International Journal of Information and Communication Technology Education
Volume 16 • Issue 2 • April-June 2020

Students' Difficulties in Identifying the Use of Ternary Relationships in Data Modeling

Rami Rashkovits, Yezreel Valley College, Jezreel Valley, Israel
Ilana Lavy, Yezreel Valley College, Jezreel Valley, Israel

ABSTRACT

The present study examines the difficulties novice data modelers face when asked to provide a data model addressing a given problem. In order to map these difficulties and their causes, two short data modeling problems were given to 82 students who had completed an introductory course in database modeling. Both problems involve three entity sets with relationships between them, either ternary or binary. The students' solutions were classified according to the types of errors they committed. More than half of the students provided faulty solutions. After an analysis of these results, open interviews were conducted with a selected group of students in order to figure out the reasons underlying the students' erroneous decisions regarding the data model. Among the reasons for their erroneous solutions were insufficient experience, lack of reflection on their solution, and lack of immediate feedback. In addition, the authors suggest instructional modifications derived from the research results.

KEYWORDS

Cardinality of Relationship, Data Model, Novice Data Modelers, Ternary Relationship

INTRODUCTION

The research literature includes an academic discussion on the difficulties novice modelers may encounter when designing a data model addressing given requirements in general (Dey et al., 1999) and regarding ternary relationships in particular (Hitchman, 2003; Batra, 2007). They encounter many difficulties, mostly concerned with cognitive complexity; among them, No Flexibility for errors, lack of immediate feedback, and information overload (Batra, 2007). As a result, data models designed by novice modelers tend to be inaccurate and erroneous, and hence the cause for the faulty behavior of information systems.

During their studies, novice data modelers study how to design a data model addressing given requirements. They study how to identify entities and how to set relationships between them. They also learn how to transform the entities and relationships into tables, fields and keys in order to form a relational schema.

One of the main challenges novice modelers face during the design phase is the identification of relationships between the entities involved. Novice data modelers find the setting of relationships between entities as their main challenge, mostly when non-binary relationships are involved (Batra, 1994).

DOI: 10.4018/IJICTE.2020040104

However, existing research is mainly focused on theoretical rather than empirical aspects of data modeling. That is, there has been little exploration of empirical data gathered from novice data modelers. Such empirical findings might shed light on the causes of the difficulties novice data modelers encounter during the design phase, and help instructors to improve their practice.

The aim of this study is to explore the difficulties novice data modelers encounter as novice data modelers regarding relationships between three entities. For this purpose, students who had completed a database course were asked to fill out a questionnaire including two problems dealing with various requirements, necessitating that their solutions use both binary and ternary relationships.

The research questions derived with the above aim are:

1. What are the types of error relating to the use of ternary relationships?
2. What are the underlying reasons for these errors?

THEORETICAL BACKGROUND

In this section, we present a brief theoretical survey of data modeling complexity, ternary relationships and students' difficulties in data modeling.

Data Modeling Complexity

Building a data model for an information system is a complex task, especially for novice data modelers (Topi & Ramesh, 2002). Novices encounter difficulties mainly in modeling relationships between entities (Batra & Antony, 1994). The main source of errors in data modeling by novices is attributed to cognitive complexity (Batra, 2007). Cognitive complexity, in the context of data modeling, refers to all the factors that make it difficult for one to grasp and understand all aspects of the problem at hand. These difficulties depend on the problem's structure as well as on the knowledge and previous experience of the designer. Four major sources of cognitive complexity were identified: problem solving principles, design principles, information overload, and systems theory (Batra, 2007). Though these papers refer to extensive factors, we focus only on those affecting data modeling, with special focus on ternary relationships. As to problem solving principles, the ones that are relevant to our research are connectivity and time delay. The first refers to a high degree of interrelatedness between the entities involved. The latter refers to the time gap between the design of the data model and its use. Time delay results in a lack of immediate feedback, therefore only upon use can one understand the quality of the design. As to design principles, the ones that are relevant to our research are the gap between the problem space and the solution space, no flexibility for errors, and a lack of knowledge of normalization rules. The gap between the problem space and the solution space refers to the ability to prune to a minimum the relationships described in the problem by the verbal constraints. No flexibility for errors refers to the fact that only one 'right' solution exists according to the normalization rules. A lack of knowledge of normalization rules refers to the lack of experience of novice data modelers as regards to normalization rules and their consequences. As to information overload, the only factor that is relevant to our research is noise, which refers to the presence of irrelevant information (Batra, 2007). In the process of analysis and evaluation of the results obtained in the present study, we use the factors listed above.

