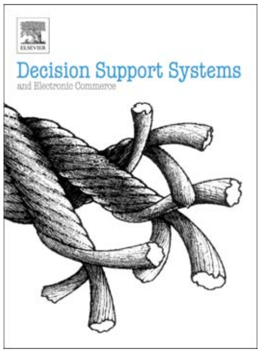
12-1-2011

Comparing the Understandability of Alternative Data Warehouse Schemas: An Empirical Study

David Schuff
Temple University

Karen Corral
Boise State University

Ozgur Turetken
Ryerson University



Comparing the Understandability of Alternative Data Warehouse Schemas: An empirical study

David Schuff, Karen Corral, Ozgur Turetken

PII: S0167-9236(11)00104-7
DOI: doi: 10.1016/j.dss.2011.04.003
Reference: DECSUP 11893

To appear in: *Decision Support Systems*

Received date: 7 November 2009
Revised date: 9 February 2011
Accepted date: 11 April 2011

Please cite this article as: David Schuff, Karen Corral, Ozgur Turetken, Comparing the Understandability of Alternative Data Warehouse Schemas: An empirical study, *Decision Support Systems* (2011), doi: 10.1016/j.dss.2011.04.003

Comparing the Understandability of Alternative Data Warehouse Schemas: An empirical study

David Schuff*
Department of Management Information Systems
Fox School of Business
Temple University
207G Speakman Hall
1810 North 13th Street
Philadelphia, Pennsylvania 19122
Tel: 215-204-3078
david.schuff@temple.edu

Karen Corral
Department of Information Technology and Supply Chain Management
College of Business and Economics
Boise State University

Ozgur Turetken
Ted Rogers School of Information Technology Management
Ryerson University

* *corresponding author*

ABSTRACT

An easily understood data warehouse model enables users to better identify and retrieve its data. It also makes it easier for users to suggest changes to its structure and content. Through an exploratory, empirical study, we compared the understandability of the star and traditional relational schemas. The results of our experiment contradict previous findings and show schema type did not lead to significant performance differences for a content identification task. Further, the relational schema actually led to slightly better results for a schema augmentation task. We discuss the implications of these findings for data warehouse design and future research.

Keywords: Data warehousing, schema understandability, experiment, cognitive effort

COMPARING THE UNDERSTANDABILITY OF ALTERNATIVE DATA WAREHOUSE SCHEMAS: AN EMPIRICAL STUDY

1. Introduction

The data warehouse, the core tool in a business intelligence strategy, continues to increase in importance within the information technology function. According to IDC, data warehouse platform software and service sales were up 12% in 2008 to $7.6 billion, and projected to continue growing at a rate of 7.4% annually [34]. Gartner also predicts that data warehouses in industries such as telecommunications, retail, and distribution will grow in size to hundreds of terabytes [5].

There are two predominant designs used to build these large information stores: the relational model and the dimensional model. Warehouses built using either model can be used to deploy an organization's data in a form ready for analysis by its users (i.e., a set of integrated and "cleansed" data from multiple data sources). The difference between the relational and dimensional models is in the structure of the logical schema used to represent each. Relational models are represented by the traditional relational schema, while the dimensional model is represented using a variant called the "star schema" (so called because of its appearance). These schemas are the logical models derived from the conceptual (ER) model, and are the mechanism by which users understand the structure of the database.

It has been asserted that the structure of a star schema, with its focus on business "facts," is a simpler representation than that of a traditional relational schema; so much so that the dimensional model is advocated as "the only viable technique for designing end-user delivery databases" [23]. However, there is controversy surrounding this assertion (e.g., see [6, 21]). The main criticism of the star schema is that it is overly restrictive because its structure forces the data warehouse designer to choose a narrow focus (sometimes even a single subject). It is often difficult to retrieve data not related to that original focus. The result is a data warehouse optimized for some users to the exclusion of others [18]. Haughey [15] contends that "the world is not a star" and that star schema cannot effectively reflect complex business scenarios. Additionally, Jukic [21] proposes that model choice is a complex issue, and is essentially a tradeoff between simplicity and flexibility. As these researchers suggest, the star schema may not be as semantically accurate as the traditional relational schema in certain cases, such as when direct relationships between some of the dimensions exist. However, for many practical problems, it is possible to create semantically equivalent alternative schemas using each approach. There are technical reasons for selecting one method over the other [21], but in this study, we will only be examining the impact of the data model on understandability of the data warehouse where the content that is contained in the data warehouse is equivalent regardless of the way it is presented.

Data models and their associated schemas are powerful communication tools that are used between users and analysts and especially between analysts and designers [33]. Data models represent the structure of the database and the data available within it [24]. There are several scenarios where it would be beneficial for business users, who typically are not technology experts, to understand a database model as operationalized through a schema. One example is during the systems analysis and design process, where verifying the validity of a database schema is an essential step in application development. A database designer can gather initial data requirements from the business users, but showing those users the resulting schema is a way of making sure that these requirements have been properly understood. Essentially, end users "sign off" on the schema, indicating that it meets the data requirements of the application. In situations where the schema cannot be understood, the client must rely on the developer's explanation of what is contained in the database, which may be subject to the same threats of misinterpretation as the original requirements gathering process. In addition, integrating the end user into the development process has ongoing practical significance as new data become available in the form of new data entities and relationships that can be added to the warehouse. The ability of business users to suggest modifications to the data warehouse according to the changing data needs of the organization is greatly aided by how well they understand what data the current version of the warehouse provides.

Understandability of the schema also enables end users to more significantly contribute to the structural changes in the data warehouse when faced with the changing needs of the organization. Ideally, those non-experts that are closest to the business should be able to determine whether a data warehouse fulfills existing business needs and suggest modifications that the IT department can implement. The ability of a schema to facilitate such “graceful extensibility” [23] has been suggested as an important characteristic in data warehouse design.

A second scenario where it is useful for business users to understand the schema is situations where the user can query the database directly. Pre-canned reports and interfaces are not adequate for all users [22]. For example, when end users’ data needs are not static, simply being able to use a query-by-example (QBE) tool is more expedient than construction of additional applications to facilitate the data retrieval. This ultimately increases the flexibility afforded to users in their interactions with the warehouse. QBE tools are common in cube browsing products such as Cognos’ Business Intelligence or Microsoft’s Analysis Services.

The objective of this paper is to provide additional insight as to whether the underlying schema of a data warehouse (star or traditional relational) affects the understandability of that data warehouse. Since prior theory provides conflicting guidance regarding the superiority of one schema over the other, we take an exploratory approach. Through two controlled experiments, we compare the relative understandability of the traditional relational schema to the dimensional schema. Previous studies have addressed schema comprehension using recall as a surrogate metric for understandability [7,35] to compare these schemas [9]. In this study, we build upon this research by conducting an empirical investigation that compares these two alternative schemas through two tasks that are more involved than simple recall.

2. Background

2.1 Evaluation of data model understandability

According to McGee, “in order for a data model to be used, it must be understood” [26, p. 372]. He identifies three properties of data models that enhance their ability to be learned and understood: (1) *simplicity* refers to the number of structure types (e.g., tuples and relations) and the number of rules that govern the assembly of those structure types, (2) *elegance* describes the ability to create the model using the smallest number of structure types, and (3) *picturability* is the degree to which the model lends itself to a visual representation.

When comparing graphical representations to tabular representations, past research hypothesized graphical models to be advantageous for comprehension due to the additional semantic meaning conveyed by a drawing. Graphical semantic models have been associated with higher levels of comprehension [20] and graphical representations with simpler graphic styles (for example, lists within graphic elements instead of separate graphic elements for each item) have been found easier to interpret [28]. Other research has found that a model is more easily learned if it has greater syntactical clarity [19]. In a comparison of the extended E-R model (EER) to the tabular relational model, users could more effectively model relationships using the EER model because its lower semantic distance (i.e. how close the meaning of the diagram's components represent the constructs that are modeled) more clearly conveys relationships among entities [1]. Evidence from the prior literature supports McGee's assertion [26] that the most effective modeling techniques are those that are graphical and simple while describing all of the structure types.

