relationships between the variables of interest at a more detailed level and present them better in relation to other, potentially significant variables.

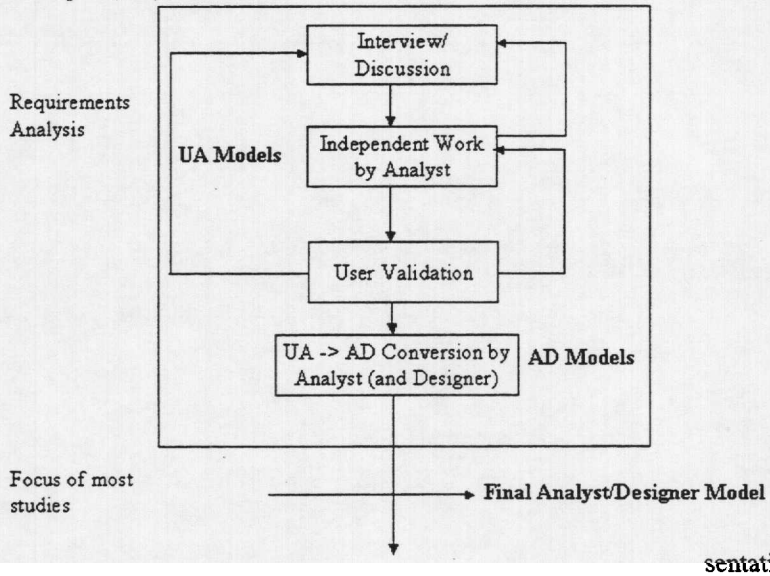
Differentiating User-Analyst and Analyst-Developer Models

In addition to suggesting a new underlying framework, we would also like to encourage the research community to investigate a broader range of uses for conceptual data modeling. As can be seen based on our findings to date, research on conceptual data modeling has focused on determining: a) the appropriate set of constructs that can be used to represent the structure of the database that reflects the requirements (the semantic power of the model) and b) how easy it is for these constructs to be used by IS professionals. The critical issue to note is that such a focus essentially means that all our research on conceptual data modeling has looked at issues related to the use of conceptual data modeling purely as a representational tool and/or a tool for communication between analysts and developers. A consequence of this focus has led to the blurring of the distinction between conceptual and logical data models; it has become common practice to use the same modeling constructs, for example, the ER model, to create schemas that are ostensibly at different levels.

To alleviate the confusion surrounding the terms conceptual and logical data modeling, we propose a new categorization of data models (schemas) based on their intended use. Such a classification should encourage the development of different models (with different constructs) for the traditional conceptual and logical modeling phases.

Figure 3 depicts the various places in the analysis and design phases where a data model can potentially be used. Based on this figure, we broadly classify data models

Figure 3: Suggested division between user-analyst (UA) and analyst-developer (AD) models



(schemas) into: a) models whose primary purpose is to facilitate user-analyst communication (UA models), and b) models that are used primarily to facilitate analyst-designer communication (AD models). Based on this classification, our findings indicate that the research to date on data modeling (including UML class diagrams) falls mostly into the domain of AD models.

At the beginning of this paper, we noted that one of the key uses of a conceptual model is as a communication tool between analysts and clients (users). This aspect of conceptual data modeling has been largely ignored in prior research. Thus, the range of research questions related to UA models is completely open. Below, we present some possible directions for such research, some of which are akin to the types of opportunities available when the ER model was first introduced by Peter Chen in 1976. The first challenge is in defining the characteristics of models that might be suitable for facilitating UA interaction.

What makes a good UA model? Navathe (1992) presents five criteria that can be used to differentiate between conceptual data models: expressiveness, simplicity, formality, minimality and unique interpretation. Given that the objective of a UA model is to promote user understandability of the model, simplicity is clearly of utmost importance. From an analyst perspective, the richer the set of constructs available in the model, the more likely it is that the analyst can capture the requirements fully and succinctly. Thus, more expressive models are also suited as UA models. For example, a data model that has support for capturing causation would clearly help analysts convey more information to the users and thus ensure that any signoff is based on an in-depth understanding of what is being conveyed. This is not to say that formality is not important in such a model. However, the fact that the user is likely to be knowledgeable of the domain and that the analyst has an opportunity to help clarify any misinterpretations reduces the importance of formality in an UA model. Hence, an UA model could sacrifice unambiguity to promote expressiveness and simplicity. A good AD model, on the other hand, should focus on the formality and unique interpretation because it is the model that is going to be used by designers to create the database (and the corresponding applications).