Ternary Relationships

The entity-relationship model (ER) (Chen, 1976), is commonly used to design databases (Lenzerini & Nobili, 1990). An ER model consists of entity-sets and relationships between them, representing real-world objects and their connections. Entity-sets include attributes of the objects, and specify an identifying key. Relationships connect two or more objects. The connections stand for a relationship representing a real-world association between the objects. However, not all real-world associations

between the involved objects must be present in the data model, only those necessary to support the system requirements. Binary relationships have multiplicity constraints representing limitations on the participation of an object in the relationship. Multiplicity constraints can be of type 1:1, 1:N, or N:M. Moreover, relationships may include additional attributes describing the connection itself.

The purpose of an ER model is to capture the business requirements and to construct an adequate relational database to support the business transactions. A set of transformation rules enables the translation of an ER model into a normalized relational data model (Chen, 1976). The entity-sets and the relationships between them are transformed into data tables. Each one includes data fields (attributes), storing records (objects) identified by a primary key. As to relationships, some are converted into tables while others are expressed using a key passed to one of the tables (foreign key).

The cardinality of a relationship can be binary (e.g., lecturer teaches course) or n-ary (e.g., physician write a prescription to a patient). A binary relationship connects two objects, while an n-ary relationship connects three or more (n) objects. In this study we focus on the understanding of novice data modelers regarding ternary relationship (n=3).

Related Work

In database design, relationships involving more than two entities are considered rare and, therefore, have not received adequate attention (Dey et al., 1999). The present paper provides a general framework for the analysis of relationships in which binary relationships simply become a special case. This framework helps a designer to identify ternary and other higher-degree relationships that are commonly represented, often inappropriately, as either entities or binary relationships. Generalized rules are also provided for representing higher-degree relationships in the relational model. This uniform treatment of relationships should significantly ease the burden on a designer by enabling him or her to extract more information from a real-world situation and represent it properly in a conceptual design.

Batra et al. (1990) found that novice data modelers encounter difficulties in modeling relationships between entities while designing data model via an ER model. The most common errors were related to the degree of a relationship and the selection of the types of entity participating in a relationship (Batra & Antony, 1994).

Both degree and connectivity errors are found frequently in novice solutions. Most novice errors are degree errors; usually, the form of this error is that the novice is unable to decide which entities participate in a relationship (Batra & Antony, 1994). A protocol study by Batra and Antony (1994) found that novice designers are satisfied if the entities in an application are somehow related to each other; they rarely consider the consequences of their design. For example, a heuristic commonly used in ER design is that if there is a sentence relating two entities, then there is a relationship between them in the solution. However, an indiscriminate use of this heuristic can lead to errors. For instance, if customers buy products via order, no database relationship exists between Customers and Products tables even if the phrase "customers buy products" is in the requirements description, the right solution would set a relationship between Customers and Orders tables, and another one between Orders and Products tables.

The Study

In this section, we present data about the study participants, the research aim and derived research questions, the task the participants had carry out, the data collection methods, and the analysis tools.

The Study Participants

Eighty-two second year students studying for a B.Sc. degree in the Information Systems department participated in the study. The research was conducted right after completing the course 'Introduction to Data Bases'. The participants were provided with a task with two phases. In the first phase they had to draw up entity-relationship diagrams and the corresponding relational data models according to given business scenarios. Then, they were asked to write SQL queries based on their own solutions.