2.2 The traditional relational schema versus the star schema

The traditional relational schema and the star schema are both logical data models. They differ in two important ways: the selection of tables, and the way in which the relationships are constructed between those tables. For example, consider two simple schemas (modeling scripts) for an airline reservation database. The first is a traditional relational schema (Figure 1) and the second is a star schema (Figure 2). These two schemas highlight that the different models can be equivalent from an information-content perspective.

<< ==== INSERT FIGURE 1 HERE ==== >>

The relationships in a traditional relational schema are constructed based on the logical relationships between these tables, but without specific emphasis on any one table or relationship. In other words, there are no structural rules defining the organization of the relationships between the tables. The star schema's structure is more constrained – it is based on a set of relationships between descriptive tables and a central table that represents the subject of the database (a reservation). A series of one-to-many relationships exist between the central "fact" table and the associated dimensions. Essentially, Figure 2 describes a reservation as a particular flight, with a particular passenger, on a particular airline, at a particular time.

<< ==== INSERT FIGURE 2 HERE ==== >>

Cognitive science provides some theoretical guidance suggesting a difference in understandability between the relational and star schema diagrams. Semantic network theory states that humans store concepts in memory as linked units [1,8]. Therefore, the representation of a collection of objects as a semantic network should be intrinsically easy for people to understand. Further, prior research has shown that the structure of human memory is organized into "chunks" which serve to increase memory capacity (e.g., [2,24,27]). This organization stores not only the data elements themselves, but also the relationships between the elements that are stored. The implication of this for comprehension of data models is that models that organize their elements into logical groupings (chunks) with clear associations between those elements will be more intuitive and therefore easier to understand.

Looking at these alternative schemas in Figures 1 and 2 using McGee's criteria of simplicity, elegance, and picturability [26], it can be argued that simplicity and elegance of the two models in these examples are comparable. They both use the same components (tables, attributes, and cardinality notation)[1], and therefore have the same number of structure types. However, the two schemas differ with regard to picturability. The star schema is able to convey more clearly than the traditional relational schema its most important information. Not only is the user able to see the database's structure through the positioning of its entities and relationships around the fact, but the visual centrality of the fact within the diagram makes clear the subject of the database. Because multiple dimensions link back to the same fact (a "reservation," in the airline example), the fact itself is reinforced. This grouping of dimensions and the fact create a chunk that visually conveys relationships within the schema.

Previous research has found evidence to support this. Drawing primarily on the concepts of semantic network theory and chunking, the star schema pattern would be easier for users to recall [9]. In a

[1] There are several accepted standards for representing the relationships between table schemas. Both the traditional relational schema and the star schema can be created using any of these standards.

lab experiment, subjects could recall a star schema diagram more accurately than an equivalently complex diagram of a traditional relational schema. Subjects also recalled the star schema in a pattern consistent with the semantic meaning of the diagram. When reconstructing the diagram, subjects first recalled the fact table, followed by its surrounding dimensions. This implies that the focus of the warehouse (the fact) was reinforced by its associated dimensions.

However, it is less clear whether or not the advantage of the star schema is scalable, and would carry over to more complex models. There are two reasons for this. First, while chunking can enable people to process more information at once [27], this capacity is still limited. Given a more complex data warehouse with many dimensions, the benefits of a star schema's presentation may be diminished by the sheer number of elements. Second, as the schemas become increasingly complex, the difference in picturability is likely to become less pronounced. A complex data warehouse typically consists of several smaller star schemas that share a common (conforming) dimension. In that case, the schema will have several foci, making the diagram more complicated to understand.

As an example, compare the relatively simple schemas in Figures 1 and 2 to the more complex schemas used for the experiment in our second study (Figures 3 and 4). The difference in picturability of these two diagrams in Figures 3 and 4 appears to be much less pronounced than in the simpler schemas of Figures 1 and 2. While there is still no focus in the traditional relational schema, the star schema now has five foci (Internship, Club Membership, Job Offer, Enroll, and Application). Further, at least visually, the conforming dimension (Student) becomes a sixth focal point of the schema.

<< ==== INSERT FIGURE 3 HERE ==== >>

<< ==== INSERT FIGURE 4 HERE ==== >>

Therefore, given the lack of a definitive theoretical rationale, there is an open question as to whether the structural advantages of the star schema truly exist. They may actually diminish significantly when the schema is complex, and therefore have limited advantage under realistic, enterprise-wide scenarios. We have constructed two studies that specifically address that issue. Our studies use complex schemas and tasks that go beyond the recall of a simple model to test subjects' comprehension of the underlying data model. In the next section, we develop our hypotheses and describe the studies that compare the star and traditional relational schemas.

3. Hypothesis Development

We put forward a series of hypotheses to test whether the star schema will differ from the traditional relational schema on key evaluation metrics with regard to understandability. Gemino and Wand [11] make a distinction between model comprehension and understanding, where the latter requires an understanding of the modeled domain in addition to the grammar, and assert that *understanding* (which is more inclusive than comprehension) should guide the choice of dependent variables in empirical studies. Topi and Ramesh [33] list “user performance” and “attitudes” as the two major categories of dependent variables in studies that evaluate data models. Because of its more objective nature we chose user performance as our surrogate for user understanding, which would be reflected in both the quality of end users’ responses to experimental tasks and their effort expended to complete that task. Outcome quality (e.g., [4,10,11,16,17,25,29,31,32,33]) and effort (e.g., [4,10,11,29,31]) are commonly used indicators of success in information presentation studies, and most closely resemble the model correctness and time variables in Topi and Ramesh’s categorization [33]. People use decision aids to reduce the cognitive effort they expend when performing a task [4]. A reduction in effort can serve as a measure of success, especially when there is not a corresponding reduction in performance [29]. The more effective presentation of information can led to both a reduction in users’ effort and a simultaneous increase in task performance [29]. Therefore, by considering both performance and effort, we can arrive at a richer measure of overall success.

In the previous section we contend that the advantages of the star schema may not exist when the schema becomes complex. We conducted two studies that aim to accurately represent the understanding and problem solving tasks one might perform when working with a complete, realistic schema (with tables, attributes, and cardinality notation). The first study employs a content identification task, where subjects are required to determine whether a query can be answered by a given schema. Our second study requires subjects to augment an existing schema by adding additional entities and relationships. As we discussed earlier, these tasks have face validity as ways of measuring understandability of the underlying data model. Task choice is also important because the fit between task and technology are key influences on task success [12]. For example, Yang [36] found that the success of CASE tools depended upon their fit with the organizations existing development methodology. The type of task the user performs may influence the effectiveness of a particular diagram type (in this study, a database schema). In the context of systems analysis, Hahn and Kim [16] found that effective diagrams support the cognitive processes associated with the user's task. Due to this potential influence of task on outcomes, the representativeness of the experimental tasks to those that users actually perform is important for the task-dependent nature of the results to be practical.

Since the direction and magnitude of the effect of schema type on the outcome variables is unclear, we take an exploratory approach to the problem. Because of the lack of strong theory to suggest the superiority of one diagrammatic representation over the other, we hypothesize that an effect exists but do not specify the direction. The question of whether there is a difference in understandability between the star schema and the traditional relational schema is tested through the following hypotheses:

H1: The type of schema (traditional relational or star schema) will affect the score subjects receive on the content identification task.

H2: The type of schema (traditional relational or star schema) will affect the effort subjects expend on the content identification task.

H3: The type of schema (traditional relational or star schema) will affect the score subjects receive on the schema augmentation task.

H4: The type of schema (traditional relational or star schema) will affect the effort subjects expend on the schema augmentation task.

Support for these hypotheses indicates evidence of a difference in understandability, in line with the conventional wisdom regarding these schema types (Kimball 1996). A lack of support would suggest that this difference may not exist; this potential implication is also interesting from a theoretical standpoint.

The results of these studies should provide data warehouse designers new insights as to whether the practical claims on the superiority of one particular schema over the other are valid, and whether the implications of the earlier empirical studies [9] on the subject are generalizable to more complex models and more complex tasks.

