Understanding Conceptual Schemas: Exploring the Role of Application and IS Domain Knowledge

Vijay Khatri[1] Iris Vessey[2] V. Ramesh[1]

Paul Clay[1] Sung-Jin Park[3]

[1] Kelley School of Business, Indiana University, Bloomington, IN 47405

[2] University of Queensland and Queensland University of Technology Brisbane QLD 4000, Australia

[3] School of Public and Environmental Affairs, Indiana University, Bloomington, IN 47405

vkhatri@indiana.edu *i.vessey@business.uq.edu.au* *venkat@indiana.edu*

pclay@indiana.edu *sungjpar@indiana.edu*

Corresponding Author: Vijay Khatri
Assistant Professor of Information Systems
Kelley School of Business
Indiana University
1309 E. 10th Street
Bloomington, IN 47405-1701
812-855-2581

Acknowledgements: The authors would like to thank the participants at workshops at Indiana University, the University of Arizona, and the University of Illinois, Chicago, the University of Queensland, Queensland University of Technology, and the University of Western Australia for their comments on earlier versions of this paper, and the Editor-in-Chief, the senior and associate editors, and reviewers for their insightful and constructive comments.

January 31, 2006

Understanding Conceptual Schemas:
Exploring the Role of Application and IS Domain Knowledge

Abstract

Although information systems (IS) problem solving involves knowledge of both the IS and application domains, little attention has been paid to the role of application domain knowledge. In this study, which is set in the context of conceptual modeling, we examine the effects of both IS and application domain knowledge on different types of schema understanding tasks: *syntactic* and *semantic comprehension tasks* and *schema-based problem-solving tasks*. Our thesis was that while IS domain knowledge is important in solving all such tasks, the role of application domain knowledge is contingent upon the type of understanding task under investigation. We use the theory of cognitive fit to establish theoretical differences in the role of application domain knowledge among the different types of schema understanding tasks. We hypothesize that application domain does not influence the solution of syntactic and semantic comprehension tasks for which cognitive fit exists, but does influence the solution of schema-based problem-solving tasks for which cognitive fit does not exist.

To assess performance on different types of conceptual schema understanding tasks, we conducted a laboratory experiment using two equivalent conceptual schemas in familiar and unfamiliar application domains (high and low application domain knowledge) and participants with high and low IS domain knowledge. As expected, we found that IS domain knowledge is important in the solution of all types of conceptual schema understanding tasks in both familiar and unfamiliar applications domains, and that the effect of application domain knowledge is contingent on task type. Our findings for the EER model were similar to those for the ER model. Given the differential effects of application domain knowledge on different types of tasks, this study highlights the importance of considering more than one application domain in designing future studies on conceptual modeling.

Keywords: Conceptual modeling, conceptual schema understanding, the theory of cognitive fit, syntactic comprehension tasks, semantic comprehension tasks, schema-based problem-solving tasks

Understanding Conceptual Schemas: Exploring the Role of Application and IS Domain Knowledge

1. Introduction

Domain knowledge, that is, knowledge of the area to which a set of theoretical concepts is applied, is fundamental to all disciplines (Alexander 1992). Domain knowledge has long been recognized as an important avenue of inquiry in educational research (see, for example, Alexander 1992; Alexander and Judy 1988). Domains that have been studied range from physics and economics, on the one hand, to history and reading, on the other hand. Further, thinking is dominated by content and skills that are domain-specific (McPeck 1990); and lack of domain knowledge results in inelegant problem-solving strategies (Alexander and Judy 1988).

In the Information Systems (IS) discipline, the term "domain knowledge" has dual significance. First, *IS domain* knowledge provides representations, methods, techniques, and tools that form the basis for the development of application systems. Second, those application systems are developed to organize/structure solutions to real-world problems that exist in a given business area, or *application domain*. IS problem solving therefore applies theoretical concepts from the IS domain to the application domain of interest. Hence, knowledge of IS and the application domain go hand-in-hand in solving IS problems.

Although a number of studies argue that application domain knowledge impacts IS problem-solving effectiveness (see, for example, Blum 1989; Curtis, Krasner, and Iscoe 1988; Glass and Vessey 1992), few studies have directly addressed this relationship empirically; Burton-Jones and Weber (1999); Purao, Rossi, and Bush (2002); Shaft and Vessey (1995 1998); Vessey and Conger (1993) are exceptions. Fewer studies still have investigated the role of application domain knowledge on an important sub-set of IS development: conceptual modeling.[1]

A 2001 CSC survey found that organizing and utilizing data is one of the top-three IS issues for organizations worldwide.[2] Although conceptual modeling of data-related requirements is a small phase within the overall organization and utilization of data, it has a greater impact than any other phase (Witt and Simsion 2000). Further, conceptual modeling impacts system development costs, system flexibility, and the ability to meet users' requirements (Moody 1998) and is particularly important in the context of the current regulatory environment (see, for example, the Sarbanes-Oxley Law and HIPAA[3]) in which documentation of the meaning of data, or data semantics, has significant legal implications.

Prior research in the area of conceptual modeling has focused on two aspects: *development* (see, for example, Kim and March 1995) and *understanding* of a conceptual schema (see, for example, Bodart, Patel, Sim, and Weber 2001; Kim and March 1995). While the former involves developing a schema based on stated requirements, the latter assumes existence of a schema and assesses schema understanding. Our focus in this research is on the latter. In the context of conceptual schema understanding, prior research has explored a wide variety of phenomena including the effects of the modeling formalism (see, for example, Kim and March 1995) and of ontological clarity (see, for example, Bodart et al. 2001; Burton-Jones and Weber, 1999; Shanks, Tansley, Nuredini, Tobin, and Weber 2002; Shanks, Nuredini, Tobin, Moody, and

[1] By *conceptual model,* we mean a formalism that is employed to develop a *conceptual schema* (or simply, a schema). A conceptual schema is an abstract representation of the structure of the data: entities, events, and their associations within an organization (Hoffer, Prescott, and McFadden 2005).
[2] http://www.csc.com/aboutus/uploads/CI_Report.pdf (last viewed on January 12, 2006).
[3] Health Insurance Portability and Accountability Act.

Weber 2003). The effects of application of domain knowledge, however, remain largely unexplored.

In this study, we conduct an exploratory investigation into the roles that knowledge of the IS (conceptual modeling) and application domains play in conceptual modelers' understanding of conceptual schemas. By exploring the effects of IS domain knowledge at different levels of application domain knowledge, we extend prior research in conceptual modeling, in general, that has investigated the effects of IS domain knowledge in the context of a single application domain (see, for example, Batra and Davis 1992; Weber 1996).

We examine application domain knowledge by exploring its effects on different types of schema understanding tasks in the belief that the role of knowledge of the application domain may be contingent upon task type. In doing so, we use the theory of cognitive fit as our theoretical base. Our overall research question is: "How do IS and application domain knowledge influence performance on different types of conceptual schema understanding tasks?" Hence, although our principal objective was to investigate the role of application domain knowledge, we sought a more complete view of conceptual schema understanding by examining the effects of both application and IS domain knowledge.

In the next section we examine prior research that has addressed the roles of IS and application domain knowledge in the realm of conceptual modeling. In Section 3, we characterize the conceptual schema understanding tasks that have been used in experiments to date and present the theory of cognitive fit as the foundation for investigating the effects of application domain knowledge on performance on such tasks. The next section presents the experimental methodology we used to test the hypotheses, while Section 5 presents the results of our analyses. Finally, we present the implications of our findings for research and practice.

2. Prior Research

In this section, we review prior literature that has examined the role of IS and application domain knowledge in conceptual modeling.

While we are not aware of any study that has examined the effects of IS domain knowledge on conceptual schema understanding directly, the effects of IS domain knowledge have been studied in a number of related contexts such as program comprehension (see, for example, Gugerty and Olson 1986; Wiedenbeck 1986), systems analysis and design (see, for example, Agarwal, Sinha, and Tanniru, 1996) and conceptual schema development (see, for example, Batra and Davis 1992). In general, participants with better IS domain knowledge have been found to perform better than those with lesser IS knowledge.

In a similar vein, while numerous conceptual modeling studies investigating a conceptual modeling phenomenon have each been framed in a specific application domain (see Topi and Ramesh 2002 for examples), very few studies have examined the effects of application domain knowledge per se. We are aware of just three such studies.

First, Burton-Jones and Weber (1999) investigated the inter-relationship of application domain familiarity and different ways of representing relationships with attributes, referred to as *ontologically-sound* and *unsound* representations, on understanding a conceptual schema. They found that while performance with each representation was similar when the domain was familiar, performance with the ontologically-sound representation was better when the domain was unfamiliar. They attributed their findings to cognitive dissonance resulting from the semantic ambiguities of an ontologically-unsound representation.

Second, Siau, Wand, and Benbasat (1995) conducted an experiment to investigate how experts used structural constraints (cardinality) associated with (isolated) binary relationships in familiar and unfamiliar settings. They found no differences in comprehension of structural constraints in the two settings. However, participants were provided with very little information on the application domain, which raises the question of whether they had sufficient information to make an informed decision.

Third, in a related, unpublished study, Siau, Wand, and Benbasat (1995) also report that a representation in which structural constraints conflicted with the underlying semantics resulted in what they called *semantic negligence*, i.e., the subjects focused on structural constraints and ignored the underlying semantics.

At a high level, these results suggest that problem solvers rely on the knowledge from a conceptual schema, alone, whenever they feel they can. The findings of these studies also show, however, that there are occasions in which knowledge of the application domain aids, or could aid, in interpreting those representations. Further, this analysis reveals that while some prior research has recognized the importance of application domain knowledge in conceptual schema understanding, little research has been undertaken to characterize its effects directly.

3. Theory

In this section we present the theoretical foundations for our work on conceptual schema understanding. We first present the types of conceptual schema understanding tasks that have been addressed in prior conceptual modeling research. We then analyze the effects of both IS and application domain knowledge on performance on the types of conceptual schema understanding tasks we address in this research, and state our hypotheses.

3.1 Characterizing Conceptual Schema Understanding Tasks

Schema understanding tasks can be viewed as either *read-to-do* (with access to schema) or *read-to-recall* tasks (without access to the schema) (Burkhardt, Detienne, and Wiedenbeck 2002). Recall tasks have been used to investigate problem solvers' knowledge structures, that is, chunks of knowledge that are stored in internal memory and reused when appropriate. Perhaps the best known studies of this type have been those conducted by Weber (see, for example, Bodart et al. 2001). Because our research is focused on evaluating performance on schema understanding tasks in which a conceptual modeler needs to examine a conceptual schema in order to derive information from it, we used read-to-do tasks.

Characterizing conceptual schema understanding tasks as read-to-do and read-to-recall, refers to the way in which this type of research is operationalized. A more in-depth categorization of the understanding tasks themselves has focused on the cognitive nature of the task. Prior to 1999, the most common method for assessing conceptual schema understanding required problem solvers to address a series of tasks, now called *comprehension tasks* (see, for example, Kim and March 1995), that required subjects to answer questions based on modeling constructs; these questions therefore focused on surface-level understanding (Bodart et al. 2001).