The Task

To address the research questions, we have constructed two problems whose solution should shed light on the students' understanding of ternary relationships. The problems relate to three different entities and their relationships. The solution of the first problem requires the use of a ternary relationship between three entities, whereas the solution of the second problem requires the use of two binary relationships. The aim of the second problem was to examine whether students are able to identify the circumstances in which a ternary relationship between three entities is not required.

Two commercial business scenarios were presented to the study participants; each requires its own entity-relationship-diagram, and the derived relational model, including tables, fields, and primary keys. The scenarios deal with products, the countries in which the products are distributed, and the distributors who distribute the products.

The following scenarios refer to global commercial companies; each holds a different policy regarding the distribution of its products around the globe. The companies define the distributors, products and countries as follows: the distributors are identified by distributor number and characterized by name. Countries are identified by country number and characterized by name. Products are identified by product number and also characterized by name.

The students were directed to keep their relational models simple, and avoid redundant tables, attributes and keys.

Company A – Global Competition

Company A manufactures various products, which are distributed by distributors around the world. According to the company policy, each distributor is permitted to sell the company's products in each country. Yet, the company allows multiple distributors to sell products in the same country, and each may set a different price for each product type.

Company B – Exclusive Distributor

Company B manufactures various products, which are distributed by distributors around the world. According to company policy, each country has one distributor who distributes all the company's products exclusively. A distributor is allowed to distribute in many countries. The price of each product is set differently for each country.

SQL Query

In addition to ER Diagrams and relational models, the students were also requested to write SQL queries to present the price list of the products. If a product has different prices, all prices, sorted alphabetically by countries and distributors, were to be displayed, including product name, distributor name, price, and the country if the product had different prices in different countries.

Data Collection and Analysis Tools

The present study is a qualitative one and uses an interpretive paradigm which allows observing situations from the study participant's perspective (Merriam, 2009).

The research data included the student's solutions for company A and company B. After classifying their solutions according to their level of correctness, open interviews were conducted with 15 students. For each type of error identified, three students were interviewed.

The interviews transcripts were analysed through a gradual process of content analysis (Krippendorff, 2004; Neuendorf, 2002) and analytic induction (Taylor & Bogdan, 1998) in order to identify categories and typical patterns focusing on the underlying reasons for making the design errors.

RESULTS AND DISCUSSION

Reviewing the students' solutions to the Company A problem found that 39 students (out of 82) provided a correct solution (Figure 1). We identified three types of errors.

Erroneous Solutions of the Company A Problem

Two Binary Instead of One Ternary Relationship

Thirty-five students (out of 82) provided the solution for Company B (Figure 2), or a variation, as the correct solution for Company A. In this kind of solution, instead of using a ternary relationship (Figure 1) to capture the correct connections between the three entities involved, the solution includes only two binary relationships, omitting important information. In the variation presented in Figure 2, relationships are set between country and distributor, and between country and product (including price). The problem with this solution is that for each pair <country, product> only one price is set, contradicting the requirement that each distributor can set a different price. This solution also makes the distributor of a product in a country untraceable.

In another common variation, relationships are set between country and distributor, and between distributor and product (including price). The problem with this solution is that only one price per product is set for each distributor, contradicting the requirement to allow the distributor to set different prices for a given product in different countries.

