**Association for Information Systems**
**AIS Electronic Library (AISeL)**

SIGHCI 2006 Proceedings

Special Interest Group on Human-Computer Interaction

2006

# Dissecting Query Performance in Logical Data Models: Parsimony vs. Greater Ontological Clarity

Ghalib Al Ma'mri
*The University of Queensland*, g.almamri@business.uq.edu.au

Paul L. Bowen
*Florida State University*, pbowen@cob.fsu.edu

Fiona H. Rohde
*The University of Queensland*

Laurel Yang
*The University of Queensland*, f.rohde@business.uq.edu.au

Follow this and additional works at: http://aisel.aisnet.org/sighci2006

Recommended Citation
Ma'mri, Ghalib Al; Bowen, Paul L.; Rohde, Fiona H.; and Yang, Laurel, "Dissecting Query Performance in Logical Data Models: Parsimony vs. Greater Ontological Clarity" (2006). *SIGHCI 2006 Proceedings*. 10.
http://aisel.aisnet.org/sighci2006/10

This material is brought to you by the Special Interest Group on Human-Computer Interaction at AIS Electronic Library (AISeL). It has been accepted for inclusion in SIGHCI 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

# Dissecting Query Performance in Logical Data Models: Parsimony vs. Greater Ontological Clarity

**Ghalib Al Ma'mri**
The University of Queensland
g.almamri@business.uq.edu.au

**Paul L. Bowen**
Florida State University
pbowen@cob.fsu.edu

**Fiona H. Rohde**
The University of Queensland
f.rohde@business.uq.edu.au

**Laurel Yang**
The University of Queensland

## ABSTRACT

Even when data repositories exhibit near perfect data quality, users may formulate queries that do not correspond to the information requested. Users' poor information retrieval performance may arise from either problems understanding of the data models that represent the real world systems, or their query skills. This research focuses on users' understanding of the data structures, i.e., their ability to map the information request and the data model. The Bunge-Wand-Weber ontology was used to formulate three sets of hypotheses. Two laboratory experiments (one using a small data model and one using a larger data model) tested the effect of ontological clarity on users' performance when undertaking component, record, and aggregate level tasks. The results indicate for the hypotheses associated with different representations but equivalent semantics that parsimonious data model participants performed better for component level tasks but that ontologically clearer data model participants performed better for record and aggregate level tasks.

### Keywords

Ontology, Information Retrieval, Query Performance, Complexity

## INTRODUCTION

Recent advances in technology and the growing interest inorganizational knowledge has resulted in organizations collecting and storing large volumes of data (Nilakanta et al. 2006). Because users extract data to support operational, tactical, and strategic decisions, the data they retrieve must be accurate, complete, timely, and relevant. When users undertake information system tasks, they typically use data models as representations of the corresponding real world systems (Hirschheim et al. 1995). Based on alternative ontological premises, data models of the same real world system can exhibit substantial differences, e.g., in complexity or representational faithfulness. These differences affect users' performance when undertaking various tasks (Bodart et al. 2001; Bowen et al. 2004, 2006; Kim and March 1995; Khatri et al. 2006; Vessey 1991).

Ontological research into conceptual data models indicates that people perform problem solving tasks better using conceptual entity relationship diagrams (ERDs) that evidence higher levels of ontological clarity (Bodart et al., 2001; Burton-Jones & Weber 1998; Gemino, 1998; Gemino & Wand, 2005; Wand et al., 1999). Prior research also provides evidence that ontological clarity and data model size affect end-user query performance (Bowen et al. 2004, 2006). Users' performance query writing is affected by their understanding of the data models that represent the real world systems and by their technical skills.

This research focuses on logical (implementation dependent) data models. It examines whether the prior results arose from the users' understanding of the data models or from their ability to jointly map the information request and the data model during the query writing process.