Two types of comprehension tasks that have been employed in prior literature are supported by research in the education literature that identifies two different types of knowledge, syntactic and semantic, required to solve problems (Shneiderman and Mayer, 1979; Mayer 1991).[4] We refer to such tasks as *syntactic* and *semantic comprehension tasks*. Syntactic knowledge involves understanding the vocabulary specific to a modeling formalism, for example, the ER model. *Syntactic comprehension tasks* are therefore those that assess the understanding of just the syntax of the formalism (conceptual model) associated with a schema. For example, the syntax for an entity type is a rectangle.

Semantic knowledge involves understanding the meaning, or the semantics of the data, embedded in the conceptual schema. In particular, data semantics refers to a set of mappings

---

[4] While their work was set in the context of programming languages, these concepts are also relevant to conceptual models.

from a representation language to agreed-upon concepts in the real world. Thus, *semantic comprehension tasks* are those that assess the understanding of the data semantics conveyed through constructs in the schema; for example, a rectangle, the symbol for an entity type, represents a collection of entity instances, that is, objects, things, events, or places (in the "real world") (Elmasri and Navathe 1994).

More recently, researchers have investigated tasks that require a deeper level of understanding than comprehension tasks, tasks that are referred to as *problem-solving tasks* (see Gemino 1999). We refer to a problem-solving task that can be solved using knowledge represented *in* the schema (Shanks et al. 2003) as a "schema-based problem-solving task." Such tasks resemble query tasks; respondents are requested to determine whether, and how, certain information is available from the schema.

A further type of problem-solving task, which we refer to as an "inferential problem-solving task," requires conceptual modelers to use information *beyond* what is provided in the schema (Gemino and Wand 2003). A number of recent studies have used this type of task in addition to comprehension tasks (see, for example, Bodart et al. 2001; Burton-Jones and Weber, 1999; Shanks et al. 2002; Shanks et al. 2003). See Appendix A for representative examples of the types of comprehension and problem-solving tasks identified in prior research.

For our research into conceptual schema understanding, we examined the comprehension and schema-based problem-solving tasks that are pertinent to conceptual modelers in practice; that is, we focused on read-to-do tasks that employ knowledge represented *in* the schema.

3.2 Solving Conceptual Schema Understanding Tasks

Conceptual schemas are well-formalized representations of data within a specific application domain. As we have seen, they are so well formalized that the understanding tasks investigated here *can* be solved using IS domain knowledge alone. We present and test theory that suggests, however, that application domain knowledge also influences the solution of certain types of schema understanding tasks; that is, we propose that the effect of application domain knowledge is contingent upon the type of task being addressed.

We investigate the effects of IS and application domain knowledge on conceptual schema understanding using participants with high and low IS knowledge in familiar and unfamiliar application domains. Figure 1 summarizes the effects that we propose below for the three conceptual schema understanding tasks investigated in this research.

Task Type	Analysis of Effects	Type of Knowledge Needed for Problem Solving	
		Familiar Application Domain	**Unfamiliar Application Domain**
	Comprehension Tasks		
Syntactic and Semantic	• Direct relationship between problem representation and task requirements o Cognitive fit exists o No further processing required	• ISDK required • ADK has no effect	• ISDK required • ADK has no effect
	Problem-Solving Tasks		
Schema-based Problem-solving	• Indirect relationship between problem representation and task requirements o Cognitive fit does not exist o Further processing required	• ISDK required • Required processing aided by knowledge of the application domain	• ISDK required • Required processing not aided due to lack of knowledge of the application domain

Figure 1: Summary of Effects of IS and Application Domain Knowledge on Different Types of Conceptual Schema Understanding Tasks

3.2.1 Effects of IS Domain Knowledge

As we have seen, IS domain knowledge has not been investigated in the specific context of conceptual schema understanding. Research on the role of IS domain knowledge has, however, been conducted in the context of conceptual schema development (see, for example, Batra and Davis 1992; Lee and Choi 1998; Moody, Shanks, and Darke 1998; Weber 1996). Batra and

Davis (1992) and Moody et al. (1998), for example, found that the quality of schemas developed by subjects with high IS domain knowledge were generally superior to those developed by subjects with low IS domain knowledge. Based on analogy with conceptual schema development, therefore, we expect that conceptual modelers with greater IS domain knowledge will perform better on all types of conceptual schema understanding tasks than those with lesser IS domain knowledge and that the effects will be apparent in both familiar and unfamiliar application domains. We investigate the following hypotheses.

> **Hypothesis 1.1:** Conceptual modelers with high IS domain knowledge perform better on syntactic comprehension tasks than those with low IS domain knowledge in both familiar and unfamiliar application domains.

> **Hypothesis 1.2:** Conceptual modelers with high IS domain knowledge perform better on semantic comprehension tasks than those with low IS domain knowledge in both familiar and unfamiliar application domains.

> **Hypothesis 1.3:** Conceptual modelers with high IS domain knowledge perform better on schema-based problem-solving tasks than those with low IS domain knowledge in both familiar and unfamiliar application domains.

3.2.2 Effects of Application Domain Knowledge

From the viewpoint of application domain knowledge, our basic thesis is that, although application domain knowledge is not essential to task solution, it may play a role in the solution of certain types of tasks; that is, we expect that the effect of application domain knowledge will be contingent upon the type of task being addressed. We use the theory of cognitive fit to describe the role that application domain knowledge plays in solving different types of conceptual schema understanding tasks.

The theory of cognitive fit (Vessey 1991) states that performance on a task will be enhanced when there is a cognitive fit (match) between the information emphasized in the type of *problem representation*[5] used and that required by the type of *problem-solving task* under consideration; see Figure 2. When the types of information emphasized in the problem-solving elements (in this case *problem representation* and *problem-solving task*) match, the problem solver uses processes (and therefore formulates a *mental representation for task solution*) that also emphasizes the same type of information. Consequently, the processes the problem solver uses to both act on the representation and to complete the task will match, and the problem-solving process will be facilitated. However, when the information in the *problem representation* and the *problem-solving task* do not match, similar processes cannot be used both to act on the *problem representation* and to solve the problem, and the *mental representation for task solution* will need to be transformed in order to solve the problem.

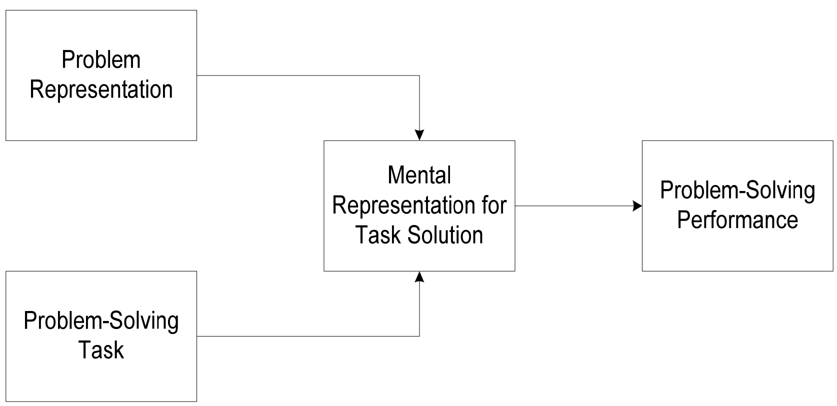


Figure 2: Cognitive Fit in Problem Solving (Vessey 1991)

In solving syntactic and semantic comprehension tasks, the way in which knowledge is represented in the conceptual schema matches that required for task solution, resulting in cognitive fit. Because cognitive fit exists, the problem solver can form a consistent *mental representation for task solution* using IS knowledge acquired from the schema alone.

---

[5] Note that references to specific constructs in our model are shown in *italics*.

Application domain knowledge therefore has little effect on *problem-solving performance* on such tasks. We state the following hypotheses.

> **Hypothesis 2.1:** Conceptual modelers are equally accurate on syntactic comprehension tasks in familiar and unfamiliar application domains.

> **Hypothesis 2.2:** Conceptual modelers are equally accurate on semantic comprehension tasks in familiar and unfamiliar application domains.

In solving schema-based problem-solving tasks, on the other hand, the knowledge represented in the schema does not match that required for task solution. Hence, cognitive fit does not exist and, instead of acquiring the required knowledge directly from the schema, conceptual modelers must transform knowledge emphasized in the *problem representation* to match that emphasized in the *problem-solving task* in order to form a mental representation that facilitates task solution (*mental representation for task solution*). The need to transform schema knowledge to solve the problem effectively increases the complexity of the task at hand. In this situation, application domain knowledge (ADK) may play a role in problem solution, thereby effectively reducing the complexity of the task under consideration.

Hence we test the hypothesis that application domain knowledge facilitates the solution of such tasks.

> **Hypothesis 2.3:** Conceptual modelers are more accurate on schema-based problem-solving tasks in a familiar than in an unfamiliar application domain.

The notions of fit and lack of fit associated with the effect of application domain knowledge on conceptual schema understanding tasks addressed theoretically by the theory of cognitive fit are supported empirically in two prior research studies of which we are aware. First, in a study in which possession of a "useful mental model" can be viewed as analogous to a situation in which cognitive fit exists, Mayer (1975) found that problem solvers who possess such a model perform better on tasks that require making inferences, tasks that require "*far-*

transfer." On the other hand, absence of such a model does not influence performance on "*near-transfer*" *tasks* because of the direct relationship between problem representation and task requirements. Second, Borthick, Bowen, Jones, and Tse (2001) found that "close" matches, in which the information required for task solution could be obtained directly from the schema, resulted in better performance than "far" matches, in which the problem solver needed to perform a number of transformations on that information.[6] Hence these studies provide empirical support for the arguments made above that are based on the theory of cognitive fit.

4. Research Methodology

We conducted a laboratory experiment to test the hypotheses presented above.

4.1 Task Setting

We investigated sales as the familiar application domain and hydrology, in this case the study of ground water, as the unfamiliar application domain. We expected that participants drawn from a business school (see the following section) would be more familiar with a sales application and less familiar with a hydrology application.

We investigated performance on the conceptual models most commonly used in practice: the ER and EER models (see Chen 1976 and Elmasri and Navathe 1994, respectively). A recent survey found that the ER model is the most commonly-used formalism for conceptual modeling, exceeding by far the usage of NIAM (now ORM) or the class diagrams of UML, for example (Davies, Green, Rosemann, Indulska, and Gallo forthcoming).

4.2 Participants

Study participants were 81 undergraduate business students, proficient in conceptual modeling, drawn from two courses offered in the business school of a large university in the U.S. mid-west.

[6] Note that Borthick et al. use the term "congruence" rather than "fit." (See, also, Vessey 2006.)

Participation was voluntary. Students could either earn up to 2% extra credit based on their performance in this experiment or complete a conceptual modeling assignment. All students chose to participate in the study. The average participant in the ER group (41 in number) had a GPA of 3.21; all participants were 20-25 years of age; 78% were male; and 73.2% had at least some work experience. The average participant in the EER group (40 in number) had a GPA of 3.19; 79.5% of participants were 20-25 years of age; 75% were male; 76.9% had at least some work experience.[7] We view our student participants as novice conceptual modelers.

4.3 Experimental Materials

Here we describe briefly the conceptual schemas for the familiar and unfamiliar application domains and then present examples of questions used in our understanding tasks.