4. Study One

4.1 Subjects, task, and procedure

The participants in the first study were 205 undergraduate Management Information Systems students. Their average age was 23.18 years and 48% of the sample was female. Because the task involved interacting directly with a schema, we recorded the experience of the subjects with databases and database models for control purposes. Experience was modeled as a categorical, binary variable (experienced or inexperienced) based on whether the subjects had completed an introductory database design and management course. The course covered database use, SQL, and schema design using the traditional relational model.

Therefore, the study employed a 2x2 between subjects design (with schema type and experience as factors), and involved a content identification task. Through a web-based tool, subjects were shown either a traditional relational schema or a star schema diagram. The subject domain of both data warehouse schemas was a hypothetical university (see Figures 5 and 6), and the schemas contained the same information. Subjects were given a series of ten English-language questions (see Appendix A), and then were asked whether or not the question could be answered based on the information given in the schema. Consistent with the notion of understandability [11], in order to answer the questions correctly the diagram must successfully convey both the grammar and subject domain of the model. The performance score was computed by simply totaling the number of correct answers for each subject.

<< ==== INSERT FIGURE 5 HERE ==== >>

<< ==== INSERT FIGURE 6 HERE ==== >>

After the task was completed, each subject completed a questionnaire in which they assessed the level of effort they expended while performing the task using the NASA/TLX (Task Load Index) instrument [14] (see Appendix B). This instrument has been used in previous information systems studies to measure workload [e.g., 13,30,31]. To complete the instrument, subjects must pair-wise compare six dimensions of effort (mental, physical, temporal, performance, frustration, and overall effort), each time selecting the one that contributed more to the effort expended completing the task. The subject then assesses the overall level of effort demanded on each dimension (on a scale from 1 to 7). The number of times each dimension of effort was selected in a pair-wise comparison is multiplied by its overall level in order to arrive at a weighted measure of perceived effort.

4.2 Results of study one

Because the dependent variables in the study (score and effort) were not significantly correlated (using Pearson's correlation test, p=0.875), we constructed two separate ANOVA models. Schema type and experience (with data modeling) were the independent variables for both models. The descriptive statistics for score are provided in Table 1. As seen in Table 2, neither the main effect of schema type (p=0.526) nor the interaction between model type and experience (p=0.181) is significant. Therefore, there was insufficient evidence to support hypothesis 1 (a difference in schema type in terms of score).

The results of the analysis for hypothesis 2 are shown in Tables 3 and 4. As seen in Table 4, the schema type (p=0.833) and the interaction between schema type and experience (p=0.397) have no significant effect on effort, therefore there is also insufficient evidence to support hypothesis 2 (a difference in schema type in terms of effort expended).

<< ==== INSERT TABLE 1 HERE ===== >>

<< ==== INSERT TABLE 2 HERE ===== >>

<< ==== INSERT TABLE 3 HERE ===== >>

<< ==== INSERT TABLE 4 HERE ===== >>

The results also indicate that experienced users have an overall advantage as we see a significant main effect of experience on both dependent variables (score and effort) favoring those experienced users.

This finding still corroborates the hypotheses testing results as subjects with the same level of experience had similar levels of performance *regardless of the schema type they were given*. The power of our test was 0.7 for a medium effect size, making it unlikely that our inability to find significant differences with the score is due to a lack of power. Similarly, the power of the test for effort is 0.99 for a small effect size. This provides compelling support for the conclusion that this lack of difference was due to the practical equivalence of the understandability of these two diagram types.

5. Study Two

5.1 Subjects, task, and procedure

The participants in the second study were 95 undergraduate Management Information Systems students (not the same students who participated in the first study). They had an average age of 24.16 years, and 41.1% of the sample was female. As in the first study, we controlled for the effect of subjects' familiarity with data models based on whether they had completed an introductory database design and management course.

As the first study, this experiment employed a 2x2 between subjects design with schema type and experience as factors. Subjects were once again given either a traditional relational schema or a star schema diagram (that contained the same information) of a data warehouse for a hypothetical university and a textual description of the scenario (see Figures 3 and 4). The scenario asked the subjects to imagine that they were database designers for a fictitious business school (see Appendix A). They had to modify the existing data warehouse to track student participation in student clubs and professional societies. The instructions listed specific information to be captured by the data warehouse, but not information regarding specific tables or the relationships between them. The subjects were allowed to either draw directly on the database diagram or on a separate piece of paper. As with the first study, this task was designed to test understanding as both subjects' mastery of the diagram's grammar and subject domain were needed to successfully augment the schema. As the first study, each subject completed a questionnaire assessing the level of effort they expended while performing the task (see Appendix B).

The diagrams given to the subjects were missing the tables related to the task (tracking club membership). Because the schemas were equivalent with regard to information content, the task of completing the missing portion of the schema was similar regardless of the diagram. In order to calculate task scores, each response was compared to an "ideal" solution. The ideal solution was the simplest way to fulfill the requirements of the task, but this was not the only solution that would be considered correct. A response was considered correct as long as it was consistent with the guidelines set forth in the task instructions. Points were deducted if there were components of the response that were incorrect, such as missing or mislabeled tables and attributes, or missing or incorrect relationships.

Each element of the diagram – tables, attributes, relationships, and cardinality – was evaluated separately on a scale of 1 ("completely incorrect or missing") to 4 ("completely correct"). A scale was used (instead of a simple "correct/incorrect" evaluation) because it provides a higher degree of differentiation between responses. It is possible that a subject might have included an element but not included it correctly (which would get rated a "2" or a "3"). For example, if a relationship should have been drawn from table A to table B, but instead it was drawn from table A to table C (and this was not correct given the rest of their solution), they would receive a score of 2 ("included but incorrect"). If there was no relationship drawn at all, they would receive a score of 1 ("completely incorrect or missing"). If they drew a relationship between table A and B (correct), and then between table A and C (incorrect), they would receive a 3 ("mostly correct"). Because there were a different number of responses for each category (e.g., the diagram had more attributes than tables) the scores for each element type (tables, attributes, relationships, and cardinality) were normalized to 25 points. The four normalized scores were summed to arrive at an overall score (out of 100). This scoring method is similar to what was done in earlier empirical research in the area, for example, the scoring based on "facets" as described in Batra et al. [3].

Different pairs of the authors coded the diagrams separately. While this introduces the possibility of experimenter bias, a predetermined key was used in order to mitigate this effect. We believe the high agreement among the two sets of ratings (0.89) confirms this. Further, the authors evaluated only the technical correctness of the solutions rather than their subjective quality. Due to the high rater agreement, the two scores for each subject were averaged to calculate the task score.

5.2 Results of study two

As in the first study, the data were analyzed using two ANOVA models, because score and effort were again not significantly correlated (using Pearson's correlation test, p=0.653). Schema type and experience were independent variables for both models (for descriptive statistics see Tables 5 and 9). To test hypothesis 3, the model was built using score as the dependent variable. The results show that although the main effect of schema type is not significant (p=0.667), there is a significant interaction effect between schema type and experience (p=0.013, see Table 6 and Figure 7). To further examine the nature of this interaction, the effect of the schema type on score was tested separately for experienced and inexperienced subjects. As seen in Tables 7 and 8, experienced subjects did significantly better with traditional relational schema diagrams ($score_{TRS}$>$score_{SS}$, p=0.040; see Table 7) while inexperienced subjects appear to have done better with the star schema diagrams ($score_{SS}$>$score_{TRS}$, p=0.148, see Table 8) although the result for the inexperienced subjects is not significant. Therefore hypothesis 3 (a difference in schema types in terms of score) is partially supported. As for hypothesis 4, the results on effort are significant in favor of the traditional relational schema ($effort_{SS}$>$effort_{TRS}$, p=0.022, see Tables 9 and 10). Therefore hypothesis 4 (a difference in schema types in terms of effort expended) is supported.