Figure 1. Company-A solution

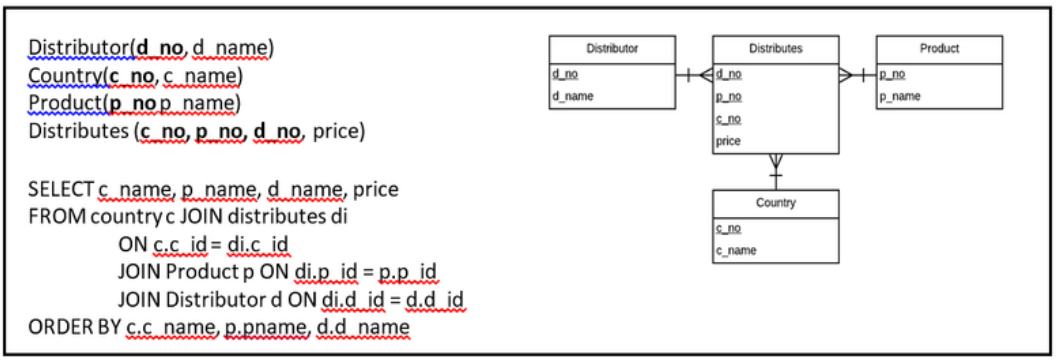
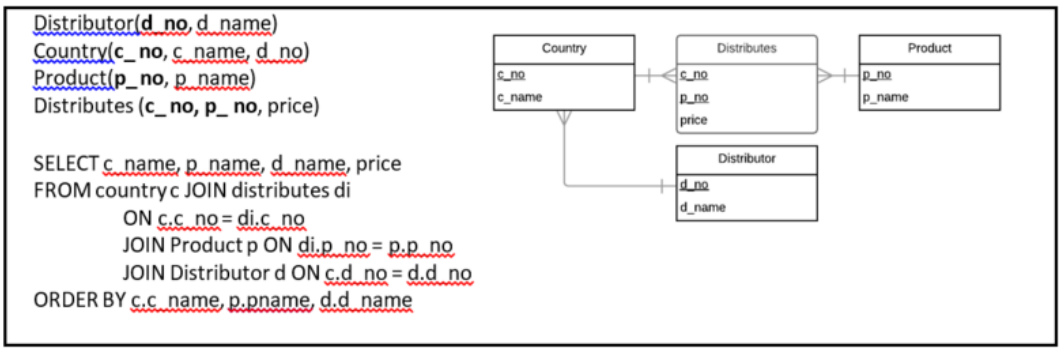


Figure 2. Company-B solution



Three Binary Relations Instead of One Ternary

Eight students (out of eighty-two) provided the solution shown in Figure 3. In this solution, the students used three binary relationships to capture all connections between the three entity-sets involved. From this solution one can extract which products are distributed by each distributor, and which products are sold in each country. However, one cannot extract the price each distributor set for a product, as only one price per <product, country> exists, contradicting the requirements.

Company A: Discussion

Figure 4 presents the distribution of Company A solutions. Thirty-nine students (out of eighty-two) managed to provide the correct solution (Figure 1). However, more than half of them provided erroneous solutions (errors 1 and 2).