4.3.1 Conceptual Schemas

The schemas in the familiar and unfamiliar application domains were syntactically equivalent: only the labels used for entity types, relationships, and attributes for the two schemas were different.

The familiar, sales schema (Figure 3) addressed concepts related to a typical order-processing application that included concepts such as SALES AREA, SALES TERRITORY, PRODUCT, PRODUCT LINE, and MANAGER. The unfamiliar, hydrology schema (Figure 4) was adapted from a schema for a ground water application created by the U.S. Geological Survey. This application included hydrological concepts such as SEEP, PLAYA, BORE HOLE, CASING, and ACCESS TUBE.[8]

---

[7] One subject did not report age and work experience.
[8] Figures 2 and 3 present the schemas for the sales and hydrology domains, respectively, using the ER model. A technical report with all of the schemas, the corresponding data dictionaries, and the understanding questions, is available at: http://www.iub.edu/~isdept/research/papers/tr143.pdf.

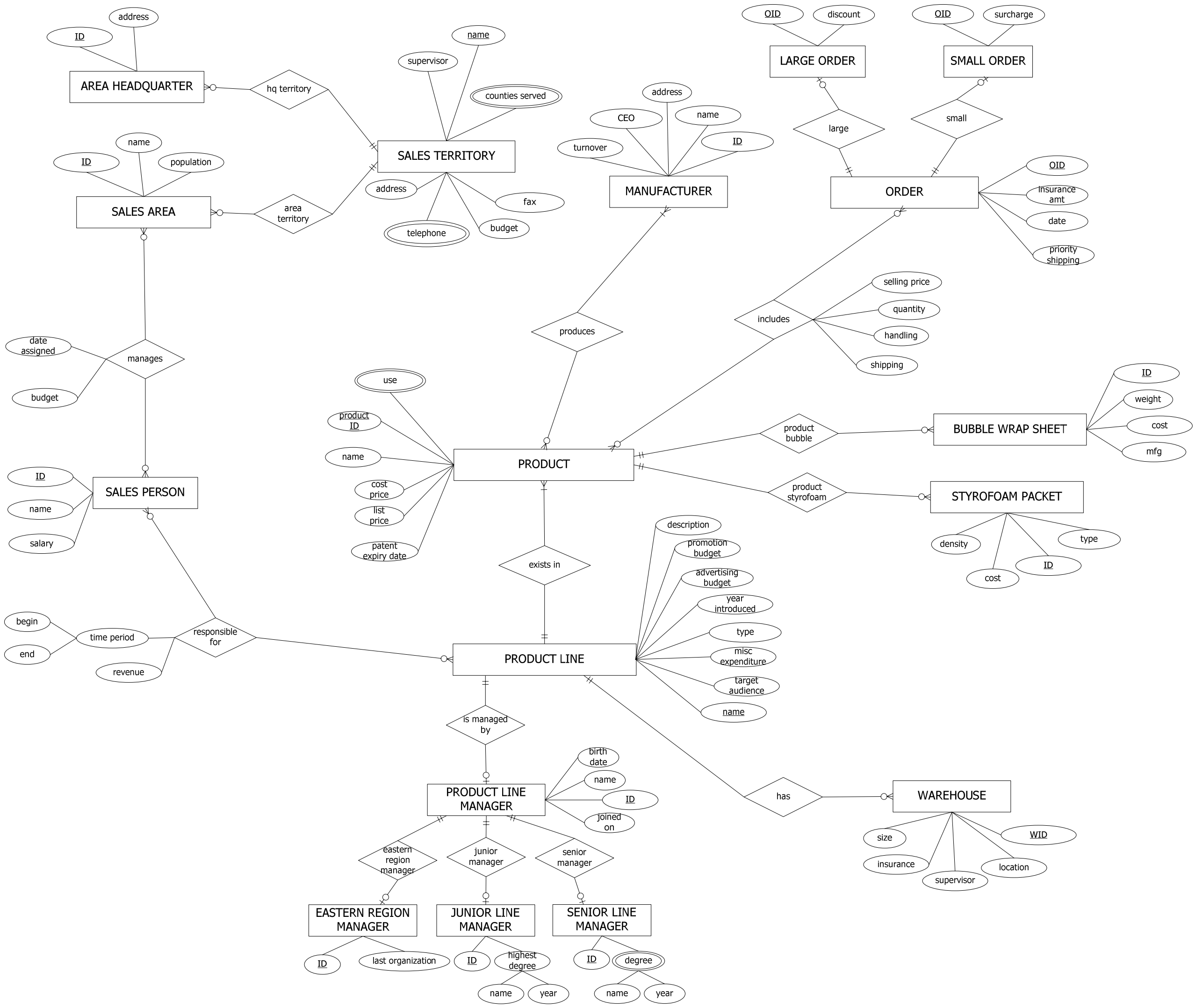


Figure 3: Sales Schema (Familiar Domain) Using the ER Model

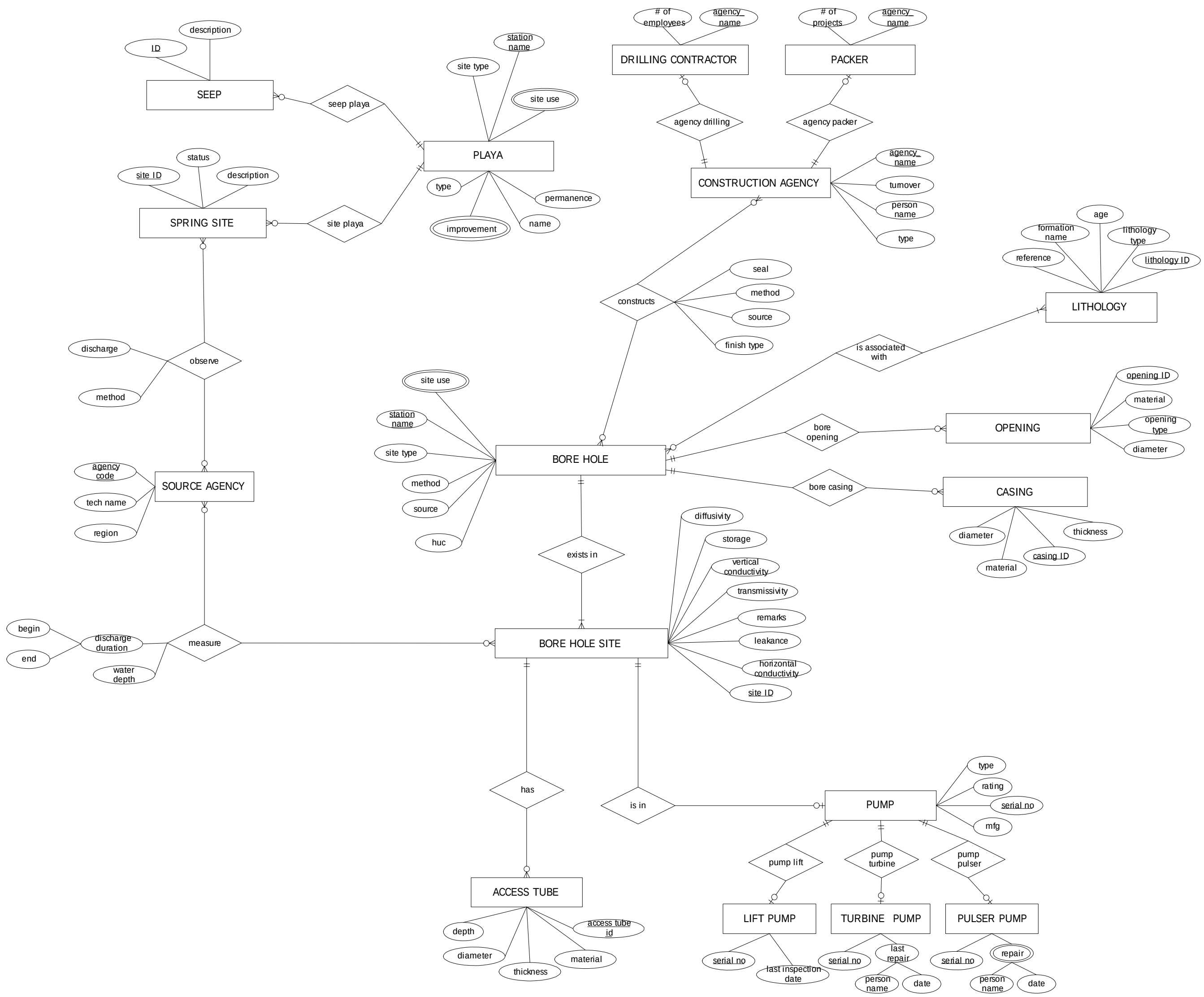


Figure 4: Hydrology Schema (Unfamiliar Domain) Using the ER Model

4.3.2 Conceptual Schema Understanding Tasks

We presented our participants with tasks in the same sequence: syntactic comprehension, semantic comprehension, and schema-based problem-solving. We presented the easiest tasks first so that participants would become increasingly familiar with the quite complex schema they were examining. In this way, we could have greater confidence that any effects of application domain knowledge on the schema-based problem-solving task were due to application knowledge itself and not to lack of knowledge of the schema per se. Question order in each category was controlled. The understanding tasks in familiar and unfamiliar application domains were based on the corresponding concepts in the two schemas and were therefore essentially equivalent.

The syntactic comprehension task was operationalized with ten multiple-choice questions that evaluated familiarity with the conceptual model syntax. For example, questions such as "What are the entity types that participate in the relationship "observe"?" and "How many attributes describe the entity type "OPENING"?" require an acquaintance with syntax, only. Note that a different font (Tahoma) was used to identify a concept from the schema.

The semantic comprehension task consisted of 20 multiple-choice questions. Ten questions were based on constructs such as entity type, attribute, and relationship. The other ten questions consisted of five questions each for super-types/sub-types and aggregates. Following is an example of a semantic question: "An access tube is related to: (a) no more than one bore hole site; (b) exactly one bore hole site; (c) at least one bore hole site; or (d) zero or more bore hole sites." Such a question evaluates the conceptual modeler's understanding of the data semantics, that is, the *meaning* of the cardinality constraint, represented on the schema.

The schema-based problem solving task consisted of six questions. Following is an example:

> "Earth scientists need to analyze the openings and casings that are associated with a bore hole. They need to ensure that a given bore hole in their analysis includes no more than two openings and one casing. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific."

We note that Shanks et al. (2003) used quite similar problem-solving tasks in their study. For example:

> "An Ontological Plastics supplier wishes to send samples of new and improved hoses to customers who regularly order hoses. Can we determine the number of hoses each customer has had delivered in the previous 3 months and the date of each delivery."

The difference lies in the fact that we asked explicitly "how" the problem could be solved, a requirement that was implicit in Shanks et al.'s task. Appendix B presents further examples of questions used in our understanding tasks.

4.4 Experimental Design

We used a 2 x 2 mixed design with knowledge of the IS domain as a between-subjects factor and familiarity with the application domain as a within-subjects factor for each of the ER and EER models. Participants were first randomly assigned to two groups (ER and EER). Finally, the presentation sequence of the two schemas was counterbalanced, thereby effectively controlling for any order effects.

