# Task Characteristics in Data Modelling

**Shoba Tegginmath**
Auckland University of
Technology,
School of Computing and
Mathematical Sciences,
2-14 Wakefield Street.
Auckland, New Zealand
Shoba.tegginmath@aut.ac.nz

**Natalia Bedran**
Officer at GCCC,
Brisbane Area, Australia
Bedran777@hotmail.com

This quality assured paper appeared at the 2nd annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ2011) incorporating the *24th Annual Conference of the National Advisory Committee on Computing Qualifications*, Rotorua, New Zealand, July 6-8. Samuel Mann and Michael Verhaart (Eds).

## Abstract
An experimental case study on how task characteristics affect student performance was conducted with the objectives of improving data modelling training in academia, and assisting students in achieving skills desired in industry. The research focused on modelling techniques employed in the IT industry as compared to those used in computing and IT education, by investigating differences in student performance in completion of two tasks, comprehension and verification, based on IDEF1X data modelling notational system. The two tasks were found to be interrelated and of equal complexity, although the comprehension task was more time-consuming than the verification task. The research also identified the constructs that were poorly interpreted by students in the comprehension task. The majority of students were able to solve both tasks, revealing both to be appropriate for training and textbooks on data modelling to help students in the acquisition of expert data-modelling skills.

## Keywords
Conceptual modelling, database analysis and design, systems, teaching

## Introduction
Many of the information systems and web applications in use today are based on databases and competence in using an Entity Relationship (ER) diagram is a necessary skill in system design and database

maintenance (Chilton, McHaney and Chae, 2006). The ER diagram (Chen, 1976) remains one of the most widely used data modelling techniques (Moody, 2002) to assist the database designer in visualizing data structures and associations. Other popular techniques are the extended ER model (Teorey, Yang and Fry, 1986), with the added concept of generalization (inheritance) in the modeling technique, and Information Engineering. However academic books available in the higher education market indicate that the Chen notation is the predominant notation (Suleiman and Garfield, 2006) in computing and IT education.

According to some, ER modeling techniques may not be used indefinitely (Chilton et al., 2006). Today however, many universities find the list of possible topics to teach in a database systems course quite large and most can only find room in the curriculum to teach one Database Systems course (Wagner, 2005). There is still a need for ER modelling skills although the application of these skills is moving away from creating new data models from scratch to working with existing data models. Teaching a set of notational symbols for constructs to represent data structures is no longer enough for students. At the same time, studying the syntax and rules of data modelling is also not sufficient. Bock and Yager (2005) recommend the use of examples that result in knowledge extraction by induction as this is more important than memorizing rules.

Software tools automate the data model design process but do not have the capability, based on expertise, to detect and correct conceptual modelling errors. In these situations there is a need for design reviewers who know what to look for. Learning from experts who gain their practical experience in a number of application domains and possess a sort of library of generalised conceptual data models which they reuse as a template (Venable, 1996), there is also support for acquisition of pattern recognition skills in students to help them create data models in new domains and prepare them for real-world software development (Batra and Wishart 2004, Wagner 2005).

A framework for human factors research in data modelling, proposed by Topi and Ramesh (2002) classifies Human, Data Model, and Task categories as independent and control variables, and Performance and User Attitudes as natural dependent variables. The Human category implies user characteristics like education, age, general work experience, intellectual ability, cognitive style, problem-solving approach, training and experience in database, programming, and data modelling. Data Model represents the differences between the data modelling notations. The Task category refers to task characteristics which include task type (understanding, problem solving, and memorization) and task complexity. Performance includes model correctness (the degree to which the model corresponds to a predetermined "correct" solution); time needed to develop the solution and understanding of the notation. User Attitudes is used to describe user's confidence, preference to employ a particular data model, perceived value and ease-of-use of the modelling formalism. The authors (id.) recommended future research on determining the influence of a specific modelling formalism, task types, and individual user characteristics on user performance and attitudes. The authors highlighted a deficiency in "understanding why certain formalisms work well in

some situations and not in others" (Topi & Ramesh, 2002, p.9).