## THEORY AND HYPOTHESIS DEVELOPMENT

The theory of ontology as formalized by Bunge (1977) and applied to information systems by Wand and Weber (1993), has been a major focus in examining business and conceptual modeling domains (Burton-Jones et al. 1998; Gemino, 1998; Bodart et al. 2001; Gemino and Wand 2005). Bodart et al. (2001) and Weber (2003) assert that removing optional properties from conceptual data models makes them ontologically clearer (hereafter referred to as ontologically clearer data models – OCDM). The majority of the aforementioned research revealed that participants using *conceptual* OCDMs performed better.

Recent research (Bowen et al. 2004, 2006) tested users' data retrieval performance using SQL queries based on logical data models. The Bowen et al. (2006) findings were consistent with prior conceptual data model research which found that participants using OCDMs outperformed participants using PDMs (parsimonious data models – optional properties allowed). Bowen et al. (2004), however, examined larger models and found participants using PDMs outperformed participants using OCDMs.

Ogden's model of query writing (1985) consists of three stages: query formulation, translation, and writing. Query errors could arise from any stage: stage one, developing the information request; stage two, information request mapping with the data model; or stage three, poor query writing skills. This research extends Bowen et al. (2004, 2006) by examining stage two errors for parsimonious and ontologically clearer smaller and larger data models.

**Parsimonious Data Models (PDM) vs. Ontologically Clearer Data Models (OCDM)**

Due to optional attributes and optional relationships, PDMs and OCDMs models yield different numbers of entities and relationships. OCDMs remove optional properties from the data model by using subtypes (Weber, 2003). The removal of optional properties within the data model typically results in an increase in the number of entities and relationships in the model. To formulate queries for OCDMs, users have to reassemble the fragmented data resulting in queries containing more terms and complicated logic. In contrast, PDMs contain optional relationships and attributes. PDMs require less assembly of data but often require the appropriate use of IS NULL/IS NOT NULL in the WHERE clause of an SQL statement or IN or NOT IN sub-queries. Thus, while PDMs are smaller and require less data reassembly, additional complexity is imposed through the challenges associated with the inclusion of optional relationships and optional attributes.

**Information System Tasks**

The query translation process requires a user to take a given information request and "decide what elements of the data model are relevant, as well as the necessary operations" (Siau and Tan, 2006). To undertake the process the user must have an understanding of the model before determining which elements to elicit. Component level tasks measure and evaluate people's overall understanding of structural relationships or cardinalities in the data model. Record level questions are designed elicit the basic objects from the model required to satisfy the information request and some of the basic restrictions required on rows and columns. Aggregate level questions are designed elicit the objects from the model required to satisfy the information request and some of the more advanced restrictions required on sets of rows.

**Complexity, Size, and Performance**

Prior research has indicated that increased complexity has negative effects on performance (e.g., Campbell, 1988; Chan et al. 1998; Siau, et al. 2004). The principles of parsimony (Occam's razor), bounded rationality (Simon, 1957), and minimum description length (Hansen and Yu, 2001; Rissanen, 1978) all imply that, past some point, increases in size lead to impaired performance. Relative to spatial tasks, complexity can be perceived to consist of component complexity, coordinative complexity, and dynamic complexity (Wood, 1986). Relative to the relationship and one subclass for those instances that do not). Within these situations all the information in one model is inferable from the other and thus the two models are informationally equivalent (Siau and Tan, 2006). Making the information explicit may make it easier to locate, however, the OCDM contains more entities which may cause difficulty because of the increased search space. PDMs add extra complexity on end users when exercising exclusion clauses, e.g., IN or NOT IN subqueries. Therefore, from a parsimony perspective, participants using research in this paper, PDMs could be viewed as having greater component complexity, i.e., individual entities are more likely to contain more and more complex attributes. Conversely, OCDMs are likely to exhibit greater coordinative complexity, i.e., increasing clarity by creating subtypes produces more entities which, in turn, requires the query formulation to perform more data reassembly (more joins). These instantiations exhibit different functionalities, strengths, and weaknesses (Siau, 2004).