Note that, by using two conceptual models, our research design allowed us to replicate our findings in different settings, thereby providing a stronger test of our theory than would have been possible had we examined a single representation only.

4.4.1 Investigating IS Domain Knowledge

To investigate our hypotheses related to application domain knowledge, recall that we operationalized high and low levels of application domain knowledge by having participants respond to understanding tasks in both familiar and unfamiliar application domains. To investigate our hypotheses related to IS domain knowledge, we needed to form groups of participants with high and low expertise in the IS domain, in this case, in conceptual modeling.

To form the groups, we used participants' scores on the database course that the participants had taken in either that semester or a prior semester. Because the course was taught by a number of instructors, we normalized participants' end-of-semester score by the average score for their section; that is, a normalized score of 1 indicated an average score in that section. The normalized database scores were used to form five groups of respondents. The middle quintile was dropped from further analyses and participants in the top two quintiles were interpreted as having high-IS knowledge and those in the bottom two as having low-IS knowledge.

The use of quintiles helped to create two groups that discriminated well with relatively few subjects being dropped. The low-IS group (Lo-IS) had average normalized scores on the database course of 0.85 and 0.87 for the ER and EER groups, respectively, while the corresponding scores for the high-IS group (Hi-IS) were 1.23 and 1.21, respectively.

4.4.2 Manipulation Checks

We conducted manipulation checks to assess differences between the ER and EER models and application domains, and assignment of participants across treatments. There were no significant differences, based on independent samples *t*-tests, in any of the control variables (age, gender, education, grade point average, work experience, self-reported knowledge of sales and hydrology domains) across both ER and EER groups, and groups with different schema presentation sequences, indicating that the participants were effectively randomized across treatments. A paired-samples *t*-test comparing the self-reported scores of participants' familiarity with each application domain showed that participants were more familiar with sales than hydrology (for the ER group, $t = 11.76$; $p < 0.01$; $df = 40$; and for the EER group, $t = 7.60$; $p < 0.01$; $df = 38$).

4.5 Pilot Study

We conducted a pilot study with graduate students who had conceptual modeling experience. This study used both the ER and EER models in three domains, sales, hydrology and abstract, which contained no meaningful terms (we used, for example, entity type names such as BETA and PHI; see Parsons 2002), and therefore conveyed even less application domain knowledge than our unfamiliar domain. The pilot study helped us to eliminate ambiguity in question wording, test the experimental procedures, and determine the length of time that the experiment would take to complete. It also revealed that an experiment using three application domains would be overly time consuming. We therefore dropped the abstract domain from the study proper.

4.6 Experimental Procedure

The study was conducted in four sessions, two each for participants using the ER and EER models. Each session began with one of the researchers providing a review of conceptual modeling concepts to the subjects for approximately 25 minutes. The participants then completed background questionnaires that sought demographic information as well as measured their a priori familiarity with the sales and hydrology application domains. Upon completion of the background questionnaire, the subjects were given an information sheet that described the syntax for their assigned model (ER or EER), as well as the schema for the first application domain. They were then asked to study the schema and the data dictionary (provided on paper) and to answer, in sequence, questions related to syntactic comprehension, semantic comprehension, and schema-based problem-solving tasks. All answers were typed into a document at a workstation. Within each task type, the questions were presented in three different sequences.[9] This procedure was then repeated for the second domain. Participants received different sequences of questions for the familiar and unfamiliar application domains.

4.7 Assessing Performance

Because the syntactic and semantic comprehension tasks took the form of multiple-choice questions with a single correct answer, the assessment was straightforward. Responses to the schema-based problem-solving questions were coded by two coders, naïve to our experimental

[9] A between-subjects' analysis for the three sequences indicated that there was no difference in performance among the sequences.

objectives, who were trained to use a well-formulated coding scheme to provide the data for further analysis.

We used a number of strategies to ensure the validity and integrity of the coded data. First, our two coders used subjective judgment of *overall quality* (OQ) (on a five-point scale with 1 = low and 5 = high) to assess participant responses. Use of subjective judgment allowed us to capture the nuances of a participant's response. However, to ensure that the coders had a common basis on which to make judgments, they first evaluated each response on: a) the extent to which relevant information was derived from the schema (data identification); and b) how well the response was constructed logically (reasoning). Coding on data identification and reasoning was done using a four-point scale (1, 2, 3, NA, with 1 = low, 3 = high, and "NA" = unable to evaluate).

Second, we prepared an extensive coding manual that described the coding scheme and the coding procedure, as well as presenting numerous examples. Third, we had two coders independently code the responses of all the participants. Fourth, we controlled for the well-known phenomenon of coder drift. Fifth, the coders resolved their differences to produce the data in its final form. Appendix C presents details of these strategies.

To evaluate coder reliability, we assessed reliability on overall quality using inter-rater reliability and confidence ratings. Inter-rater reliability was assessed using both raw agreement and the kappa statistic (Cohen 1960), which corrects raw agreement for agreement due to chance. The raw agreement was 97.42%, while the kappa statistic was 0.97 ($p < 0.01$). According to Landis and Koch (1977), a kappa value exceeding 0.81 is "almost perfect." Coder confidence in their overall score was measured on a three-point scale as a further measure of the viability of the coding (Orwin 1994; Kazdin 1977). As shown in Table 1, the coders were highly confident of their ratings.

Code	Code Interpretation	Statistics
1	I guessed the code	0.21%
2	I think my code has chance of being correct	2.86%
3	I am certain or almost certain that my code is correct	96.93%

Table 1: Confidence Rating for Schema-Based Problem-Solving Questions

5. Analysis and Results

In this section, we report the results of testing our two sets of hypotheses. In each sub-section, we present the results for the ER group, followed by those for the EER group.[10]

5.1 Testing for the Effects of IS Domain Knowledge

Here we investigate Hypotheses 1.1, 1.2, and 1.3, that participants with high IS domain knowledge outperform those with low IS domain knowledge on the three schema understanding tasks investigated in this research, in both familiar and unfamiliar application domains.

Tables 2a and 2b present the means, standard deviations, and statistical comparisons for each type of understanding task for participants with high and low IS domain knowledge in each of the application domains for the ER and EER models, respectively.[11] We report both *p*-values and effect sizes (Cohen's *d*).

Understanding	Familiar Application Domain	Unfamiliar Application Domain

[10] Tests of normality, using skewness and kurtosis, showed minor violations in a limited number of instances. The results with parametric and non-parametric tests were, however, essentially similar and we therefore report the results of the parametric tests.
[11] For comparison purposes, all scores that appear in this paper are normalized by the maximum scores possible for each of the tasks.

Task Type	Lo-IS (18)	Hi-IS (16)	*p*-value[1]	Effect Size[3] (Cohen's *d*)	Lo-IS (18)	Hi-IS (16)	*p*-value	Effect Size (Cohen's *d*)
Syntactic	0.96[2]	0.99	0.04**	Medium	0.94	0.97	0.10*	Medium
Comprehension	(0.05)	(0.03)		(0.62)	(0.07)	(0.06)		(0.46)
Semantic	0.78	0.81	0.23	Small	0.78	0.83	0.10*	Medium
Comprehension	(0.14)	(0.08)		(0.26)	(0.12)	(0.08)		(0.42)
Schema-based	0.62	0.76	<0.01***	Large	0.55	0.69	0.01***	Large
Problem-solving	(0.14)	(0.14)		(0.96)	(0.19)	(0.15)		(0.82)

Table 2a: Performance on Schema Understanding Tasks for the ER Group by IS Domain Knowledge

[1] *** significant at $p \leq 0.01$; ** significant at $p < 0.05$; * significant at $p < 0.1$.
[2] Means (and standard deviations) are presented in each case.
[3] Effect size: large (d=0.8); medium (d=0.5); and small (d=0.2).

Perusal of Table 2a for the ER group indicates that participants with high IS knowledge performed better than those with low IS knowledge for all types of understanding tasks, in both the familiar and unfamiliar domains, and that the effect size for IS domain knowledge was uniformly medium/large.[12] All differences were significant except that for the influence of IS domain knowledge on the semantic comprehension task in the familiar domain. With the exception, then, of one of the six tests, these results suggest that IS domain knowledge has similar effects on performance regardless of task type and familiarity with the application domain.

Understanding	**Familiar Application Domain**				**Unfamiliar Application Domain**			
Task Type	**Lo-IS (15)**	**Hi-IS (15)**	***p*-value[1]**	**Effect Size[3] (Cohen's *d*)**	**Lo-IS (15)**	**Hi-IS (15)**	***p*-value**	**Effect Size (Cohen's *d*)**
Syntactic	0.92[2]	0.97	0.04**	Large	0.91	0.97	0.01***	Large
Comprehension	(0.08)	(0.06)		(0.67)	(0.09)	(0.05)		(0.84)
Semantic	0.75	0.82	0.04**	Large	0.66	0.82	<0.01***	Large
Comprehension	(0.11)	(0.11)		(0.66)	(0.17)	(0.09)		(1.22)
Schema-based	0.52	0.61	0.09*	Medium	0.45	0.54	0.05**	Medium
Problem-solving	(0.15)	(0.19)		(0.49)	(0.12)	(0.16)		(0.64)

Table 2b: Performance on Schema Understanding Tasks for the EER Group by IS Domain Knowledge

[1] *** significant at $p \leq 0.01$; ** significant at $p < 0.05$; * significant at $p < 0.1$.
[2] Means (and standard deviations) are presented in each case.
[3] Effect size: large (d=0.8); medium (d=0.5); and small (d=0.2).

[12] Although the p-values for syntactic and semantic comprehension tasks in the unfamiliar application domain were marginally significant at a less conservative alpha value of 0.1, the effect size was uniformly medium.

Perusal of Table 2b shows similar results for the EER group. All comparisons were significant, and the effect sizes were also medium/large. Hence, overall, 11 of 12 results suggest that IS domain knowledge affects *problem-solving performance* on all types of understanding tasks in both familiar and unfamiliar domains, supporting Hypotheses 1.1, 1.2, and 1.3.

5.2 Testing for the Effects of Application Domain Knowledge

Next, we investigate Hypotheses 2.1, 2.2, and 2.3, that the influence of application domain knowledge on performance on schema understanding tasks is contingent on the type of understanding task under investigation in both familiar and unfamiliar application domains.

Table 3 presents the normalized means, standard deviations, and statistical comparisons (paired-samples *t*-tests) for each type of understanding task in each of the application domains for both the ER and EER groups.[13] Perusal of the table for the ER group revealed no significant performance differences for syntactic and semantic comprehension tasks; further, the effect sizes were small. Hence, Hypotheses 2.1 and 2.2 received null support. On the other hand, performance on the schema-based problem-solving task in the familiar domain was superior to performance in the unfamiliar domain, with a medium effect size. This finding supports Hypothesis 2.3.

Perusal of Table 3 for the EER group shows that our findings are similar to those for the ER group both for significant effects and effect sizes, again supporting Hypotheses 2.1, 2.2, and 2.3.

Our findings for both conceptual models therefore support our hypotheses on the effect of application domain knowledge on conceptual schema understanding, that the influence of

[13] Having familiarity with the application domain as a within-subjects variable effectively controls for differences in knowledge of the IS domain.

application domain knowledge on performance on schema understanding tasks is contingent upon the type of task being addressed.