Figure 3. Three binary relationships

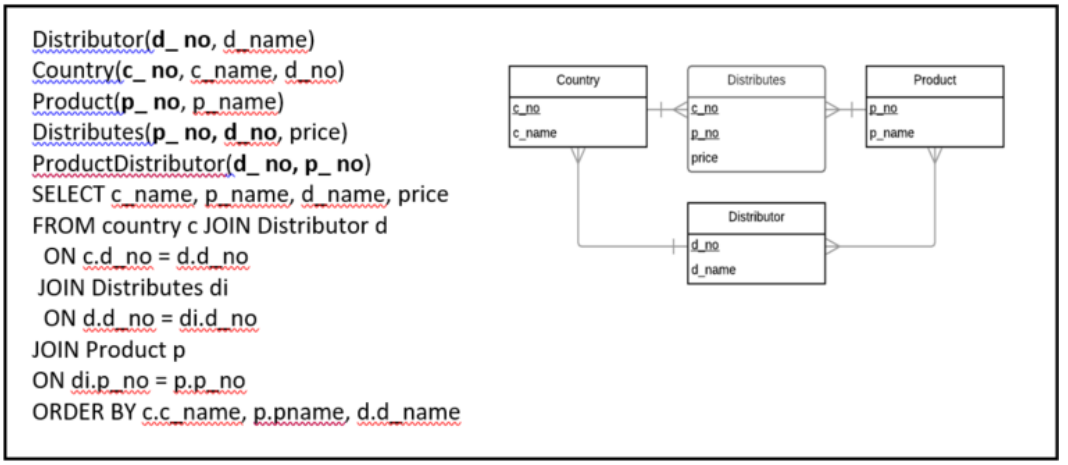
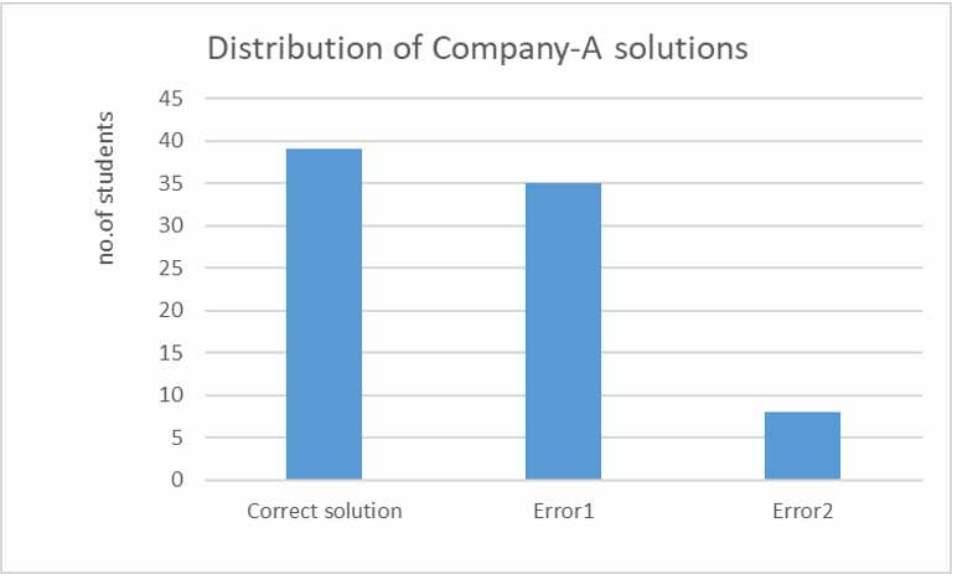


Figure 4. Distribution of solutions - company A



In what follows, we present representative excerpts from the solutions:

My solution was created gradually, after reading each sentence; I tried to figure out which model component fits. For instance, when I came across a distributor who distributed products, I created two entity-sets, and constructed a binary relationship between them. Then it was said that the distributor could distribute in any country, and I added the binary relationship between distributor and country. In retrospect, I had to go back to the question to make sure that my solution addresses all the constraints.

From the above excerpt, one can see that some students singled out and mapped the objects to the model components, then figured out the relationships between these objects, but they continued onward without reflecting on the obtained results. They avoided going back to re-examine the solution in light of the problem requirements to verify its correctness and integrity via examples of data and queries:

If I had the chance to run the query I wrote against real data I believe I would immediately get the design error and fix it. But when the solution is "on paper" it's harder for me to detect such errors.

Although the students were asked to plan the relational schema and formulate a query based on it, they still were not confronted with their faulty design because they could not view the query outcomes. This is in line with Batra (2007), who referred to time delay as one of problem solving principles influencing cognitive complexity in data modeling.

In the students' reflections after solving the given problems, another issue came up. Here is a representative excerpt:

I understand ternary relationship theoretically, but I find it difficult to implement. The common relationships are of binary type; therefore, we were less exposed to other types of relationships, although they are more difficult to understand. Specifically, the cardinality issue is quite confusing. I must admit that I do not understand thoroughly the difference between two binary relationships and a ternary relationship, and I still believe that my solution can also be accepted as a good one.