*Modelling Formalism*

Although most data is deceptively complex, developing a simple but comprehensive and accurate data model is paramount (Churcher, McLennan, and McKinnon, 2000). In a learning environment it is important to have a data modelling formalism that is understood and favoured by students. Four extensively used methodologies in the data modelling literature were found to be – Chen Style, Bachman Style, Information Engineering Style, and Martin Style (Pons, Polak, and Stutz, 2006). Although these notational systems are based on underlying commonality, they vary in their salient features such as the use of different relationship notations, the various ways of listing attributes, and special signs to depict cardinality among linked entities. Pons, Polak, and Stutz (2006) examined the differences in student comprehension of ER diagrams while employing these notations; they focused on evaluation, and comparison and contrast of the notational details. Their questionnaire measured understanding of the entities and their attributes, and comprehension of entity relationships. No statistically significant differences in performance among the four data modelling notations were found.

The Extended ER notation although considered an "academic form" has been used in most experimental studies of data modelling understanding and the generalisability of these results is questionable (Moody 2002).

IDEF1X is an information processing standard developed by the US federal government for the design of database schemata. IDEF1X is a high-level data modelling version of the Entity Relationship (ER) modelling approach, used in industrial database applications (Krogstie, Halpin and Siau, 2005). It is considered to be a popular notation (Zikopoulos, Baklarz, Katchnelson and Eaton 2007) used by experts in Industry (Rivero, Doorn and Ferraggine, 2006). Many software tools such as Erwin and Visio support IDEF1X. Therefore IDEF1X was used in this research to test how novices (students) would take to this industry-standard language, considered to be a stable and rigid language (Lankhorst, 2005).

*Task Types*

Previous research has compared different data modelling formalisms and compared novice and expert performance. Except for a few published studies (Kim and March, 1995; Moody, 2002) on how task characteristics play a role in determining performance, relatively little attention has been paid to tasks and task analysis (Zhang, Yeliz & Eseryel, 2005). Data modelling tasks can be classified into four types: (i) Transforming a narrative description into in to a model; (ii) Translating a given model into a narrative problem description (reverse engineering); (iii) Comprehension task – asking questions about a data model; and (iv) Verification task – studying the model for correctness and completeness against business requirements. Tasks used in previous studies have not been found to be representative of tasks that users are required to perform in practice; Most experimental studies have focused on task(i) (Moody 2002). The intention of this study was to look into the relevance and appropriateness of data modelling notation and task type to computing practice. As tasks (i) and (ii) are rarely performed in practice (id.), this study

concentrated on the last two tasks. Thus by studying the types of tasks that are required in today's computing practice, we looked to maximize the external validity of this study.

Unfamiliarity and lack of syntactic and semantic understanding of data modelling constructs lead to problems such as a misuse of data modelling constructs, incorrect statements of understanding, and generalisation structures being ignored. Syntactic comprehension assesses the understanding of the syntax of the language associated with the model (Khatri, Vessey, Ramesh, Clay, & Park 2006) while semantic comprehension means that the statements in the model are understood by users. Experts check whether the data model supports all of the listed domain requirements by verifying the data model. This helps experts to develop complete data models with fewer errors, thereby lowering maintenance costs. In contrast, novices have no practical experience and tend to be reluctant to discover flaws and make changes in their initial data models. Novice designers skip the process of evaluation and improvement of their initial solutions (Batra and Wishart 2004). In other words, novices, or students, are unaware of different kinds of common errors. Novices do not know appropriate quality control techniques and how to apply them.

Students experience difficulties in understanding data modelling constructs and demonstrate an inability to perform quality control over the models well (Venable, 1996). Therefore, we decided it essential to find out whether students find it is easy or difficult to comprehend data modelling elements and perform quality assurance of the model.

The objective performance of students in experimental tasks and their attitude towards the tasks and their own performance were studied. The research questions addressed in this study were:

1. Which task, comprehension or verification, requires more effort to perform?
2. Which task, comprehension or verification, requires more time to complete?