Understanding Task Type	ER Group				EER Group			
	Familiar (41)	Unfamiliar (41)	*p*-value[1]	Effect Size (Cohen's *d*)	Familiar (40)	Unfamiliar (40)	*p*-value	Effect Size (Cohen's *d*)
Syntactic Comprehension	0.96[2] (0.06)	0.95 (0.07)	0.44	Small (0.15)	0.95 (0.07)	0.94 (0.07)	0.42	Small (0.14)
Semantic Comprehension	0.79 (0.11)	0.80 (0.11)	0.60	Small (0.07)	0.79 (0.12)	0.76 (0.15)	0.11[3]	Small (0.22)
Schema-based Problem-solving	0.68 (0.15)	0.60 (0.18)	<0.01***	Medium (0.50)	0.57 (0.17)	0.52 (0.16)	0.01**	Medium (0.35)

Table 3: Performance on Schema Understanding Tasks for the ER and EER Group by Application Domain Familiarity

[1] *** significant at $p \leq 0.01$; ** significant at $p < 0.05$; * significant at $p < 0.1$.
[2] Means (and standard deviations) are presented in each case.
[3] Note that while the *p*-value is close to 0.1, the effect size is small.

6. Discussion and Implications

This research explored the role of knowledge of the IS and application domains in conceptual schema understanding. We examined the overall research question: "How do IS and application domain knowledge influence performance on different types of conceptual schema understanding tasks?" To do so, we conducted an experiment in which we manipulated both IS and application domain knowledge.

In this section we discuss our findings and present the contributions of our research. We conclude with the implications of our research for both future research and for practice.

6.1 Discussion of the Findings

Our research resulted in three specific findings. First, we found that IS domain knowledge consistently affects problem solving on all types of schema understanding tasks in both familiar and unfamiliar applications domains. Second, we found that the effect of application domain knowledge on problem-solving performance is contingent on the type of conceptual schema understanding task under investigation. We found no evidence to suggest that application domain knowledge has an effect on performance on tasks for which cognitive fit exists, that is, on tasks that involve extracting knowledge that is represented directly in the conceptual schema, such as the knowledge required in syntactic and semantic comprehension tasks. We found, however, that application domain knowledge does aid problem solvers in solving more demanding tasks. Specifically, performance was better in the familiar domain for the schema-based problem-solving task in which cognitive fit between the schema and the task requirements did not exist, that is, tasks that require transformation of the knowledge in the conceptual schema into a form that facilitates task solution. Finally, our findings for the EER model were similar to those for the ER model thus providing stronger support for our hypotheses

Our research makes a number of contributions to IS research. First, by demonstrating the differential effects of application domain knowledge on different types of tasks, this study highlights the importance of considering more than one application domain context in designing future studies on conceptual modeling.

Second, this study highlights a further dimension of fit, that of the "extent" of fit between *problem-solving task* and *problem representation,* which adds to the growing body of literature attesting to the utility of the theory of cognitive fit (see Vessey 2006).

Third, we characterize conceptual schema understanding tasks as involving either comprehension or problem solving. Further, each of these types of tasks is made up of two sub-types, syntactic or semantic in the case of comprehension tasks and schema-based and inferential in the case of problem-solving tasks. Using this characterization of tasks in future research will help to develop a cumulative tradition in the area of conceptual schema understanding.

Finally, we employed a particularly rigorous approach to coding subjective data in this research, as well as assessing the reliability of coding using both Cohen's Kappa and confidence ratings. Other researchers can use our approach as a model for their own studies that use this type of data.

Our study has the following limitations. First, we conducted our investigation using students who were relatively inexperienced in using real world conceptual schemas. We characterize them as novice conceptual modelers. Second, we investigated just two of the possible conceptual models of many that might be investigated. Note, however, that little research in conceptual modeling attempts to replicate results with two different models as we did in this experiment. Third, our study was conducted in a laboratory setting, which means that it suffered the typical limitations of all experiments. However, we were able to control for many aspects that might have come into play had we conducted our study in a professional setting.

Fourth, Hypotheses 2.1 and 2.2 required us to test the null. We provide the following support for our approach: 1) these hypotheses follow naturally from the theory used in this research: Hypotheses 2.1 and 2.2 help contrast the situation when cognitive fit between knowledge presented in the problem representation (in this case, the conceptual schema) and that required for task solution exists and when it does not exist; 2) Hubbard and Armstrong (1997) strongly recommend "telling the story," something that we cannot do without testing Hypothesis 2 for each of our schema understanding tasks; 3) these hypotheses help to provide additional insight into the boundaries of prior research that has not investigated the effects of knowledge of the application domain; and 4) prior research suggests that in order to determine null support researchers should replicate effect sizes and/or findings (Nickerson 2000); in this experiment, we obtained similar findings and effect sizes for the ER and EER models. Note also that prior research suggests that non-significance does not tend to replicate more consistently than significance (Schmidt and Hunter 1997).

Finally, we presented the tasks to participants in the same order (syntactic, semantic, schema-based problem solving) so that we could have greater confidence that any effects of application domain knowledge on the more demanding schema-based problem-solving tasks were due to application knowledge itself, and not to lack of knowledge of the schema. Because we presented our tasks always in the same sequence, fatigue may have played a role in our findings. In the presence of fatigue effects, however, performance in the second domain would have been worse than in the first domain for each of the problem-solving tasks, a phenomenon we did not observe. Further, subjects were given different types of tasks in the same order in familiar and unfamiliar domains; thus, in comparing performance across IS and applications domains, the effects of fatigue were essentially controlled.

6.2 Implications of the Findings

Here we examine the implications of our findings for future research and for practice.

6.2.1 Implications for Research

Our findings have a number of implications for researchers. First, research needs to be conducted to establish the boundaries of our theory. Whetten (1989) suggests that a theory should elaborate on: a) what factors (variables, constructs, concepts) should be logically considered part of the theory; b) how those factors are related; c) why those beliefs are valid; that is, what is the theory behind them; and d) who, where, and when; that is, what are the conditions under which the theory holds. While we have addressed each of these issues to a certain extent in our initial investigation of this phenomenon, much work remains to be done to establish its applicability on a broader scale.

Some of Whetten's points relate to the typical arguments for generalizability of research findings. In our case, our findings are generalizable only to novice IS conceptual modelers; future research should therefore be undertaken to examine the applicability of our findings to the general population of IS professionals who have years of experience both in the IS domain and, most commonly, in multiple application domains. Research also needs to be conducted to determine whether our theory is valid for other types of conceptual models. While we investigated two conceptual models, ER and EER, future research should address other types of models such as ORM and the class diagrams of UML.

More far-reaching, however, is the need to determine the "why" to which Whetten alludes because "why" establishes the validity and future applicability of the theory. First, future research needs to address, for example, the underlying characteristics that render application domain knowledge useful. Notions from cognitive psychology relating to declarative and procedural knowledge may prove a useful starting point. Syntactic and semantic comprehension tasks involve declarative or factual knowledge, while schema-based problem-solving tasks require procedural knowledge, that is, knowledge required to apply declarative knowledge (Anderson 1983). Examining these tasks in this light may lead to a deeper understanding of the theory associated with the role of application domain knowledge in IS problem solving.

Second, future research should also seek to characterize in greater depth the processes by which application domain knowledge influences problem solving when cognitive fit does not exist.

Finally, preliminary evidence suggests that provision of ontologically-sound conceptual models may compensate for lack of domain knowledge (Burton-Jones and Weber 1999). Our current analyses suggest that such effects may be due to cognitive fit. Further research therefore needs to be conducted to characterize the dimensions of fit in this context.

6.2.2 Implications for Practice

Our research also has several implications for practice. First, if an organization has scarce resources in a particular application domain, our research provides guidelines to suggest the types of tasks on which they are best deployed. While simple syntactic and semantic comprehension tasks may be conducted largely without the need for application domain knowledge, it is beneficial to the conduct of schema-based problem-solving tasks. Management should therefore seek to place application domain specialists on those IS projects that require transformations of the knowledge represented in the schema, such as those involving querying databases.

Second, organizations providing training for conceptual modelers should focus not only on tool knowledge and IS domain knowledge, but also on application domain knowledge.

Third, the growing body of evidence pointing to the importance of application domain knowledge in certain types of IS problem solving suggests that tool builders should investigate ways to incorporate characteristics of the application domain into their tools, for example, through the use of domain-specific modeling patterns and templates.

7. Conclusions

The role of the application domain is an issue that has been largely neglected in research on conceptual modeling. Our research addresses the role of IS and application domain knowledge in understanding conceptual schemas both theoretically, via the theory of cognitive fit, and empirically. Specifically, we address the role of IS and application domain knowledge on the performance of conceptual modelers on different types of conceptual schema understanding tasks. Our research shows that while IS domain knowledge is important to the solution of all types of schema understanding tasks, application domain knowledge affects the solution of just schema-based problem-solving tasks, tasks for which conceptual modelers must transform knowledge in the schema into a form suitable for task solution.

References

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

Alexander, P.A., “Domain Knowledge: Evolving Themes and Emerging Concerns,” *Educational Psychologist,* 27 (1), 1992, pp. 33-51.

Alexander, P.A. and Judy, J.E. "The Interaction of Domain-Specific and Strategic Knowledge in Academic Performance," *Review of Educational Research,* 58 (4), 1988, pp. 375-404.

Anderson, J.R. *The Architecture of Cognition*. Mahwah, N.J: Lawrence Erlbaum Associates, 1983.

Batra, D. and Davis, J. G., "Conceptual data modelling in database design: similarities and differences between expert and novice designers," *International Journal of Man-Machine Studies*, 37, 1992, pp. 83-101.

Blum, B.A., "A Paradigm for the 1990s Validated in the 1980s," *Proceedings of the AIAA Conference*, 1989, pp. 502-511.

Bodart, F., Patel, A., Sim, M., and Weber, R. "Should Optional Properties Be Used in Conceptual Modelling? A Theory and Three Empirical Tests," *Information Systems Research* (12:4) 2001, pp. 384-405.

Borthick , A.F., Bowen, P.L., Jones, D.R., and Tse, M.H.K. "The Effects of Information Request Ambiguity and Construct Incongruence on Query Development," *Decision Support Systems*, (32: 1) 2001, pp. 3-25.

Burkhardt, J.-M., Détienne, F., and Wiedenbeck, S. "Object-Oriented Program Comprehension: Effect of Expertise, Task and Phase," *Empirical Software Engineering* (7:2) 2002, pp. 115-156.

Burton-Jones, A. and Weber, R. "Understanding relationships with attributes in entity-relationship diagrams," *Proceedings of the Twentieth International Conference on Information Systems*, 1999, pp. 214-228.

Chen, P.P. "The Entity-Relationship Model - Toward a Unified View of Data," *ACM Transactions of Database Systems* (1:1) 1976, pp. 9-36.

Cohen, J. "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement* (20) 1960, pp. 37-46.

Curtis, B., Krasner, H., and Iscoe, N., "A Field Study of the Software Design Process for Large Scale Systems," *Communications of the ACM*, 31 (11), 1988, pp. 1268-1287.