We learned how to plan a data model and I tried my best to do so. In retrospect, if we had received examples of wrong data models, and would have been asked to detect the errors, I think it may have been more useful. I learn better from mistakes...

From the first part of the first excerpt we can learn about the reason underlying the students' difficulties in identifying ternary relationships, which stems from too few examples and a non-exhaustive discussion of the differences between the types of relationships. The second part of the first excerpt shows that the students did not internalize the fact that the problem of designing a schema has a single solution (Batra, 2007). This may be attributed to the fact that in math and computer science problems, there is usually more than one way to solve a problem, and students are asked to look for more than one method of solution and examine its implications.

Batra (2007) presents four major sources of complexity. We believe that there is an additional source, experience in data design. From both excerpts, we may learn that students feel that they are not experienced enough in designing a data model, since it requires skills that are gained only via experience acquired over time while working in the field. Being novice data modelers they use naïve heuristics that may lead them to an erroneous design. Experienced designers use advanced heuristics gained via engagement in similar problems (Purao, Storey & Han, 2003).

The following excerpt raises an additional issue:

I'm not familiar with the field of commerce. Concepts such as distribution and exclusivity are not sufficiently clear to me. Hence, I apparently did not understand the nuances in the text referring to the commercial relationships embedded in the problem and their implications on the desired data model.

From the above excerpt, we may learn the when students are introduced to a problem in a domain that is not familiar to them, the complexity of the problem is greater. This may cause misinterpretations of the concepts embedded in the problem, which might result in a faulty solution (Holland et al., 1986).

Erroneous Solutions of the Company B Problem

Reviewing the students' solutions to the Company B problem revealed that 34 students (out of 82) provided a correct solution (Figure 1). However, we identified three types of errors.

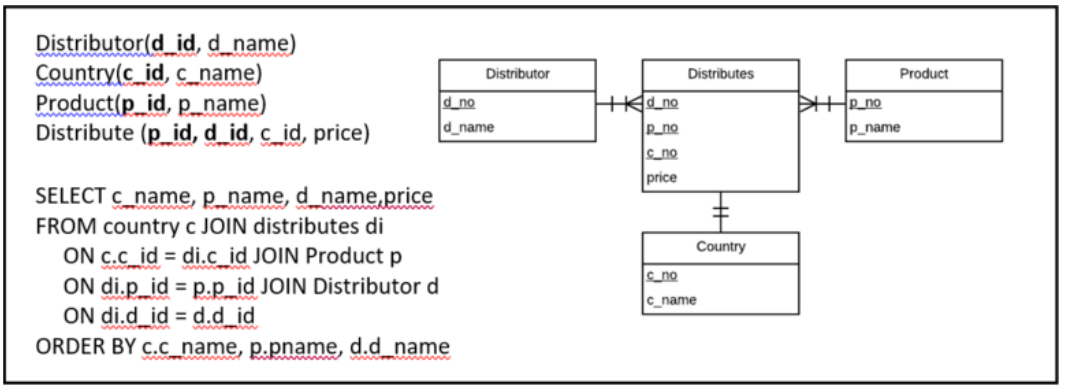
Ternary 1:N:M Instead of Two Binary Relationships

Thirty-seven students (out of eighty-two) provided a ternary-based solution. Twenty students provided the solution shown in Figure 1 (error 1), and seventeen students provided the solution shown in Figure 5 (error 2). The difference between these two solutions is the '1' multiplicity constraint connecting the country to the ternary relationship. Both solutions are erroneous since they do not comply with the requirement of one exclusive distributor per country. The solution shown in Figure 1 supports multiple distributors per product in every country. The solution shown in Figure 5 prevents more than one distributor per product in a country; however, it does not prevent other distributors from distributing other products in that country, contradicting the exclusiveness requirement. Moreover, the latter solution also enables a distributor to act in multiple countries, contradicting the requirement of avoiding having distributors distribute in more than one country.