## Methodology

This experimental case study consisted of a series of tests as part of the performance evaluation. The experimental evaluations were conducted within a real-life context, in the Auckland University of Technology (AUT) environment. AUT undergraduate students attending the Logical Database Design course took part in the study. Each class consisted of approximately twenty students. In accordance with Ethics Approval the lecturers for the course were not present in the room while the study was being conducted and all measures to protect the confidentiality of participants were adopted.

The material prepared included: a set of training notes on the IDEF1X data model; two tasks based on IDEF1X data model notation and a description of organizational data requirements; answers for each task; a grading scheme; three questionnaires using Likert-type scales. Questionnaire1 requested demographic and experience information. Questionnaire2 tested comprehension and was administered after students completed the comprehension task. The questionnaire examined student attitudes on basic concepts—how easy/difficult

they found comprehension of entities, attributes and relationships. Questionnaire3 was administered after students completed the verification task. This questionnaire examined students' ability to understand, verify and modify a model to accurately represent documented requirements for the system. Thus students had the opportunity to revise the models and familiarize themselves with aspects that should be checked in order to finish with a good data model.

The training notes were developed using the material from studies by Bruce (1992); Kroenke and Gray (2006); Kusiak, Letsche, and Zakarian (1997); Logic Works (1997); NATO (2002); and Unhelkar (2005). The questionnaires were constructed using several previous studies such as Pons, Polak, and Stutz (2006), but were modified to fit the context of the tasks. The complete set of materials is available from the first author.

A pilot study was conducted two weeks before the actual experiments; the Research and Design Project class was used for the pilot study. The pilot test revealed helpful insights on procedural issues and confirmed that the time for the training and experiments was adequate. As no significant changes were made after the pilot study the results of the pilot study were included in the final data analysis. The actual experiments were conducted during the normal class schedule with five streams.

The paired-sample t-test was used to check for statistical significance. Each student was measured in terms of his/her response to two different tasks. The variables of interest were the number of correct answers and time, measured on an interval scale. Two tasks were performed and two samples relating to the same groups of students were paired.

## Results

The study involved 64 students. The questionnaire results on demographics and answers about students attitude towards the tasks were analyzed to get insights into students' self-reflection on the tasks and difficulties experienced. Overall, the students had some knowledge of programming, building models, writing HTML code, accessing data, developing databases, documenting and solving problems, and they assessed their ability to use software packages (Excel, Powerpoint) at a fairly high level.

The students reported that they did not have difficulties with attributes in the comprehension task which at least suggests that questions about attributes were easy for them to answer. This was confirmed by the actual results of their performance presented in Table 1.

**Table 1: Correctness of the comprehension task answers**

| Construct | % Correct |
|---|---|
| Attribute | 66 |
| Relationship | 55 |
| Categorization | 54 |
| Entity | 50 |
| Cardinality | 45 |

The questions on attributes had the highest percentage of correct answers. It has been found in most empirical studies that "novice designers do not run into much

difficulty in modelling entities and attributes" (Batra & Antony, 2000, p.27). This study confirms that attributes are the most understood constructs of the model by students while cardinality is the most misinterpreted, which is consistent with prior research (Moody et al., 2003).

The verification task had two implicit subtasks: model comprehension and discrepancy checking. Student performance on finding discrepancies between the documented requirements and data model showed that the semantic questions were answered with a lower percentage of correctness (Table 2). Students did not fully understand the meaning of the data embedded in the model. It was easier for students to find syntax errors compared to semantic errors. Practical training that focuses on the semantic understanding of business requirements and their conformance to data models is warranted. Poels et al. (2005) and Nelson and Monarchi (2007) agree that semantic quality is more difficult to evaluate; this study corroborates their statements.