<< ==== INSERT TABLE 5 HERE ====>

<< ==== INSERT TABLE 6 HERE ====>

<< ==== INSERT TABLE 7 HERE ====>

<< ==== INSERT TABLE 8 HERE ====>

<< ==== INSERT TABLE 9 HERE ====>

<< ==== INSERT FIGURE 7 HERE ====>

<< ==== INSERT TABLE 10 HERE ==== >>

These results indicate that all subjects (experienced or inexperienced) given the traditional relational schema expended less effort in completing the task than those given the star schema. Experienced subjects performed better with the traditional relational schema – those who were given that schema received a higher task score than those given the star schema. A possible explanation for this is that the experienced subjects are a group much more likely to have had experience with the traditional relational schema. Completion of the course used as criteria to classify subjects as experienced was heavily based on that schema. These subjects were more successful (performing better on the task while expending less effort) with the diagram with which they were more familiar. In addition, the increased role experience plays in modeling, as compared with simply retrieving information from a database, may have further accentuated the impact of prior modeling experience on their performance with the traditional relational schema.

6. Discussion

The purpose of the two studies reported in this paper was to determine whether the star schema differed from the traditional relational schema with regard to its understandability. This is an important step in demonstrating the relative effectiveness of these schemas as a delivery mechanism of data to end users. There was evidence to support this basic notion in previous studies, and we have expanded upon that work by conducting two controlled experiments, which required subjects to understand the semantic content of the schema to effectively perform the tasks. The results from the two studies are summarized in Table 11.

<< ==== INSERT TABLE 11 HERE ===== >>

We found evidence that the differences in understanding for the two schema types are task-dependent [12]. In the first study (which involved a content identification task), no differences were found with regard to either performance or effort expended, whether the subjects were given the star schema or the traditional relational schema. This finding is interesting because the formulation of queries is representative of the type of tasks typically performed by a data warehouse user. The retrieval of data from a warehouse is consistent with Kimball's [23] view of the star schema as a delivery mechanism of data to end users. However, the star schema appears to be no better than the traditional relational schema in enabling users to formulate queries.

For the schema augmentation task (the second study), users who were more experienced with data modeling appeared to do better (while still using less effort) when given the traditional relational schema. This may simply be a reflection of their course-specific experience with that schema. It is possible that if this group had experience with the star schema in their course instead of the traditional relational schema, the experienced group would have favored the star schema. Therefore, what is most interesting is that inexperienced users did not have significantly different performance levels when using the different schema types. The lack of a difference found for these inexperienced users provide, at best, mixed evidence of a difference between the schema types. Since most end users of data warehouses are likely to be unfamiliar with data modeling, the inexperienced group is more representative of the typical end user.

The results of the experiments provided conflicting evidence to the results of the Corral et al. study [9] (where the task required recall of the schema) as to what "technology" (i.e. the underlying schema used in design) best supports these tasks. In that context, the star schema appears to aid a simple task such as recall [9], but these benefits do not appear to translate to the more complex tasks used in this paper. In the studies presented here, the results suggest the advantage of the star schema is not scalable. Recall remains a good first step, providing evidence regarding the understandability of a diagrammatic representation. However, further studies (such as this one) regarding whether this manifests itself in an improvement in task performance can provide additional insight in model comprehension in general.

As with any research, this study has limitations. First, the use of student subjects may limit the generalizability of the findings. However, several aspects of the design of these studies minimize this issue. Student subjects typically differ from "real" end users because they lack domain knowledge. To alleviate this problem, in both experiments we used a domain with which students were familiar (a university). Additionally, since the level of experience among student subjects vary, we controlled for experience and incorporate its effects into our analysis.

A second limitation, as stated previously, is that conclusions drawn on a lack of statistical evidence to reject the null hypothesis should be made with caution. It is important to note that failure to find a statistically significant difference does not prove that the effect does not exist. It simply means that we were unable to find that effect. In study one, we had adequate power to detect a moderate-sized effect, making it likely that there was no effect of practical significance to be found. While we are confident that the controlled nature of the experiment and the high rater reliability strengthens our ability to rule out alternative explanations, this study by itself still should not be considered conclusive. Instead, our findings indicate a need for additional studies to further explore the relative efficacy of these schemas.

Third, there is reason to believe that the complexity and size of the schema might have played a role, since the schemas used in study two had more entities and relationships than the schemas used in study one. Future studies should more rigorously examine these two effects by creating experimental conditions that test them separately. Specifically, one could hold schema type constant and vary the size and complexity of the schema. Studies in this area should also consider different, more elaborate tasks, which test a wider range of interactions with a data warehouse to more fully understand this relationship.

7. Conclusions

For a data warehouse to be effective, its content must be easily understood by those who use it. This study provides insight regarding schema choice and its effect on understandability. Kimball [23] contends that using the star schema as the underlying model for a data warehouse should facilitate understanding more effectively than the traditional relational schema. However, there is controversy surrounding that statement [6,21,23]. Our results challenge Kimball’s [23] contention as we found that users performed no better when using the star schema for a content identification task, and experienced users actually performed worse at a more sophisticated schema augmentation task. There are still technical reasons to use a dimensional model and the star schema – for example, a cube is constructed and indexed for the efficient retrieval of large amounts of data. The implication of our findings is that those technical reasons [21], not the understandability of the two alternative schemas, should be the stronger determinant in the choice of a data model.

Our results also imply that the use of "cube browsing" tools may be no more effective than relational query-by-example tools to end users. If users do not understand the content of a dimensional database any better than they understand a traditional relational database, it is unlikely that they will be able to effectively interact with the warehouse using sophisticated business intelligence tools. Just as users of relational databases use high-level graphical interfaces with pre-defined queries, the users of dimensional databases may require access to a set of pre-defined "views" of the data cube. The burden of constructing these views will still remain with the Information Technology function. Certainly, many organizations use business intelligence tools simply as reporting tools, requiring little of the end user. Future research could focus on the construction of visual metaphors, which provide users more flexibility without requiring direct interaction with the dimensional database.

Finally, the results of this study suggest that training is an important determinant of end user success in working with a data warehouse. Experience has a strongly significant effect on task performance (positive) and effort (negative) in the first study, and a strongly significant effect on task performance (positive) in the second study. More importantly, when it comes to inexperienced users, we could not find any evidence as to the superiority of one particular schema over the other. This would imply that data warehouse administrators should not, from a usability standpoint, spend time redesigning their data warehouse to improve understandability. Instead, they should train their users in the basic understanding of database schemas, regardless of their type.

REFERENCES

[1] J.R. Anderson, Cognitive Psychology and its Implications, 3rd ed., W.H. Freeman and Co., New York, 1990.

[2] M.H. Ashcraft, Human Memory and Cognition, Scott Foresman and Co., Glenview, IL, 1989.

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

[4] I. Benbasat and P. Todd, The effects of decision support and task contingencies on model formulation: A cognitive perspective, Decision Support Systems 33(4) (1996) 241-252.

[5] A. Bitterer, Management update: Steer clear of common data warehousing pitfalls (Gartner Group Research, November 16, 2005).

[6] M. Breslin, Data warehousing battle of the giants: Comparing the basics of the Kimball and Inmon Models, Business Intelligence Journal 9(1) (Winter 2004) 6-20.

[7] M. Brosey and B. Shneiderman, Two experimental comparisons of relational and hierarchical database models, International Journal of Man-Machine Studies 10(6) (1978) 625-637.

[8] A.M. Collins and M.R. Quillian, How to make a language user, in: E. Tulving and W. Donaldson (Eds.), Organization of Memory, Academic Press, New York, 1972, 309-351.

[9] K. Corral, D. Schuff and R.D. St. Louis, The impact of alternative diagrams on the accuracy of recall: A comparison of star-schema diagrams and entity-relationship diagrams, Decision Support Systems 42(1) (2006) 450-468.

[10] W.H. DeLone and E.R. McLean, Information system success: The quest for the dependent variable, Information Systems Research 3(1) (1992) 60-95.

[11] A. Gemino and Y. Wand, A framework for empirical evaluation of conceptual modeling techniques, Requirements Engineering 9 (2004) 248-260.

[12] D.L. Goodhue and R.L. Thompson, Task-technology fit and individual performance, Management Information Systems Quarterly 19(2) (1995) 213-237.