Thus, a UA model, unlike an AD model, needs to be driven more by its understandability by users. Therefore, an UA model needs to be motivated more by cognitive needs than an AD model. Another key issue that needs to be kept in mind when developing UA models is that these models should not be constrained by relationships that can be supported by databases or even application programs. The objective of these models is to create a set of representations using the types of relationships expressed in a requirements document so that they can be interpreted and verified by users. Thus, one needs to go beyond the traditional data modeling relationships, such as aggregation, generalization/specialization relationships, etc. Ramesh and Browne (1999) present a number of tools and techniques from the cognitive psychology and behavioral decision making literature that can potentially be used during requirements determination. Examples of representations that can potentially be adapted for use as a UA model include knowledge maps, influence diagrams, cognitive maps and affinity diagrams. In general, future research is clearly needed to improve our understanding of the efficient use and creation of appropriate modeling techniques for user-analyst communication. The extended framework presented in Figure 2 can then be used to conduct further research involving various aspects of such models.

New Areas of Focus

Finally, we would like to propose two additional foci for conceptual data modeling research: a) basic research on concept formulation, categorization, and usage;and b) applied research on data modeling processes.

First, we need a better understanding of the psychological processes in data modeling and the ways the tools affect these processes. This will enable us to find a firm theoretical basis for human factors research on data modeling. Researchers in this area should be interested not only in the characteristics of the current models, but the reasons underlying the potential performance differences between various approaches to data modeling. Batra's (1993) framework of error behaviors and the introduction of the GEMS model to this domain by Batra and Antony (2001) are excellent steps in the right direction. As Siau (1999) points out, cognitive science is potentially a very useful reference discipline; especially, the

research in cognitive science that has its roots in cognitive psychology or in artificial intelligence (Batra, 1993; Henderson & Peterson, 1992; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Smith & Medlin, 1981). Applied research in this field has been done, for example, in marketing and organizational behavior (for representative examples see Day & Lord, 1992; Fiol & Huff, 1992; Ozanne, Brucks, & Grewal, 1992).

The essence of all modeling is in the identification of concepts and categorization of them (Booch, 1994, Chapters 1-4; Coad & Yourdon, 1991, Chapter 1). The links between theoretical research on categorization and data modeling are still somewhat weakly defined although Parsons and Wand's (1997; 2000) work is a very important contribution and an excellent example of the type of research that is needed in this area. An additional important contribution would be a conceptual analysis of the characteristics of various data modeling techniques compared with categorization theories (see Henderson and Peterson (1992) for a concise introduction) and an empirical verification of the results of this research. The central focus of this research should be on the relationships between individual abilities, individuals' histories, situation characteristics, perceptions of reality, and categorization behavior. On the other hand, it is very important to note that data models are not (or at least should not be) created in a social vacuum; a data model describes a collective cognitive view about an organization. If reality is socially constructed (Berger & Luckmann, 1967) and information processing is greatly affected by social structures and forces (Salancik & Pfeffer, 1978; Weick, 1979), a closer analysis of the impact of social forces on data modeling (Ram & Ramesh, 1998) is warranted.

Second, and in addition to research focusing on fundamental psychological and social psychological processes, rigorous applied empirical research and theory development is also needed, but with a broader focus, i.e., work that is applicable to both the UA and AD models described in the previous section. In applied research, two important characteristics of the real world modeling tasks have to be taken into account. First, the process of model building, validation, and implementation is almost always iterative. Models are not built in a very limited amount of time and accepted without conceptual and empirical testing, or if they are, at least the implementation (and the implicit, but not the documented data model) will be changed if modeling errors lead to application errors. Second, the elicitation, representation, and validation phases of the modeling process are normally closely integrated, and the separation of them in research environments is often artificial.

In addition to broader tasks, a richer set of methodologies is also needed. A quantitative analysis of results obtained in a laboratory environment is not enough. In addition, qualitative techniques and field data are needed. For example, Batra and Davis (1992) used protocol analysis (Ericsson & Simon, 1993) in a laboratory environment. In-depth case studies in field environments — for exploratory and later for theory testing purposes — are also necessary to analyze the real effects of data modeling in organizational environments.

CONCLUSION

Conceptual data modeling forms an important foundation for systems development. In this paper, we have reviewed the existing research on conceptual data modeling and described avenues for further work in this area. We suggest that future research should pay significantly more attention to the role of conceptual modeling in facilitating user-analyst communication; so far, existing research has almost exclusively investigated issues related to analyst-developer models. In addition, we emphasized the importance of building a stronger theoretical foundation based on the work in cognitive science and other relevant reference disciplines.

ACKNOWLEDGMENTS

An earlier version of this paper was published in the Proceedings of the Sixth CAiSE/IFIP8.1 Intl. Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD'01). We gratefully acknowledge the highly valuable comments by the EMMSAD'01 participants and the reviewers and editors of all versions of this paper.