Impact of Information Representation on Performance

The two models are compared using the theory of informational equivalence (Siau, 2004). Two representations are informationally equivalent if "all the information in one is inferable from the other, and vice versa" (Siau, 2004 p. 77).

Same information and representation in PDMs and OCDMs

When all attributes and all relationships are mandatory, PDMs and OCDMs present the same semantic content and exhibit the same representation. Thus, the two models are informationally equivalent (Siau and Tan, 2006). For some specific information retrieval tasks, if the relevant portions of both the PDM and the OCDM are the same relative to both semantic content and representation then performance differences, if any, arise because of the portions of each model that users must ignore, i.e., mentally discard. From a parsimony perspective, participants using PDMs should perform better, but, from a fineness perspective, participants using OCDMs should perform better. Hence, the following hypotheses are stated in the null form.

H1: If *the relevant portions of the ERDs required to complete the tasks are the same, using PDMs or using equivalent OCDMs will not have a significant impact on users' performance for (a) component level, (b) record level*, or (c) *aggregate level tasks.*

Same Information but Different Representation in PDMs and OCDMs – Optional vs. Mandatory Relationships

Sometimes information required for completing tasks using both PDMs and OCDMs is exactly the same relative to semantic content, but instantiated by different representations within the data model, e.g., due to the removal of optional relationships. Instead of using one entity and an optional relationship as in PDMs, OCDMs transform the optional relationship into two subclasses (one subclass for those instances that participate in the PDM's

PDMs should perform better, but from a fineness perspective, participants using OCDMs should perform better. Hence, the hypotheses are stated in the null form.

H2: *If the relevant portions of the ERD required to complete the tasks are the same in semantic content but different in representation, using PDMs or using equivalent OCDMs will not have a significant impact on users' performance for (a) component level, (b) record level*, or (c) *aggregate level tasks.*

Non-Equivalence of Information in PDMs and OCDMs – Optional vs. Mandatory Attributes

Sometimes the representation provided by PDMs and OCDMs for the same real world situation may not convey the same semantic information. These differences typically arise when the use of optional attributes occurs. By using subtypes to represent optional attributes and thus all possible classifications, OCDMs provide more complete and less ambiguous information. The subtypes improve users' understanding (Weber, 2003) and users are less likely to make erroneous assumptions. As one representation may contain information that is not inferable from the other, the two representations are not informationally equivalent (Siau and Tan, 2006).

Users provided with PDMs have to recognize the additional abstraction role an attribute may have relative to its parent entity. By using subtypes to represent all possible classifications, OCDMs typically provide more complete and less ambiguous information than PDMs. OCDMs, however, exhibit higher relational complexity if the required information is located in more than one subclass. Also, as the size of OCDMs grows, the additional detailed information can cause information overload. Therefore, from a parsimony perspective, participants using PDMs should perform better, but from a fineness perspective, participants using OCDMs should perform better. Hence, the following hypotheses are stated in the null form.

H3: *When OCDMs provide more complete information, using PDMs or using equivalent OCDMs will not have a significant impact on users' performance for (a) component level (b) record level*, or (c) *aggregate level tasks.*

## METHOD

To test the hypotheses two experiments were conducted. Forty business and IT students participated in the first experiment and forty three in the second. Participants received a monetary incentive of AUD $30 to take part in each experiment. Both experiments employed materials based upon domains used in prior experiments (but not with these participants). All participants received the scenario and the ERD and data dictionary for either the PDM or the OCDM.