Davies, I., Green, P., Rosemann, M., Indulska, M., and Gallo, S. "How do practitioners use conceptual modeling in practice?" *Data and Knowledge Engineering*, *forthcoming*.

Elmasri, R. and Navathe, S.B. *Fundamentals of Database Systems*, (Second ed.) Benjamin/ Cummings Publishing Co., Redwood City, CA, 1994.

Gemino, A., "Empirical methods for comparing system analysis modelling techniques," *Unpublished PhD thesis*, University of British Columbia, Vancouver, B.C., Canada, 1999.

Gemino, A. and Wand, Y., "Evaluating modeling techniques based on models of learning," *Communications of the ACM* (46:10) 2003, pp. 79-84.

Glass, R. L. and Vessey, I., "Toward a Taxonomy of Application Domains: History," *Journal of Systems and Software* 1992, pp. 189-199.

Gugerty L. and Olson G., " Debugging by skilled and novice programmers," *Proceedings of the SIGCHI conference on Human factors in computing systems*, Boston, Massachusetts, 1986, pp. 171-174.

Hoffer, J. A., Prescott, M. B., and McFadden, F. R., *Modern Database Management*, Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

Hubbard R. and Armstrong J. S., "Publication Bias Against Null Results," *Psychological Reports*, 80, 1997, pp. 337-338.

Kazdin, A.E. "Artifact, bias and complexity of assessment: the ABCs of reliability," *Journal of Applied Behavior Analysis* (10:1) 1977, pp. 141-150.

Kim, Y.-G. and March, S.T. "Comparing Data Modeling Formalisms," *Communications of the ACM* (38:6) 1995, pp. 103-115.

Landis, J.R. and Koch, G.G. "The measurement of observer agreement for categorical data," *Biometrics* (36) 1977, pp. 159-174.

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

Mayer, R. E., "Different Problem-Solving Competencies Established in Learning Computer Programming with and without Meaningful Models," *Journal of Educational Psychology*, (67: 6), 1975, pp. 725-734.

Mayer, R.E. *Thinking, Problem Solving, Cognition* W. H. Freeman and Company, New York, NY, 1991, pp. 560-578.

McPeck, H., "Critical Thinking and Subject Specificity: A Reply to Ennis," *Educational Researcher*, (19: 4) 1990, pp. 10-12.

Moody, D. L., "Metrics for Evaluating the Quality of Entity Relationship Models," 17th International Conference on Conceptual Modeling (ER 1998), Singapore, Tok Wang Ling, Sudha Ram, Mong-Li Lee (Eds.), November 16-19, 1998, pp. 211-225.

Moody, D.L., Shanks, G.G., and Darke, P. "Improving the Quality of Entity Relationship Models - Experience in Research and Practice," *17th International Conference on Conceptual Modeling*, Singapore, 1998, pp. 255-276.

Nickerson, R. S., "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, (5:2) 2000, pp. 241-301.

Orwin, R.G. "Evaluating coding decisions," in: *The Handbook of Research Synthesis,* H. Cooper and L.V. Hedges (eds.), Russell Sage Foundation, New York, 1994, pp. 139-162.

Parsons, J., Effects of local versus global schema diagrams on verification and communication in conceptual data modeling, *Journal of Management Information Systems* (19: 3) 2002, pp. 155-183.

Purao, S., Rossi, M., and Bush, A., "Toward an Understanding of the Use of Problem and Design Spaces During Object-Oriented System Development," *Information and Organization*, 12, 2002, pp. 249-281.

Schmidt, F. L. and Hunter, J. E., “Eight common but false objections to the discontinuation of significance testing in the analysis of research data." In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? Hillsdale, NJ: Erlbaum, 1997, pp. 37-64.

Shaft, T. and Vessey, I., "The Relevance of Application Domain Knowledge: The Case of Computer Program Comprehension," *Information Systems Research* (6:3) 1995, pp. 286-299

Shaft, T. and Vessey, I., "The Relevance of Application Domain Knowledge: Characterizing the Computer Program Comprehension Process," *Journal of Management Information Systems* (15:1) 1998, pp. 51-78.

Shanks, G, Tansley, E., Nuredini, J., Tobin, D, Moody, D., and Weber, R., "Representing Part-Whole Relationships In Conceptual Modeling: An Empirical Evaluation," *Proceedings of the Twenty-Third International Conference on Information Systems*, 2002, pp. 89-100.

Shanks, G, Nuredini, J., Tobin, D, Moody, D., and Weber, R., "Representing Things and Properties in Conceptual Modeling: An Empirical Investigation," *European Conference on Information Systems*, 2003.

Shneiderman, B. and Mayer, R.E. "Syntactic/Semantic interactions in programmer behavior: A model and experimental results," *International Journal of Computer and Information Science* (8) 1979, pp. 219-238.

Siau, K., Wand, Y. and Benbasat, I., "A Psychological Study on the Use of Relationship Concept -- Some Preliminary Findings," *Proceedings of the 7th International Conference on Information Systems Engineering (CAiSE)*, Jyvaskyla, Finland, 1995, pp. 341-354.

Topi, H. and Ramesh, V. "Human Factors Research on Data Modeling: A Review of Prior Research, an Extended Framework and Future Research Directions," *Journal of Database Management* (13:2) 2002, pp. 3-19.

Vessey, I., "Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature," *Decision Sciences*, (22) Spring 1991, pp. 219-240.

Vessey, I. and S. Conger, "Learning to Specify Information Requirements: The Relationship between Application and Methodology." *Journal of Management Information Systems*, (10:2) Fall 1993, pp. 177-201.

Vessey, I. The theory of cognitive fit: One aspect of a general theory of problem solving? in P. Zhang and D. Galletta (eds.), *Human-computer interaction and management information systems: Foundations, Advances in Management Information Systems Series*, Armonk, NY: M.E. Sharpe, 2006.

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

Whetten, D.A., “What Constitutes a Theoretical Contribution?” *Academy of Management Review*, (14:4) 1989, pp. 490-495

Wiedenbeck, S., Processes in computer program comprehension, " In E. Soloway and S. Iyengar (eds.), *Empirical studies of programmers*, 1986, pp. 48-57.

Witt, G.C. and G.C. Simsion, *Data Modeling Essentials: Analysis, Design and Innovation*, The Coriolis Group, 2000.

Appendix A

Types of Schema Understanding Tasks	Examples
Comprehension	
Syntactic	Kim and March (1995) How many attributes describe the INSTITUTION entity? (a) 0; (b) 1; (c) 2; (d) 3; (e) 4 Parsons (2002) List the properties of alphas.
Semantic	Siau, Wand, and Benbasat (1995) Choose one of the following two options that more correctly reflects the participation of the Employee entity type in the Assigns relationship type: (a) must assign; (b) may assign Kim and March (1995) Every employee has a unique employee #. (a) True (b) False Bodart, Patel, Sim, and Weber (2001) Must a research program have at least two projects? (a) yes; (b) no; (c) not sure Parsons (2002) Some Gammas may caudate some Sigmas. True/False Shanks, Nuredini, Tobin, Moody, and Weber (2003) Can an employee be assigned to manage more than one customer at a time? (a) yes; (b) no; (c) not sure
Problem-Solving	
Schema-based	Shanks, Nuredini, Tobin, Moody, and Weber (2003) An ontological plastics supplier wishes to send samples of new and improved hoses to customers who regularly order hoses. Can we determine the number of hoses each customer has delivered in the previous 3 months and the date of each delivery?
Inferential	Bodart, Patel, Sim, and Weber (2001) A research project that was supposed to be completed last month has not been completed. What reasons can you provide for delay in completion? Write down as many reasons as you can think of.

Appendix B

This material with sample understanding questions is presented here for ease of reference only. A technical report with all of the schemas, the corresponding data dictionaries, and the understanding questions, is available at: http://www.iub.edu/~isdept/research/papers/tr143.pdf.

Hydrology	Sales
Syntactic Comprehension Task	
What is the minimum:maximum cardinality of the relationship between "BORE HOLE" and "CONSTRUCTION AGENCY"? (a) 0:M and 0:M; (b) 1:M and 0:M; (c) 0:M and 1:M; (d) 1:1 and 0:M	What is the minimum:maximum cardinality of the relationship between "PRODUCT" and "ORDER"? (a) 0:M and 0:M; (b) 1:M and 0:M; (c) 0:M and 1:M; (d) 1:1 and 0:M
How many entity types participate in the relationship "measure"? (a) 2; (b) 3; (c) 4; (d) 5	How many entity types participate in the relationship "responsible for"? (a) 2; (b) 3; (c) 4; (d) 5
What are the entity types that participate in the relationship "observe"? (a) "SOURCE AGENCY" and "BORE HOLE SITE"; (b) "BORE HOLE SITE" and "SPRING SITE"; (c) "CONSTRUCTION AGENCY" and "BORE HOLE"; (d) "SPRING SITE" and "SOURCE AGENCY"	What are the entity types that participate in the relationship "manages"? (a) "SALES PERSON" and "PRODUCT LINE"; (b) "PRODUCT LINE" and "SALES AREA"; (c) "ORDER" and "PRODUCT"; (d) "SALES AREA" and "SALES PERSON"
Which of the following is a multi-valued attribute? (a) "last repair"; (b) "improvement"; (c) "material"; (d) "thickness"	Which of the following is a multi-valued attribute? (a) "highest degree"; (b) "telephone"; (c) "weight"; (d) "supervisor"
Which term describes the attribute "discharge duration"? (a) Multi-valued; (b) Composite; (c) Identifier; (d) None of the above	Which term describes the attribute "time period"? (a) Multi-valued; (b) Composite; (c) Identifier; (d) None of the above
Semantic Comprehension Task	
A source agency engages in measurement: (a) at exactly one bore hole site; (b) at the most one bore hole site; (c) no more than one bore hole site; (d) none of the above	A sales person is responsible for: (a) exactly one product line; (b) at the most one product line; (c) no more than one product line; (d) zero or more product lines
A given pump must necessarily be related to a(n): (a) bore hole; (b) access tube; (c) bore hole site; (d) lithology	A product line manager must necessarily be related to a(n): (a) product; (b) warehouse; (c) product line; (d) manufacturer
A packer is characterized by: (a) site use; (b) person name; (c) # of employees; (d) none of the above	A small order is characterized by: (a) counties served; (b) insurance amount; (c) discount; (d) none of the above
A given lift pump may also be a: (a) turbine or pulser pump; (b) pneumatic pump; (c) centrifugal pump; (d) cannot determine	A given eastern region manager may also be: (a) junior line manager or senior line manager; (b) chief financial officer; (c) chief information officer; (d) cannot determine
A given playa has: (a) exactly one site use; (b) at least one site use; (c) no more than one site use; (d) multiple site uses	A sales territory serves: (a) exactly one county; (b)at least one county; (c) no more than one county; (d) multiple counties
Schema-Based Problem-Solving Task	
Geologists and hydrologists need to decide which bore holes to include in their groundwater study. These decisions are based on measures of leakance, horizontal conductivity, and vertical conductivity at a bore hole site within a bore hole. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.	Managers in the finance and marketing divisions need to decide which products to keep in their product portfolio. These decisions are based on measures of advertising budget, miscellaneous budget, and target audience for a given product line. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.
Management wants to know the number of different types of construction agencies and the aggregate turnover of drilling contractors and packers. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.	Management wants to know the number of different types of orders and the aggregate insurance amount for small and large orders. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.
A scientist wants to create a report that correlates the repair history of pulser pumps with their ratings. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.	A manager wants to create a report that correlates education level of senior line managers with the number of years of service of employees. Based on the material provided, can you find an answer to the above problem? If yes, describe how you would find the answer. Please be specific.