**Table 2: Correctness of the verification task answers**

| Type of Question | % Correct |
|---|---|
| Syntactic | 66 |
| Semantic | 47 |

Questions on the perceived level of overall difficulty of tasks were evaluated against the criteria of time taken to complete and task correctness. Task correctness was measured as the degree to which the student's answers matched predefined "correct" answers. The students found neither task difficult or easy to perform. Understanding of data modelling constructs and performing quality control over the model required an equal amount of students' mental effort.

T-tests were run using SPSS with 0.05 significance level selected. Table 3 shows the results for the number of correct answers for the two tasks - comprehension and verification. The t value of -0.547, with 63 degrees of freedom and a probability of 0.586 is not significant at the 0.05 level. Therefore there is no significant difference in student performance between the comprehension task and the verification task. Scores for the simple tasks are always better than the scores for the most complex tasks (Shoval and Even-Chaime 1987). Therefore this study found the two tasks, comprehension and verification, to be of equal complexity.

**Table 3: Paired samples t-test – Pair 1: Correct Answers**

| | Mean | SD | SE | t-value | Sig |
|---|---|---|---|---|---|
| No. C<br>No. V | -.281 | 4.111 | .514 | -.547 | .586* |

*Not significant at 0.05

Table 4 reports on the time taken for the two tasks - comprehension and verification.

**Table 4: Paired samples t-test – Pair 2: Time taken**

| | Mean | SD | SE | t-value | Sig |
|---|---|---|---|---|---|
| C Time<br>V Time | 5.031 | 13.317 | 1.665 | 3.023 | .004* |

*significant at 0.01

The t value of 3.023, with 63 degrees of freedom and a probability of 0.004 is significant at the 0.01 significance level. There is a significant difference in time required to complete the comprehension and verification tasks. The mean time (in minutes) for the comprehension task was 27.30 and for the verification task 22.27. Thus, significantly more time is required for the comprehension task than for the verification task.

Further, correlation analysis was performed using Pearson product to determine any interrelationships between the comprehension and verification tasks. A correlation of 0.513 between the numbers of correct comprehension and verification task answers was obtained. This is a significant positive correlation at the 0.01 level, indicating a linear relationship between the two variables. Thus, students who scored highly on the comprehension task tended to score highly on the verification task. As noted earlier, the verification task consists of two subtasks: comprehension and discrepancy finding. The comprehension task bore a similarity to the comprehension subtask of the verification task with which the students became familiar after completing the first task. Agarwal et al. (1996) highlighted that task similarity has significant effects on performance. This study also has shown that better student performance in the comprehension task leads to better student performance in the verification task.

## Conclusion

The goal of this research was to discover the difference between student performances on two tasks. This study found a significant difference in completion time for the two tasks studied. This study also found that better student performance in the comprehension task leads to better student performance in the verification task; verification of models directly depends on how models are understood. Inability to understand a model will mean that the verification process can fail.

Neither the comprehension nor verification task was found to be harder to perform. So both tasks can be used in training and textbooks on data modelling in order to help students to acquire expert skills. Use of the two tasks should provide students with marketable, practical skills in performing reviews. The comprehension task helps students make significant advances in their understanding of data modelling constructs. The verification task helps them learn about specific kinds of errors that occur and makes them aware of the importance of quality control techniques.

## References


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

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

Batra, D., & Wishart, N., A. (2004). Comparing a rule-based approach with a pattern-based approach at different levels of complexity of conceptual data modelling tasks. *International Journal of Human-Computer Studies, 61*(4), 397-419.

Bock, D. B., & Yager, S., E. (2005). Using the data modelling worksheet to improve novice data modeller performance. *Journal of Information Systems Education, 16*(3), 341-350.

Bruce, T. A. (1992). Designing quality databases with IDEF1X information models. New York: Dorset House.

Carlis, J., & Maguire, J. (2001). Mastering data modelling. A user-driven approach. US: Addison-Wesley.

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

Chilton, M. A., McHaney, R., & Chae, B. (2006). Data modelling education: The changing technology. *Journal of Information Systems Education, 17*(1), 17-20.