For each experiment, forty-five information requests were developed. The experimental task required participants to answer a series of component, record, and aggregate level questions via a computer interface. Figure 1 summarizes the allocation of questions for each information task and information representation. Both experiments used a 2 x 2 between subjects design. The data structure, i.e., PDM vs OCDM, was the main treatment. To control for any question order effect, the experiment questions were assigned to one of two orders. The first question order was 1 (H1a), 6 (H1b), 11 (H1c), 16 (H2a), 21 (H2b), etc. The second question order was 1 (H1a), 16 (H2a), 31 (H3a), 6 (H1b), 21 (H2b), etc. The question orders and data structure were randomly assigned to four groups. The order to which participants were assigned to each group was determined by a coin toss.

Each experiment was conducted in four 2.5 hour sessions. The first half hour consisted of a training session designed to familiarize participants with the type of questions and interface. For the remaining two hours, participants answered as many questions as they could via the interface. Each participant's answers were recorded via the interface.

The dependent variable to measure end users' performance was accuracy and was proxied by the participants' percentage scores on each question. The independent variable was the treatment i.e., parsimonious or ontologically clearer. Two covariates, number of IS/IT courses completed and GPA, were included in the analysis.

## RESULTS

Results indicate that, in absolute terms, participants achieved higher scores when using the OCDM (irrespective of size). Within each model, in absolute terms, the participants achieved almost the same score when given the same information and representation, the participants using the OCDM achieved higher scores when given the same information using different representations, and the results were mixed when given the non equivalent information using different representations.

| | Component Tasks | Record Tasks | Aggregate Tasks |
|---|---|---|---|
| Same Information<br>Same Representation | H1a<br>Questions 1-5 | H1b<br>Questions 16-20 | H1c<br>Questions 31-35 |
| Same Information<br>Different Representation | H2a<br>Questions 6-10 | H2b<br>Questions 21-25 | H2c<br>Questions 36-40 |
| Different Information<br>Different Representation | H3a<br>Questions 11-15 | H3b<br>Questions 26-30 | H3c<br>Questions 41-45 |

**Figure 1. Research Model with Question Allocation**

### Same information and representation in PDMs and OCDMs

Comparing the parsimonious with the ontologically clearer group, MANCOVA results reported in Table 1 indicate, for both models, that percentage scores were not

significantly associated with the level of ontological clarity for component (H1a), record (H1b), or aggregate tasks (H1c). Thus, none of these three null hypotheses can be rejected.

#### Same Information but Different Representation in PDMs and OCDMs – Optional vs. Mandatory Relationships

Comparing the parsimonious with the ontologically clearer group, MANCOVA results, reported in Table 1, indicate for both models that the percentage scores were significantly associated with the level of ontological clarity for H2a component, H2b record, and H2c aggregate tasks. The LS means results indicate that participants using the small OCDM achieved marginally significant higher scores than those using the equivalent PDM for H2a component level tasks. The reverse was observed for component level tasks for the larger model. The LS means results for H2b record tasks and H2c aggregate tasks indicate, however, that for both the small and the larger models the participants using the OCDM achieved higher scores than participants using the equivalent PDM.

#### Non-Equivalence of Information in PDMs and OCDMs – Optional vs. Mandatory Attributes

Comparing the parsimonious with the ontologically clearer group, MANCOVA results, reported in Table 1, indicate that for both the small and the larger models the percentage scores were not significantly associated with the level of ontological clarity for H3a component, H3b record, or H3c aggregate tasks. None of these three null hypotheses can be rejected. The LS means results reveal no consistent pattern.

## CONCLUSIONS

The results reveal that participants, in absolute terms, achieved higher scores using the OCDM. Statistically significant differences were detected between participants query performance using the PDM and the equivalent OCDM when same information was presented in different manners. As the two models were considered to be informationally equivalent this result is unexpected. For the smaller model the ontologically clear group performed better for all tasks. The results, however, for the larger model revealed the parsimonious group performed better for the component and record tasks. The results for the larger model also revealed the ontologically clearer group performed better for the aggregate level tasks.