Appendix C

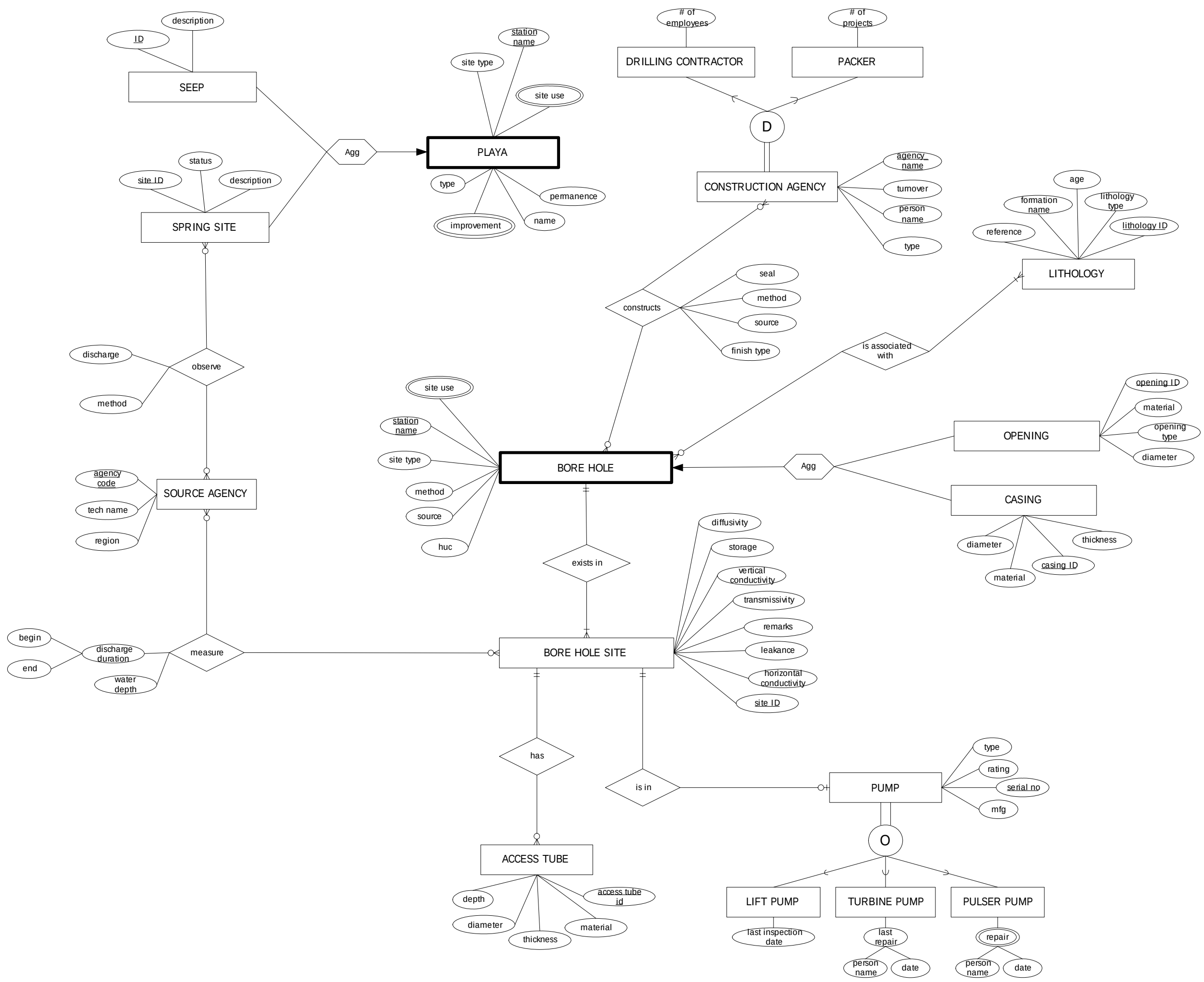
We used a number of strategies to reduce coding error (see, for example, Orwin 1994). First, to ensure that consistent criteria were employed in evaluating all responses, both the coders evaluated each response based on: 1) the extent to which relevant information was derived from the schema (data identification); and 2) how well the response was constructed logically (reasoning). The judgment of quality on data identification and reasoning was made using a four-point scale (1, 2, 3, NA), with 1 indicating a low score and 3 a high score, and “NA” indicating inability to evaluate the response on the criteria.

Second, we used two independent coders with substantive expertise who were naïve to the issues under investigation. The coders were doctoral students in Information Systems and Computer Science, both of whom had substantial experience with conceptual modeling. Neither coder participated in any discussions regarding the project and both were therefore unaware of the constructs under investigation.

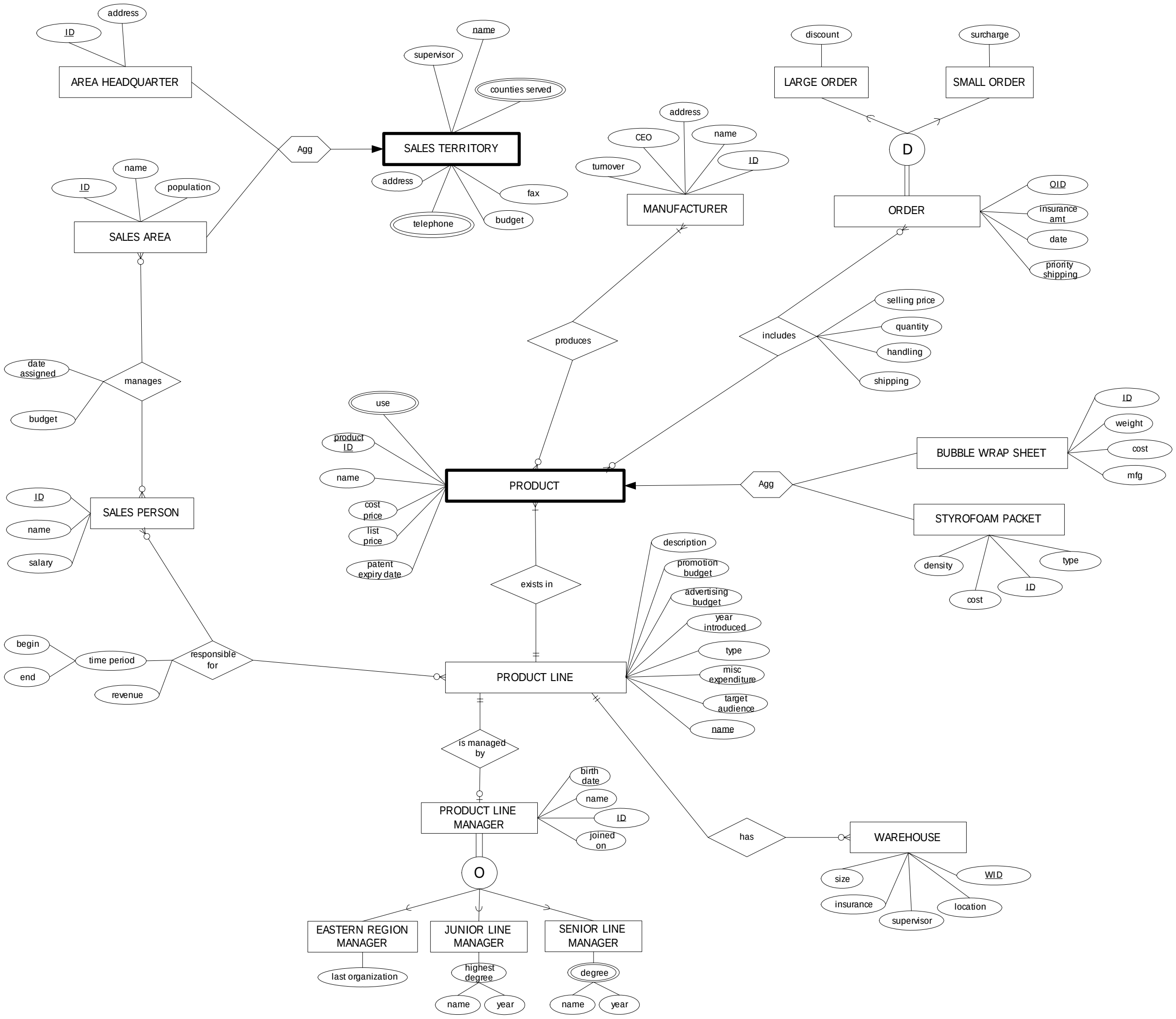
Third, we developed a coding manual, extensively trained the coders, and pilot tested the coding protocol prior to engaging in coding proper. The coding manual provided coders with detailed information on coding procedure, which included details of coder training, the coding itself, and resolution of the coding, as well as providing numerous examples to support the coding scheme. Coder training was monitored by the first author. Before starting to code the study responses, the coders coded sample responses, from both the familiar and unfamiliar domains, which were not part of the study data. On completion of the sample coding, all three persons discussed both their ratings and rationale so that the coders fully understood the nuances of the responses and how they were to be coded.

Fourth, while coder training is assumed to ensure that consistent coding definitions are applied accurately, prior researchers have observed a phenomenon known as coder drift, that is, a tendency to change the way in which coding definitions are applied over time (see, for example, Kazdin 1977). To control for coder drift, the coders coded responses to sample answers to each question prior to coding participant responses. The coders again compared and discussed their responses and the researcher reviewed the overall coding scheme and its application to coding the specific question.

Fifth, following the coding of every 15th answer, the two coders resolved their OQ ratings. If their ratings were not consistent, the coders discussed and tried to arrive at a consensus rating. If the coders could not come to a consensus rating, the researcher assigned codes to such responses. This occurred for just 2 of the 972 responses.