ENDNOTES

1) The concept of "entity" refers in this context not only to static objects but also to relevant activities and events within the domain of interest.
2) We acknowledge that our sample may not include some relevant papers published in the proceedings of specialized conferences.
3) Only those studies on object-oriented modeling have been included that have data modeling as their primary focus.

REFERENCES

Agarwal, R., Sinha, A. P., & Tanniru, M. (1996). The role of prior experience and task characteristics in object-oriented modeling: An empirical study. *International Journal Of Human-Computer Studies,* 45, 639-667.

Amer, T. (1993). Entity-relationship and relational database modeling representations for the audit review of accounting applications: An experimental examination of effectiveness. *Journal of Information Systems,* 7(1), 1-15.

Anderson, J. R. (1993). Rules of the Mind. Hillsdale, NJ: Lawrence Erlbaum Associates.

Batra, D. (1993). A framework for studying human error behavior in conceptual database modeling. *Information and Management,* 25(3), 121-131.

Batra, D., & Antony, S. R. (1994). Effects of data model and task characteristics on designer performance - a

References continued on page 18

Continued from page 12

laboratory study. *International Journal Of Human-Computer* Studies, 41(4), 481-508.

Batra, D., & Antony, S. R. (2001). Consulting support during conceptual database design in the presence of redundancy in requirements specifications: an empirical study. *International Journal Of Human-Computer Studies,* 54(1), 25-51.

Batra, D., & Davis, J. G. (1992). Conceptual data modeling in database design - similarities and differences between expert and novice designers. *International Journal Of Man-Machine Studies,* 37(1), 83-101.

Batra, D., Hoffer, J. A., & Bostrom, R. P. (1990). Comparing representations with relational and EER models. *Communications of the ACM,* 33(2), 126-139.

Batra, D., & Kirs, P. J. (1993). The quality of data representations developed by nonexpert designers: An experimental study. *Journal of Database Management,* 4(4), 17-29.

Batra, D., & Sein, M. K. (1994). Improving conceptual database design through feedback. *International Journal Of Human-Computer Studies,* 40(4), 653-676.

Batra, D., & Srinivasan, A. (1992). A review and analysis of the usability of data management environments. *International Journal Of Man-Machine Studies,* 36(3), 395-417.

Berger, P., & Luckmann, T. (1967). *The Social Construction of Reality.* New York: Doubleday.

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

Booch, G. (1994). *Object Oriented Analysis and Design With Applications* (2nd ed.). Redwood City, CA: Benjamin/Cummings.

Brosey, M., & Shneiderman, B. (1978). Two experimental comparisons of relational and hierarchical database models. *International Journal Of Man-Machine Studies*, 10, 625-637.

Burton-Jones, A., & Weber, R. (1999). Understanding relationships with attributes in entity-relationship diagrams. *Proc. of the Twentieth International Conference on Information Systems*, Charlotte, NC.

Cambell, D. (1992). Entity-relationship modeling: one style suits all. *Data Base,* 23(5), 12-18.

Cao, Q., Nah, F., & Siau, K. (2000). A meta-analysis of relationship modeling accuracy: comparing relational and semantic models. *Proc. of the 6th Americas Conference on Information Systems*, Long Beach, CA.

Chen, P. (1976). The entity-relationship model - toward the unified view of data. *ACM Transactions On Database Systems*, 1(1), 9-36.

Coad, P., & Yourdon, E. (1991). *Object-Oriented Analysis* (2 ed.). Englewood Cliffs, NJ: Prentice-Hall.

Day, D. V., & Lord, R. G. (1992). Expertise and problem categorization: The role of expert processing in organizational sense-making. *Journal of Management Studies,* 29(1), 35-47.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis. Verbal Reports as Data.* Cambridge, MA: The MIT Press.

Fiol, C. M., & Huff, A. S. (1992). Maps for managers: Where are we? Where do we go from here? *Journal of Management Studies,* 29(3), 267-285.

Hardgrave, B. C., & Dalal, N. P. (1995). Comparing object-oriented and extended-entity-relationship data models. *Journal of Database Management,* 6(3), 15-21.

Henderson, P. W., & Peterson, R. A. (1992). Mental accounting and categorization. *Organizational Behavior and Human Decision Processes,* 51(1), 92-117.

Hoffer, J. A. (1982). An empirical investigation into individual differences in database models. *Proc. of the Third International Conference of Information Systems,* Ann Arbor, MI.

Hull, R., & King, R. (1987). Semantic database modeling: survey, applications, and research issues. *ACM Computing Surveys,* 19(3), 201-260.