[13] M. Grise and R.B. Gallupe, Information overload: Addressing the productivity paradox in face-to-face electronic meetings, Journal of Management Information Systems 16(3) (2000) 157-185.

[14] S.G. Hart and L.E. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, in P.A. Hancock and N. Meshkati, Eds., Human Mental Workload (North-Holland, New York, 1988, 239-250).

[15] T. Haughey, Is dimensional modeling one of the great con jobs in data management history? Part 1, Information Management Magazine, 2004, downloaded 9/12/2009, http://www.information-management.com/issues/20040401/1000939-1.html.

[16] J. Hahn and J. Kim, Why are some diagrams easier to work with? Effects of diagrammatic representation on the cognitive integration process of systems analysis and design, ACM Transactions on Computer-Human Interaction 6(3) (1999) 181-213.

[17] M. Hertzum and E. Frokjaer, Browsing and querying in online documentation: a study of user interfaces and the interaction process, ACM Transactions on Computer-Human Interaction 3(2) (1996) 136-161.

[18] W.H. Inmon, The Problem with Dimensional Modeling, DMReview.com, 2000, downloaded 6/24/2009, http://www.dmreview.com/article_sub.cfm?articleId=2184.

[19] S.L. Jarvenpaa and J.J. Machesky, End user learning behavior in data analysis and modeling tools, Proceedings of the 15th International Conference on Information Systems, Atlanta, Georgia, 1986, 152-167.

[20] S.H. Juhn and J.D. Naumann, The effectiveness of data representation characteristics on user validation, Proceedings of the 14th International Conference on Information Systems, Atlanta, Georgia, 1985, 212-226.

[21] N. Jukic, Modeling strategies and alternatives for data warehousing projects, Communications of the ACM 49(4) (2006) 83-88.

[22] R. Kimball, The Data Warehouse Toolkit. John Wiley, New York, 1996.

[23] R. Kimball, A dimensional modeling manifesto, DBMSmag.com, 1997, downloaded 6/24/2009, http://www.dbmsmag.com/9708d15.html .
[24] R.L. Leitheiser and S.T. March, The influence of database structure representation on database system learning and use, Journal of Management Information Systems, 12(4) (1996) 187-213.
[25] K.H. Lim, I. Benbasat and P.A. Todd, An experimental investigation of the interactive effects of interface style, instructions, and task familiarity on user performance, ACM Transactions on Computer-Human Interaction 3(1) (1996) 1-37.
[26] W.C. McGee, On user criteria for data model evaluation, ACM Transactions on Database Systems 1(4) (1976) 380-387.
[27] G.A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, The Psychology Review 63(2) (1956) 81-97.
[28] J.C. Nordbotten and M.E. Crosby, The effect of graphic style on data model interpretation, Information Systems Journal 9(2) (1999) 139-156.
[29] D.G. Roussinov and H. Chen, Document clustering for electronic meetings: An experimental comparison of two techniques, Decision Support Systems 27(1-2) (1999) 67-79.
[30] D. Schuff, O. Turetken and J. D'Arcy, A Multi-Attribute, Multi-Weight Clustering Approach to Managing 'E-Mail Overload', Decision Support Systems 42(3) (2006) 1350-1365.
[31] C. Speier and M.G. Morris, The influence of query interface design on decision-making performance, Management Information Systems Quarterly 27(3) (2003) 397-423.
[32] J.K.H. Tan and I. Benbasat, The effectiveness of graphical presentation for information extraction: A cumulative experimental approach, Decision Sciences 24(1) (1993) 167-191.
[33] H. Topi and V. Ramesh, 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) 3-19.
[34] D. Vesset and B. McDonough, Worldwide Data Warehouse Platform Software 2009-2013 Forecast, 2009, downloaded 9/12/2009, http://www.idc.com/getdoc.jsp?containerID= 217442.
[35] R. Weber, Are attributes entities? A study of database designers' memory structures, Information System Research 7(2) (1996) 137-162.
[36] H. Yang, Adoption and implementation of CASE tools in Taiwan, Information & Maangement 35(2) (1999), pp. 89-112.

APPENDIX A: EXPERIMENTAL TASKS

Study One: Query Task

For a data warehouse built from the diagram shown, could you answer the following questions:

1. Which students had internships last year with GE?
2. How many "A"s did Professor John Doe give last semester?
3. How many jobs offered to students involved travel?
4. How many accounting majors have taken the "Introduction to Java" course?
5. How many CIS faculty got their Ph.D. from a Research I institution?
6. What percentage of internships offered no payment to students?
7. How many students failed "Introduction to Accounting" last semester?
8. How many CIS students transferred from another institution?
9. Which faculty had Mary Smith as a student?
10. Which students have been offered jobs with Motorola?

Answers:

(1) Yes, (2) Yes, (3) No, (4) Yes, (5) No, (6) Yes, (7) Yes, (8) No, (9) Yes, (10) Yes

Study Two: Schema Augmentation Task

You are in charge of designing the student database for the College of Business at Central State University. The Dean's office has set a goal to encourage student participation in the various student clubs on campus. To this end, they would like to track student membership in all clubs and professional societies.

Your task is to add the necessary entities to the current database so that it will record that information. Given the schema of the database (see the attached diagram), you will add entities and their attributes to capture the following information:

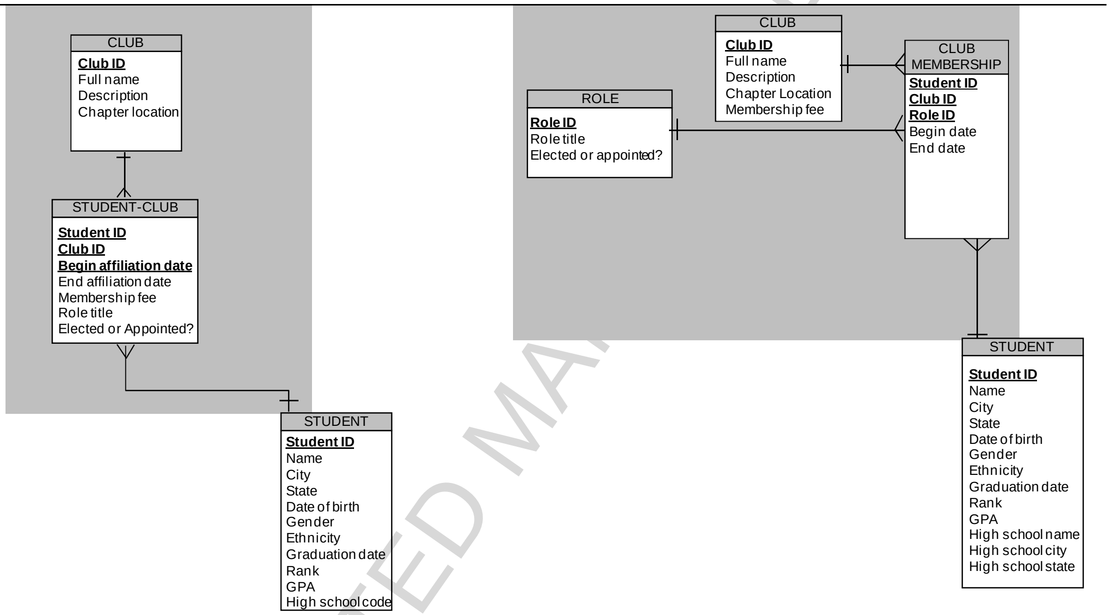
- The name of the club or professional society
- The title of the student's role in the club or professional society, and whether they were elected or appointed to that position
- A description of the club or professional society
- The location of the club or professional society
- Any fees that are part of membership
- The dates of their affiliation (beginning and end)

You can write your answer directly on the diagram, or in the space below.

Answers:

Shaded tables were left off of the schema in Figures 3 and 4. Subjects were asked to fill in the missing tables.

Traditional Relational Schema	Star Schema



Appendix B: NASA/TLX Instrument

The following are dimensions of demand which could describe the task you have just completed:

Item		Description
MD	Mental Demand	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
PD	Physical Demand	How much physical activity was required? Was the task easy or demanding, slack or strenuous?
TD	Temporal Demand	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
OP	Overall Performance	How successful were you in performing the task? How satisfied were you with your performance?
FR	Frustration Level	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?
EF	Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?