Churcher, C., McLennan, T., & McKinnon, A. (2000). Pragmatic data modelling and design for end users. *In Proceedings of Seventh Asia-Pacific Software Engineering Conference APSEC 2000.*

Khatri, V., Vessey, I., Ramesh, V., Clay, P., & Park, S. (2006). Understanding conceptual schemas: Exploring the role of application and IS domain knowledge. *Information Systems Research, 17*(1), 81-99.

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

Kroenke, D. M., & Gray, C. D. (2006). Toward a next generation data modelling facility: Neither the entity-relationship model nor UML meet the need. *Journal of Information Systems Education, 17*(1), 29-37.

Krogstie, J., Halpin, T., & Siau, K. (2005). Information modelling methods and methodologies. US: Idea Group Inc.

Kusiak, A., Letsche, T., & Zakarian, A., (1997). Data modelling with IDEF1x. *International Journal of Computer Integrated Manufacturing, 10*(6), 470-486.

Lankhorst, M. (2005). Enterprise architecture at work: Modelling, communication and analysis. Germany: Springer.

Logic Works, Inc. (1997). ERwin methods guide. Retrieved May 2, 2007 from http://72.14.253.104/search?q=cache:FEDFRWM2az8J:tangra.si.umich.edu/~radev/654/resources/ERMGALL.PDF+Erwin+methods+guide.&hl=en&ct=clnk&cd=1&gl=nz.

Moody, D. (2002). Complexity effects on end user understanding of data models: An experimental comparison of large data model representation methods. In *Proceedings of* ECIS 2002 June 6–8, Gdańsk, Poland.

Moody, D. L., & Sindre, G. (2003). Incorporating quality assurance processes into requirements analysis education. *Proceedings of the 8th annual conference on Innovation and technology in computer science education*.

Moody, D. L., Sindre, G., Brasethvik, T., & Solvberg, A. (2003). Evaluating the quality of information models: Empirical testing of a conceptual model quality framework. *Proceedings of the 25th International Conference on Software Engineering ICSE '03.*

NATO (2002). Data model documentation: IDEF1X data model diagram.

Poels, G., A. Maes, et al. (2005). Measuring user beliefs and attitudes towards conceptual schemas: tentative factor and structural equation model, Citeseer.

Pons, A. P., Polak, P., & Stutz, J. (2006). Evaluating the teaching effectiveness of various data modelling notations. *The Journal of Computer Information Systems, 46*(2), 78-84.

Rivero, L. C., Doorn, J. H., & Ferraggine, V. E. (2006). Encyclopaedia of database technologies and application. US: Idea Group Inc.

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

Suleiman, J., & Garfield, M. J. (2006). Conceptual data modelling in the introductory database course: Is it time for UML? *Journal of Information Systems Education, 17*(1), 93-99.

Teorey, T. J., Yang, D, and Fry, J. P. (1986). A logical design methodology for relational databases using the extended Entity-Relationship model. *Computing Surveys 18*(2), pp. 197-222.

Topi, H., & Ramesh, V. (2002). Human factors research on data modelling: A review of prior research, an extended framework and future research directions. *Journal of Database Management, 13*(2), 3-19.

Unhelkar, B. (2005). Verification and validation for quality of UML 2.0 models. US: John Wiley.

Venable, J. R. (1996). Teaching novice conceptual data modellers to become experts. *In Proceedings of International Conference on Software Engineering: Education and Practice,* IEEE.

Wagner, P. (2005). Teaching data modelling: Process and patterns*. ITiCSE*, *Proceedings of the 10th annual SIGCSE conference on innovation and technology in computer science education*.

Zhang, P., & Yeliz Eseryel, U. (2005). Task in HCI research in the management information systems (MIS) literature: A critical survey. *In Proceedings of the 11th International Conference on Human-Computer Interaction.* July 2005.

Zikopoulos, P., Baklarz, G., Katchnelson, L., & Eaton, C. (2007). IBM DB2 version 9 new features. New York: The McGraw-Hill Companies.