The study extends prior research in end users' query performance by investigating the query translation stage of the query composition process. The results provide insights into how two different representations, PDM and OCDM, affect this stage of query development. These results indicate that some of prior research findings where end users writing queries for larger models made more errors are likely to be attributable to the third stage of the query writing process, i.e., composing syntactically and semantically correct queries to retrieve the desired data.

Limitations of research include the general caveats associated with experiments. Second, the research relies upon stage one of the query writing model (preparation of the information request occurring accurately). The distinction between the tasks (component, record and aggregation) undertaken during the second stage of query writing (query translation) was somewhat arbitrary. Third, the small number of participants reduced the power of the statistical tests. Fourth, the possibility of a participant's non familiarity with the domain may have affected results (Khatri et al. 2006).

Future research opportunities include an investigation of data models that combine both ontologically clearer and parsimonious aspects, i.e, a mixed modeling approach. Also, additional research is needed to test whether users formulating SQL queries perform more effectively using PDMs or OCDMs.

## REFERENCES


1. Bodart, F., Sim, M., Patel, A., and Weber, R. (2001) Should Optional Properties be Used in Conceptual Modelling? A Theory and Three Empirical Tests, *Information Systems Research*, 12,4, pp 384-405.
2. Bowen, P. L., O'Farrell, R.A., and Rohde, F. H. (2004) How Does Your Model Grow? An Empirical Investigation of the Effects of Ontological Clarity and Application Domain Size on Query Performance, in *Proceedings of the 25th ICIS Conference*, Washington DC, December, 77-90.
3. Bowen, P. L., O'Farrell, R. and Rohde, F. H., (2006) Analysis of Competing Data Structures: Does Ontological Clarity Produce Better End User Query Performance, *Journal of the AIS* Volume 7 Issue 8 Article 22
4. Bunge, M. (1977) Ontology I: The Furniture of the World, Treatise on Basic Philosophy, Vol. 3. Reidel.
5. Burton-Jones, A., & Weber, R. (1998) Understanding Relationships with Attributes in Entity-Relationship Diagrams, *ICIS*, December, 214-228.
6. Campbell D. J. (1998) Task Complexity: A Review and Analysis, *Academy of Management Review*, 13, 1, 40-52.
7. Chan, H. C., Siau, K., & Wei, K. K. (1998) The effect of data model, system and task characteristics on user query performance-an empirical study. *The Data Base for Advances in Information Systems*, 29, 1, 31-49.


| | SMALLER | | | | LARGER | | | |
|---|---|---|---|---|---|---|---|---|
| **Source** | **F** | **Pr > F** | **LS Mean** | **LS Mean** | **F** | **Pr > F** | **LS Mean** | **LS Mean** |

| | Value | | PDM | OC | Value | | PDM | OC |
|---|---|---|---|---|---|---|---|---|
| Model* | 13.60 | 0.0001 | | | 10.42 | <.0001 | | |
| Error | | | | | | | | |
| Hytest | 13.25 | 0.0001 | | | 10.44 | <.0001 | | |
| GPA | 28.56 | 0.0001 | | | 0.05 | 0.8147 | | |
| NoCourses | 17.19 | 0.0001 | | | 19.87 | <.0001 | | |
| **Contrast** | | | | | | | | |
| H1a PvsOC | 0.06 | 0.8006 | 0.83 | 0.82 | 0.13 | 0.7189 | 0.68 | 0.69 |
| H1b PvsOC | 0.48 | 0.4873 | 0.64 | 0.68 | 0.88 | 0.3488 | 0.84 | 0.79 |
| H1c PvsOC | 0.47 | 0.4938 | 0.56 | 0.59 | 0.45 | 0.5007 | 0.68 | 0.71 |
| H2a PvsOC | 3.08 | 0.0795 | 0.72 | 0.81 | 4.70 | 0.0304 | 0.89 | 0.78 |
| H2b PvsOC | 16.18 | 0.0001 | 0.41 | 0.61 | 21.52 | 0.0001 | 0.53 | 0.76 |
| H2c PvsOC | 3.86 | 0.0497 | 0.41 | 0.51 | 8.71 | 0.0032 | 0.50 | 0.66 |
| H3a PvsOC | 0.04 | 0.8388 | 0.70 | 0.71 | 1.32 | 0.2511 | 0.90 | 0.85 |
| H3b PvsOC | 2.92 | 0.0876 | 0.53 | 0.61 | 0.54 | 0.4625 | 0.78 | 0.75 |
| H3c Pvs OC | 0.60 | 0.4374 | 0.51 | 0.55 | 1.61 | 0.2044 | 0.59 | 0.66 |