Jarvenpaa, S. L., & Machesky, J. J. (1989). Data analysis and learning: an experimental study of data modeling tools. *International Journal Of Man-Machine Studies,* 31, 367-391.

Juhn, S., & Naumann, J. D. (1985). The effectiveness of data representation characteristics on user validation. *Proc. of the Sixth International Conference on Information Systems*, Indianapolis, IN.

Kim, Y. G., & March, S. T. (1995). Comparing data modeling formalisms. *Communications of the ACM,* 38(6), 103-115.

Leavitt, H. J. (1965). Applied Organizational Change in Industry: Structural, Technological, and Humanistic Approaches. In J. G. March (Ed.), *Handbook of Organizations* (pp. 1144-1140). Chicago: Rand McNally.

Lee, H., & Choi, B. G. (1998). A comparative study of conceptual data modeling techniques. *Journal of Database Management,* 9, 26-35.

Liao, C. C., & Palvia, P. C. (2000). The impact of data models and task complexity on end-user performance: an experimental investigation. *International Journal Of Human-Computer Studies,* 52(5), 831-845.

McFadden, F. R., Hoffer, J. A., & Prescott, M. (1999). *Modern Database Management.* Reading, MA: Addison-Wesley.

Navathe, S. B. (1992). Evolution of data modeling for databases. *Communications of the ACM*, 35(9), 112-123.

Nordbotten, J. C., & Crosby, M. E. (1999). The effect of graphic style on data model interpretation. *Information System Journal,* 9(2), 139-155.

Ozanne, J. L., Brucks, M., & Grewal, D. (1992). A study of information search behavior during the categorization of new products. *Journal of Consumer Research,* 18(4), 452-463.

Palvia, P. C., Liao, C. C., & To, P.-L. (1992). The impact of conceptual data models on end-user performance. *Journal of Database Management,* 3(4), 4-15.

Parsons, J., & Wand, Y. (1997). Choosing classes in conceptual modeling. Communications of the ACM, 40(6), 63-69.

Parsons, J., & Wand, Y. (2000). Emancipating instances from the tyranny of classes in information modeling. *ACM Transactions On Database Systems,* 25(2), 228-268.

Ram, S., & Ramesh, V. (1998). Collaborative conceptual schema design: A process model and prototype system. *ACM Transactions On Information Systems,* 16(4), 347-371.

Ramesh, V., & Browne, G. J. (1999). Expressing causal relationships in conceptual database schemas. *Journal of Systems and Software,* 45(3), 225-232.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.

Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly,* 23, 427-456.

Shoval, P., & Even-Chaime, M. (1987). Data base schema design: an experimental comparison between normalization and information analysis. *Data Base*, 18(3), 30-39.

Shoval, P., & Frumermann, I. (1994). OO and EER conceptual schemas: A comparison of user comprehension. *Journal of Database Management*, 5(4), 28-38.

Shoval, P., & Shiran, S. (1997). Entity-relationship and object-oriented data modeling - An experimental comparison of design quality. *Data & Knowledge Engineering,* 21(3), 297-315.

Siau, K. (1999). Information Modeling and Method Engineering. *Journal of Database Management*, 10(4), 44-50.

Sinha, A. P., & Vessey, I. (1999). An empirical investigation of entity-based and object-oriented data modeling: a development life cycle approach. *Proc. of the Twentieth International Conference on Information Systems,* Charlotte, NC.

Smith, E. R., & Medlin, D. L. (1981). *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Srinivasan, A., & Te'eni, D. (1995). Modeling as constrained problem solving: an empirical study of the data modeling process. *Management Science,* 41(3), 419-434.

Teorey, T. J., Yang, D., & Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys,* 18(2), 197-222.

Weber, R. (1996). Are attributes entities? A study of database designers' memory structures. *Information Systems Research,* 7(2), 137-162.

Weick, K. E. (1979). Cognitive processes in organizations. *Research in Organizational Behavior*, 1, 41-47.

Heikki Topi received his Ph.D. in Management Information Systems from Indiana University in 1995. He is currently an Associate Professor of Computer Information Systems at Bentley College. His research focuses on usability issues in the fields of data management and systems analysis & design, management and commercial utilization of advanced telecommunications technologies with a special emphasis on mobile solutions, and the effects of time availability constraints on human-computer interaction.

V. Ramesh, Ph.D., is an Assistant Professor of Information Systems and Ford Motor Company Teaching Fellow in the Department of Accounting and Information Systems at Kelley School of Business, Indiana University. He has published over 25 papers in leading journals, books, and conferences. His areas of expertise are in database modeling and design, systems design and development, heterogeneous databases, and groupware systems.