From each of the fifteen pairs below, select the item that was the larger factor for you while performing the task you just completed (for example, for the first pair, was there more physical demand or mental demand while completing the task?).

□ PD	□ MD
□ TD	□ MD
□ OP	□ MD
□ FR	□ MD
□ EF	□ MD

□ TD	□ PD
□ OP	□ PD
□ FR	□ PD
□ EF	□ PD
□ TD	□ OP

□ TD	□ FR
□ TD	□ EF
□ OP	□ FR
□ OP	□ EF
□ EF	□ FR

For each type of demand below, rate its overall level for the task you just completed (for example, what was the level of mental demand for this task?).

Demands	Ratings for Task						
	Low						High
MD	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
PD	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
TD	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
OP	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
FR	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
EF	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7

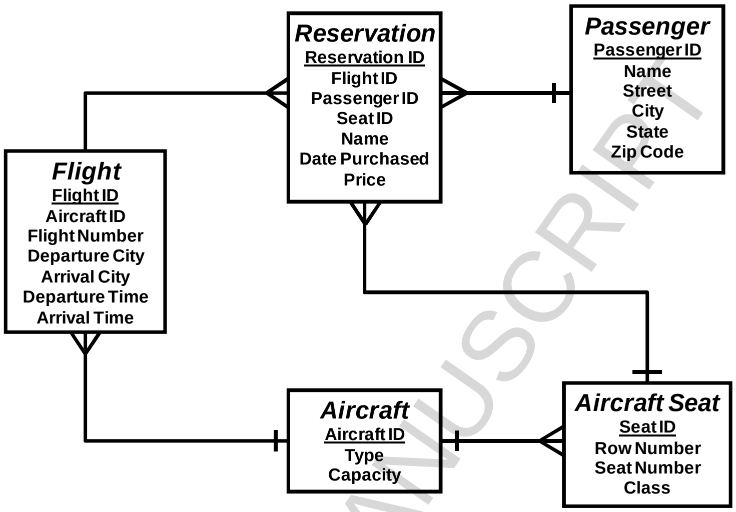


Figure 1. Simple Traditional Relational Schema

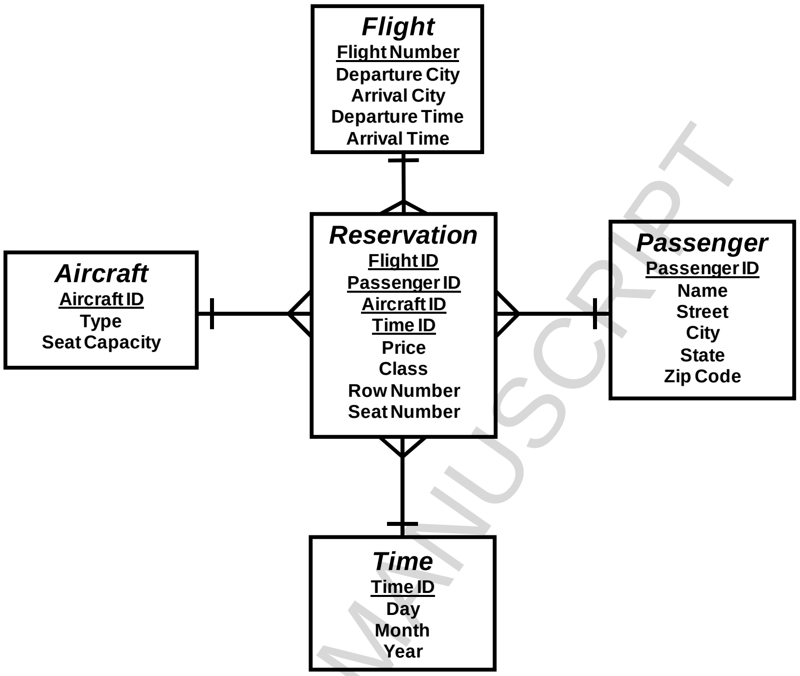


Figure 2: Simple Star Schema

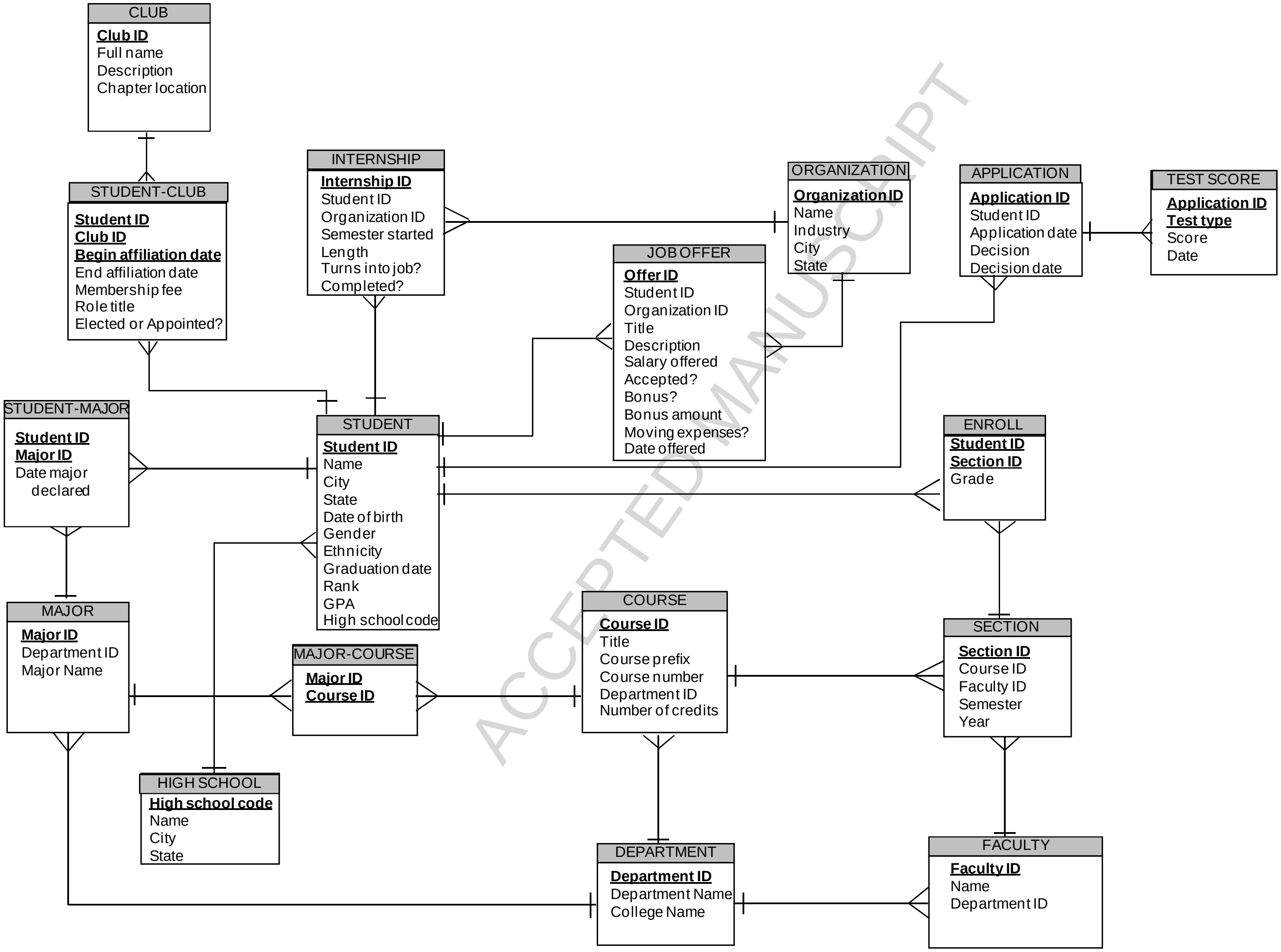


Figure 3. Complex Traditional Relational Schema (used in Study 2)