* $R^2$ for the smaller model is 0.2297 and $R^2$ for the larger model is 0.1712

***Table 1. Comparison of Findings for Parsimonious versus Ontologically Clearer Data Models (Accuracy)***


8. Gemino, A. (1998) To Be or May To Be: An Empirical Comparison of Mandatory and Optional Properties in Conceptual Modeling, *Proceedings of the Annual Conference of Administrative Science Association of Canada*, IS Division, Saskatoon, Saskatchenwan.
9. Gemino, A. and Wand, Y. (2005) Complexity and Clarity in Conceptual Modeling: Comparison of Mandatory and Optional Properties, *Data and Knowledge Engineering*, 55, 301-326.
10. Hansen, M.H. and Yu, B. (2001) Model Selection and the Principle of Minimum Description Length, *Jnl of the American Statistical Association*, 96, 454, 746-774.
11. Hirschheim, R., Klein, H., Lyytinen, K. (1995) Information Systems Development and Data Modeling: Conceptual Foundations and Philosophical Foundations. Cambridge University Press.
12. Khatri, V, Vessey, I. , Ramesh, PC and Park, SJ. (2006) Understanding Conceptual Schemas: Exploring the Role of Application and IS Domain Knowledge, *Information Systems Research*, 17, 1, 81-99.
13. Kim, Y. G. and March, S. T. (1995) Comparing Data Modeling Formalisms, *Communications of the ACM*, 38, 6, 103-115.
14. Nilakanta S, Miller L, and Zhu D.(2006)Organizational Memory Management: Technological and Research Issues, *Journal of Database M'ment*, 17, 1, 85-94.
15. Ogden, W. C. (1985). Implications of a cognitive model of database query: Comparison of a natural language, a formal language, and direct manipulation interface. *ACM SIGCHI Bulletin*, 18, 2, 51-54.
16. Rissanen, J. (1978) Modeling by Shortest Data Description, *Automatica*, 14, 465-471.
17. Siau, K (2004) Informational and computational equivalence in comparing information modeling methods. *Journal of Database M'ment*, 15, 1, 73-86.
18. Siau, K.L., H. C. Chan, and K.K. Wei (2004) Effects of Query Complexity and Learning on Novice User Query Performance with Conceptual and Logical Database Interfaces. *IEEE Trans on Systems, Man and Cybernetics, Part A*, 34, 2, 276-281.
19. Siau, K., & Tan, X. (2006). Cognitive mapping techniques for user-database interaction. *IEEE Trans on Professional Communication*, 49, 2, 96-108.
20. Simon, HA (1957) Models of Man, J. Wiley & Sons.
21. Vessey, I. (1991) Cognitive Fit: A Theory-Based Analysis of the Graph Versus Tables Literature, *Decision Sciences* 22, 2, 22:2, 219-240.
22. Wand, Y., and Weber, R. (1993) On the Ontological Expressiveness of Information System Analysis and Design Grammars. *Journal of Information Systems*, 4, 4, 299-330.
23. Weber, R. (2003) Conceptual Modeling and Ontology: Possibilities and Pitfalls, *Journal of Database M'ment*, 14, 3, 1-20.
24. Wood, R.E. (1986) Task Complexity: Definition of the Construct, *Organizational Behavior and Human Decision Processes* 37 pp. 60-82.