EER Schema for Hydrology Application



EER Schema for Sales Application

Type/Attribute	Description
PLAYA: A concentrated discharge of ground water to the surface that has an extent (i.e., area)	
station name	Name of the surface water station
site use	Purpose for which site was constructed
site type	Type of site, e.g., stream, lake or reservoir, estuary
permanence	Permanence of discharge at spring
name	Name by which the spring is known locally
improvement	Type of improvements constructed at or in association with spring
type	Type of spring, e.g., artesian, fracture, geyser, perched
SOURCE AGENCY: agency responsible for water quality, water level or withdrawal data	
agency code	Code for agency responsible for collection of data
tech name	Name of the person who made site inventory or visit
region	Region for which the source agency is responsible for collecting data
SPRING SITE: Physical point where measurements related to spring are taken.	
site ID	15-digit number assigned to the site
status	The status of the site at the time the water level was measured
description	Description of the spring site
SEEP: Ground water that discharges along hill slope	
ID	10-digit number assigned to a seep
description	Description of seep
observe: Measurement at spring site by source agency	
discharge	Discharge in gallons per minute.
method	The method used to determine the discharge, e.g., flume and estimated
BORE HOLE SITE: Physical point where the measurement related to the bore hole is taken	
site ID	15-digit number assigned to the site
transmissivity	The rate at which water of the prevailing kinematic viscosity is transmitted through the unit width of the hydrogeologic unit under a unit hydraulic gradient
horizontal conductivity	The rate (distance traveled per unit of time) at which water will move through soil or a saturated geologic formation
vertical conductivity	Vertical hydraulic conductivity is expressed in feet per day
storage	The volume of water a hydrogeologic unit releases from or takes into storage per unit surface area of the hydrogeologic unit per unit change in head
leakance	The vertical hydraulic conductivity of the hydrogeologic unit divided by the thickness of the unit
remarks	Significant remark related to test performed
diffusivity	The computed hydraulic diffusivity is expressed in feet squared per day
measure: Measurement at bore hole site by source agency	
water depth	Water level, in feet below land surface, while the well was discharging
discharge duration	Length of time that the well was pumped prior to the measurement of production levels
begin	Start date/time for discharge duration
end	End date/time for discharge duration
ACCESS TUBE: Tube that is used to gain access to a section of the bore hole	
access tube id	Identifier of access tube
material	Diameter of casing in inches
thickness	Material from which casing is made
diameter	Thickness of casing wall
depth	Depth of the access tube
PUMP: An instrument used to remove water from the bore hole	
serial no	Serial number for pump
rating	Horsepower rating of the primary power source
type	Type of pump
mfg	Name of the company that manufactured the pump
LIFT PUMP:	
last inspection date	Last date of inspection of the pump
TURBINE PUMP:	
last repair	Last repair that was done on the pump

Type/Attribute	Description
person name	Person who conducted the repair
date	Date on which the repair was conducted
PULSER PUMP:	
repair	Repairs that have been done on the pulser pump
person name	Person who conducted the repair
date	Date on which the repair was conducted
OPENING: Section of the bore hole, which is open to rock to allow water flow	
opening ID	Identifier for opening
material	Type of material from which the screen or other open section is made
opening type	Type of open section, e.g., fractured rock, mesh screen
diameter	Inside diameter of perforated or slotted pipe
CASING: Section of a bore hole with concrete, steel or plastic installed on the bore hole	
casing ID	Identifier of the casing
thickness	Thickness of casing wall, in inches
material	The material from which the casing is made
diameter	Diameter of section of casing
LITHOLOGY: Gross physical character of a rock or rock formation	
lithology ID	Lithology identifier
lithology type	Principal lithology of unit
age	Age of the rock
formation name	Code identifying the lithologic unit
reference	Source of lithology data
CONSTRUCTION AGENCY: Agency involved in building the site	
agency name	The agency that reported the data
turnover	Turnover of the agency in dollars
contractor	Name of the individual or company that did the work
type	Type of construction agency
constructs	
seal	Type of material used to seal the well against the entry of surface water, e.g., cement and clay, the depth to the bottom of seal below land surface
method	Method by which the site was constructed
source	Indicates who furnished the data
finish type	Method of finish or the nature of openings that allow water to enter the well
DRILLING CONTRACTOR: An agency with primary responsibility of drilling a bore hole	
# of employees	Number of employees
PACKER: An agency that specializes in packer placement in smooth unfractured portions of the borehole	
# of projects	Number of current projects
BORE HOLE: A physical description of the well in the ground	
site use	Purpose for which the site was created
station name	District well numbering system
site type	Type of site
method	Code for method to determine the altitude
source	Information about how the depth of the well was obtained, e.g., driller's log report, depth interpreted from geophysical logs.
huc	The hydrologic-unit code for the Office of Water Data Coordination (OWDC) cataloging unit in which the site is located. This eight-digit code consists of four parts: Hydrographic Region, Subregion code description, Accounting unit within the National Water Data String Network

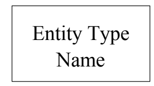
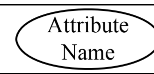
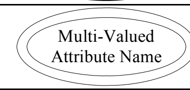
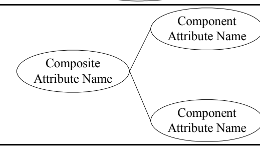
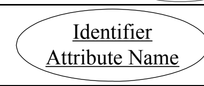
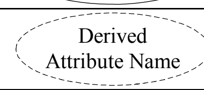
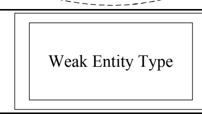
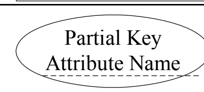
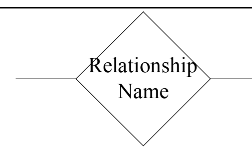
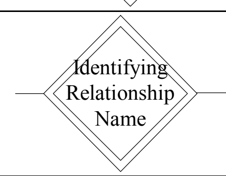
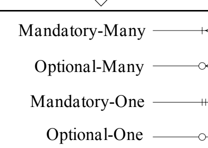
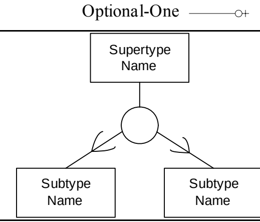
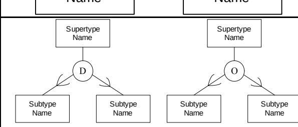
Data Dictionary for Hydrology Application

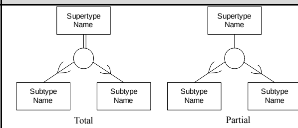
Entity Type/Attribute	Description
SALES TERRITORY: Sales region created by geography.	
name	Name of the sales territory
supervisor	Name of the supervisor for the sales territory
fax	Fax number
counties served	Counties that are served by the sales territory
budget	This year budget for sales territory
telephone	Telephone(s)
address	Address of the sales territory
SALES PERSON: person responsible for product line and sales area	
ID	Code for sales person
name	Name of the sales person
salary	Current salary of the sales person
SALES AREA: Sales region created by geography	
ID	15-digit number assigned to the sales area
name	Name of the sales area
population	Total population of the area served
AREA HEADQUARTER: Office that manages a sales area	
ID	10-digit number assigned to an area headquarter
address	Postal address of the area headquarter
manages: Sales area managed by a sales person	
date assigned	Date on which the sales person was assigned to the sales area
budget	Budget allocated to the sales person for a sales region
PRODUCT LINE: A group of products sharing a common, managed set of features that satisfy the needs of a selected market or mission area	
name	Name of the product line
year introduced	Year in which the product line was introduced
type	Type of product line, e.g., cosmetics, personal care
advertising budget	Amount of advertising budget that is allocated
promotion budget	Amount of promotion budget that is allocated
misc expenditure	Miscellaneous expenditure incurred related to product line
target audience	Primary target audience for the product line
description	Description of the product line
responsible for: Sales person is also responsible for product line	
revenue	Revenue generated by sales person for the given time period for a product line
time period	Time period over which the sales person is responsible for a product line
begin	Start date for time period
end	End date for time period
WAREHOUSE: Facility where the product line is physically stored	
WID	Identifier of warehouse
location	Location of the warehouse
supervisor	Person who is assigned to be the supervisor of the warehouse
insurance	Name of the insurance company
size	Size of the warehouse in cubic feet.
PRODUCT LINE MANAGER: Person who manages the marketing and branding of the product line	
ID	Serial number for product line manager
name	Name
birth date	Date of birth
joined on	Date of joining
EASTERN REGION MANAGER: Product line manager in the eastern region	
last organization	Last organization in which the product line manager was employed
JUNIOR MANAGER: Junior line manager	
highest degree	Highest degree for the junior line manager
name	Name of the degree
year	Year in which the degree was conferred
SENIOR LINE MANAGER: Senior line manager	
degree	Degree for the senior manager

Entity Type/Attribute	Description
name	Name of the degree
year	Year in which the degree was conferred
BUBBLE WRAP SHEET: Bubble wrap packaging for a product	
ID	ID for bubble wrap package
weight	Weight of the bubble wrap package
cost	Cost of bubble wrap package
mfg	Manufacturer of the bubble wrap package
STYROFOAM PACKET: A packet of styrofoam used for packaging a product	
ID	Identifier for styrofoam packet
type	Type of a styrofoam packet
cost	Cost of a styrofoam packet
density	Density of a styrofoam packet
MANUFACTURER: Manufacturer who produces the product	
ID	Identifier for the manufacturer
name	Name of the manufacturer
address	Address of the manufacturer
CEO	Chief Executive Officer
turnover	Last year's turnover of the manufacturer
ORDER: Order associated with the product	
OID	Identifier for an order
date	Date on which an order was placed
insurance amt	Amount for which an order is insured
priority shipping	Additional cost for priority shipping of an order
includes	
selling price	Selling price for the product in an order
quantity	Quantity of a product in an order
handling	Handling charges for a product in an order
shipping	Shipping cost for a product in an order
LARGE ORDER: Large orders	
discount	Discount that is given to large orders, in percentage of the total order
SMALL ORDER: Small orders	
surcharge	Surcharge levied on small orders
PRODUCT: The product that is sold by the organization	
use	Uses of the product
product ID	Identifier for the product
name	Name of the product
cost price	Cost price of the product
list price	Price for which the product is listed
patent expiry date	Date on which the patent of the product will expire

Data Dictionary for Sales Application

Model syntax provided to the respondents

EER Modeling Construct	Description	Graphical Notation
Entity Type	A collection of entities for which common characteristics are to be modeled.	
Attribute	Properties of members of an entity type.	
Multi-valued Attribute	An attribute that can have multiple values for an entity is referred to as a *multi-valued attribute*.	
Composite Attribute	A *composite attribute* is divided into smaller parts; the parts of the composite are called *component attributes*.	
Identifier Attribute	An attribute whose values are distinct for each individual entity is called an *identifier attribute*.	
Derived Attribute	An attribute that can be derived from another attribute is referred to as a *derived attribute*	
Weak Entity Type	Entity types that do not have their own key attributes are referred to as *weak entity types*.	
Partial Identifier Attribute	A weak entity type has a *partial key* that uniquely identifies a weak entity that is related to an *owner entity type*.	
Interaction Relationship	Associations between members of one entity type and one or more other entity types.	
Identifying Relationship	The relationship between a weak entity type and its identifying owner (or strong entity type) is referred to as the *identifying relationship*.	
Cardinality	The number of relationship instances that participate with a given entity instance	
Supertype/subtype	Subtype/Supertype implies that similar objects types (i.e., subtypes) are abstracted to form a higher order objects types (i.e., supertype).	
Disjointness constraint on supertype/subtype	*Disjointness* constraint addresses that question whether an instance of a supertype may simultaneously be a member of a two or more subtypes. Two type: *disjoint* and *overlapping*	

EER Modeling Construct	Description	Graphical Notation
Completeness constraint on supertype/subtype	*Completeness* constraint addresses the question whether an instance of a supertype may also be a member of at least one subtype. Two types: *total* and *partial*.	
Aggregate Entity Type	A type of entities whose members are physically or logically composed of members or sets of members from some other entity type(s) are called an *aggregate* entity type.	

EER Model Constructs