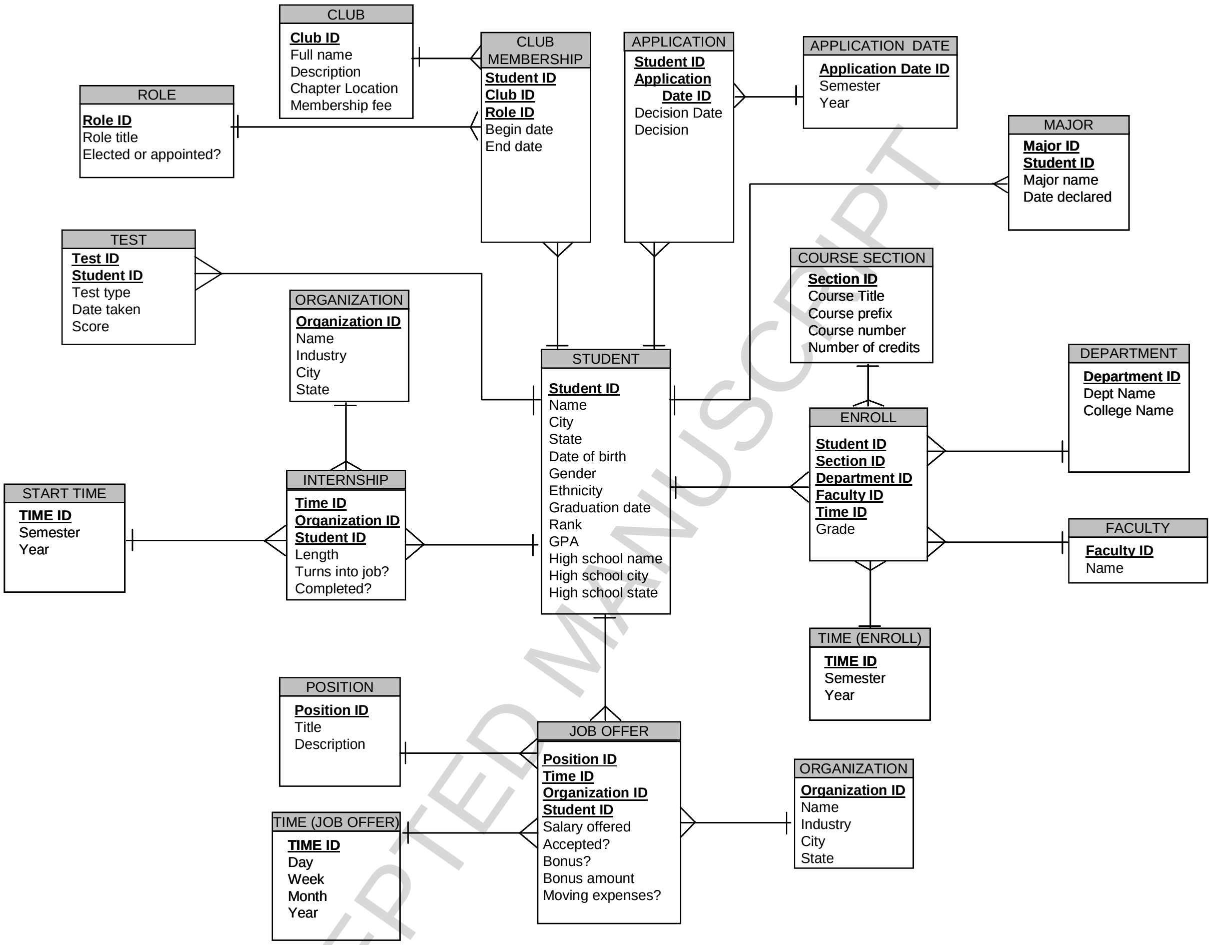


Figure 4. Complex Star Schema (used in Study 2)

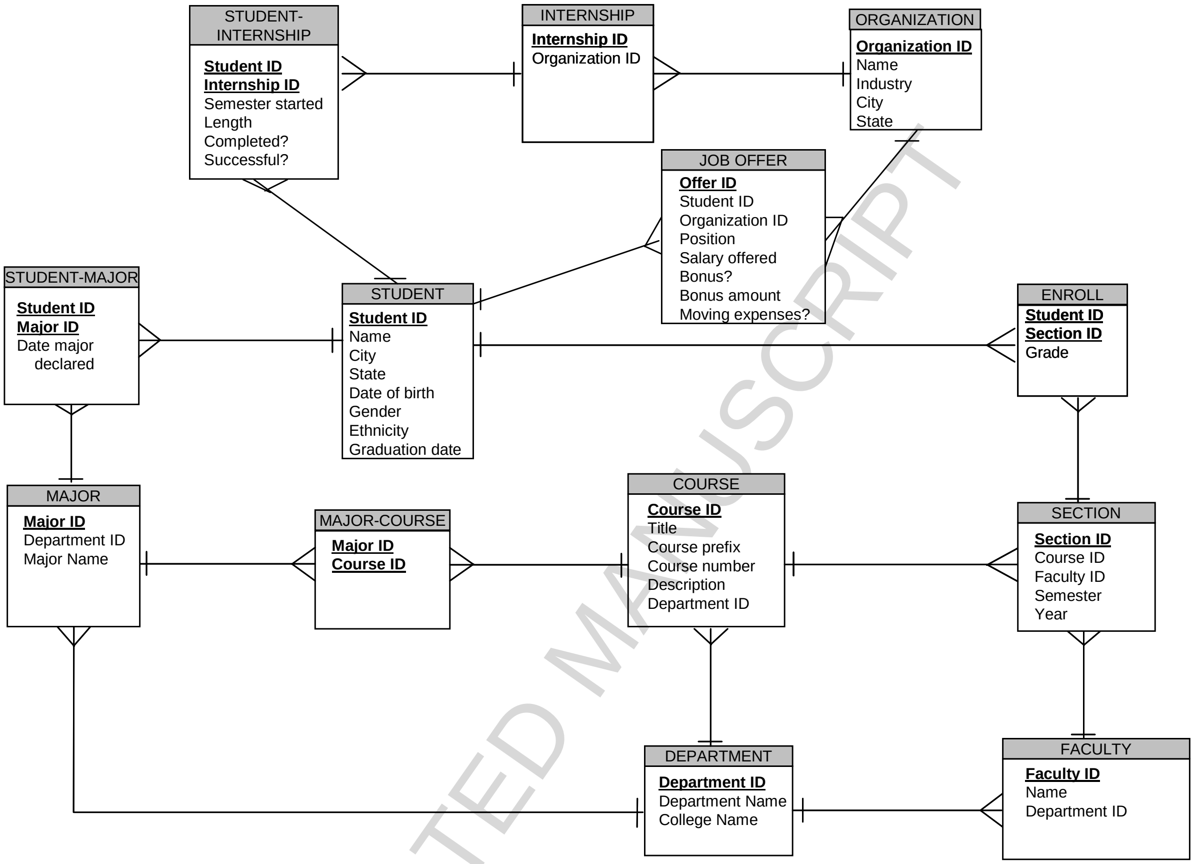


Figure 5. Traditional Relational Schema used in Study One

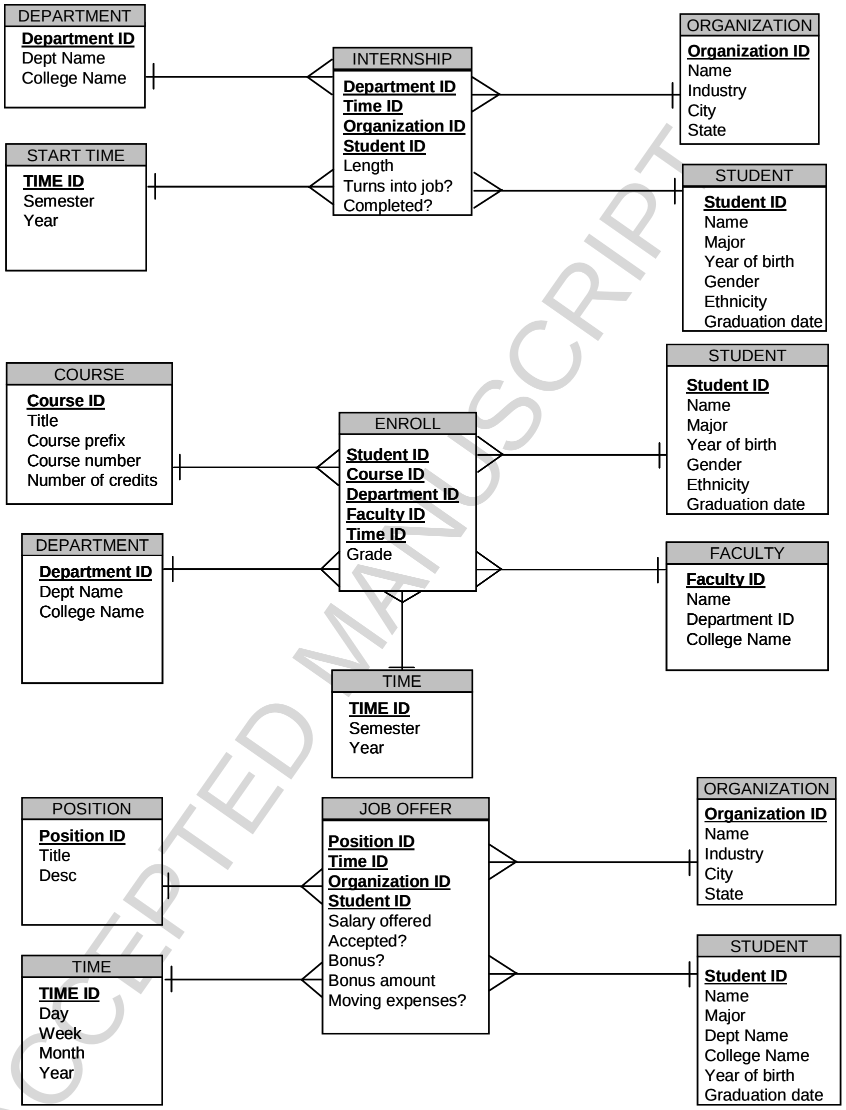


Figure 6. Star Schema used in Study One

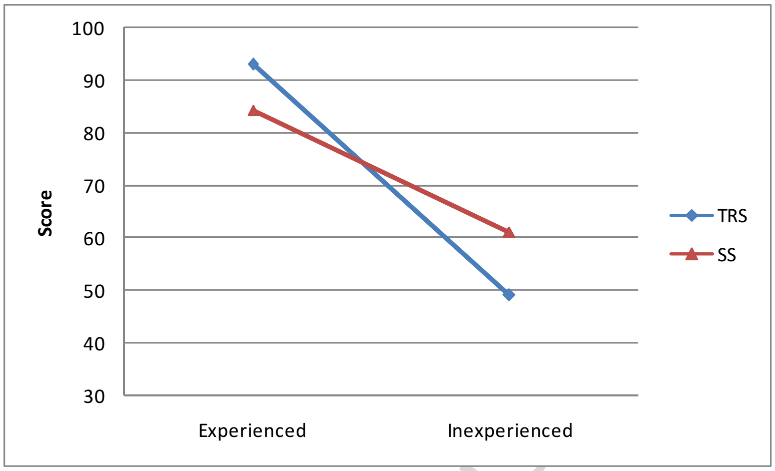


Figure 7. Interaction Diagram for Score – Schema Type by Experience (Study Two)

Table 1. Descriptive Statistics for Score (Study One)

Schema Type	Experience	Mean	Std. Deviation	N
Traditional Relational Schema	Experienced	8.215	0.879	51
	Inexperienced	7.480	1.644	50
	Total	**7.852**	**1.359**	**101**
Star Schema	Experienced	8.346	0.988	52
	Inexperienced	7.115	1.592	52
	Total	**7.731**	**1.456**	**104**
Total	Experienced	8.282	0.933	103
	Inexperienced	7.294	1.620	102
	Total	**7.790**	**1.407**	**205**

Table 2. Tests of Between-Subjects Effects for Score (Study One)

Source	Df	F	Sig.
Schema Type	1	0.403	0.526
Experience	1	28.431	0.000
Schema Type * Experience	1	1.802	0.181

R^2 = 0.133 (Adjusted R^2 = 0.120)

Table 3. Descriptive Statistics for Effort (Study One)

Schema Type	Experience	Mean	Std. Deviation	N
Traditional Relational Schema	Experienced	0.529	0.177	51
	Inexperienced	0.584	0.171	50
	Total	**0.556**	**0.176**	**101**
Star Schema	Experienced	0.504	0.167	52
	Inexperienced	0.599	0.156	52
	Total	**0.551**	**0.168**	**104**
Total	Experienced	0.517	0.172	103
	Inexperienced	0.591	0.163	102
	Total	**0.554**	**0.171**	**205**

Table 4. Tests of Between-Subjects Effects for Effort (Study One)

Source	Df	F	Sig.
Schema Type	1	0.045	0.833
Experience	1	10.035	0.002
Schema Type * Experience	1	0.722	0.397

R^2 = 0.051 (Adjusted R^2 = 0.037)

Table 5. Descriptive Statistics for Score (Study Two)

Schema Type	Experience	Mean	Std. Deviation	N
Traditional Relational Schema	Experienced	92.971	8.830	29
	Inexperienced	49.106	19.939	17
	Total	**76.760**	**25.459**	**46**
Star Schema	Experienced	84.377	20.324	32
	Inexperienced	61.222	27.119	17
	Total	**76.344**	**25.218**	**49**
Total	Experienced	88.463	16.387	61
	Inexperienced	55.164	24.231	34
	Total	**76.545**	**25.200**	**95**

Table 6. Tests of Between-Subjects Effects for Score (Study Two)

Source	df	F	Sig.
Schema Type	1	0.186	0.667
Experience	1	67.308	0.000
Schema Type * Experience	1	6.427	**0.013**

R^2 = .445 (Adjusted R^2 = .427)

Table 7. Tests of Between-Subjects Effects Score for Experienced Subjects (Study Two)

Source	df	F	Sig.
Schema Type	1	4.423	**0.040**

R^2 = 0.070 (Adjusted R^2 = 0.054)

Table 8. Tests of Between-Subjects Effects Score for Inexperienced Subjects (Study Two)

Source	df	F	Sig.
Schema Type	1	2.203	0.148

R^2 = 0.064 (Adjusted R^2 = 0.035)

Table 9. Descriptive Statistics for Effort (Study Two)

Schema Type	Experience	Mean	Std. Deviation	N
Traditional Relational Schema	Experienced	51.931	22.274	29
	Inexperienced	53.588	30.328	17
	Total	**52.543**	**25.227**	**46**
Star Schema	Experienced	62.094	24.781	32
	Inexperienced	66.706	9.999	17
	Total	**63.694**	**20.853**	**49**
Total	Experienced	57.262	23.979	61
	Inexperienced	60.147	23.211	34
	Total	**58.295**	**23.624**	**95**

Table 10. Tests of Between-Subjects Effects for Effort (Study Two)

Source	df	F	Sig.
Schema Type	1	5.462	**0.022**
Experience	1	0.396	0.531
Schema Type * Experience	1	0.088	0.767

$R^2 = 0.061$ (Adjusted $R^2 = 0.030$)

Table 11. Summary of Results

Study One: Content Identification Task			
	Test	**Result**	**Direction**
H1	$\text{score}_{ss} \neq \text{score}_{trs}$	Not supported	N/A
H2	$\text{effort}_{ss} \neq \text{effort}_{trs}$	Not supported	N/A
Study Two: Schema Augmentation Task			
	Test	**Result**	**Direction**
H3	$\text{score}_{ss} \neq \text{score}_{trs}$	Partially supported	$\text{score}_{ss} < \text{score}_{trs}$ (for experienced subjects only)
H4	$\text{effort}_{ss} \neq \text{effort}_{trs}$	Supported	$\text{effort}_{ss} > \text{effort}_{trs}$