Three Binary Relationships

Eleven students (out of eighty-two) provided solutions similar to the ones presented in Figure 3 (error 3). Adding redundant relationships can cause a discrepancy in the data. For example, if John is the sole distributor in England and distributes Product X, it will appear in the Distributes table (which connects products and countries). However, product X might not be included in the relationship table between distributors and products (the redundant relationship), and hence might cause a discrepancy. In addition, there is no information in this redundant relationship that cannot be retrieved from the other two relationships. Moreover, the students were explicitly instructed to avoid repetitive and redundant elements in their solutions.

Figure 5. 1:N:M Ternary solution



Company B: Discussion

Our research concerns the understanding and implementation of ternary relationships. For that purpose, we explored the following situations: (1) when a ternary relationship is required (the company A problem); (2) when it is not required yet is being used (the company B problem).

Figure 6 presents the distribution of Company B solutions. Thirty-four students (out of eighty-two) managed to provide the correct solution (Figure 2). However, more than half of them provided erroneous solutions (errors 1-3).

Confronting students with their faulty solutions yielded the following representative excerpts:

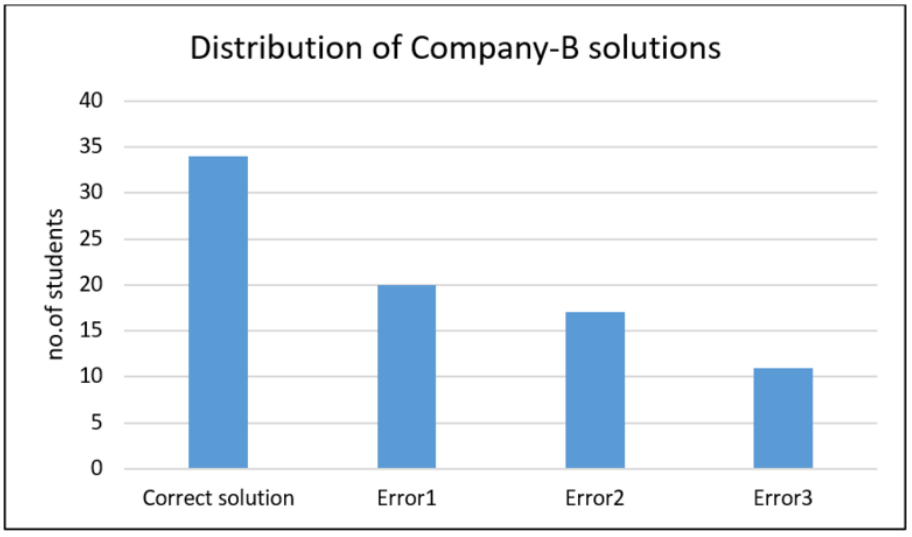
I thought the purpose of the second problem was to check if my solution is consistent with the first one. I thought that both problems have the same solution although they were phrased differently. I still believe that my solution addresses them well.

I thought that this question is a variant of the previous one. Since I used a ternary relationship in the first problem, I thought it fits here too, adding a small change in the constraints. I did not notice that the solution did not meet the exclusivity requirement of the distributor in the country. I assume that if I was given the second problem before the first one, I would not fall into this trap.

From the first excerpt, we learn that some students assumed that the goal of the questions was not only to examine their abilities to provide correct solutions to design problems, but also to examine the consistency of their solutions. In other words, instead of trying to deal with solving the second problem, the students focused their efforts on keeping consistency with their first solution.

From the second excerpt, we learn that sometimes students make prior assumptions before they try to solve a given problem. In this case, since this problem was given to them after they solved the Company A problem, in which a ternary relationship was required, some of the students who solved correctly the previous problem assumed that the company B problem was a variation of the company A problem. As a result, they tried to adapt the previous solution to fit the new requirements by changing the cardinality constraints of the ternary relationship instead of starting the design from scratch. They actually used a naive heuristic that results in an erroneous solution (Polya, 1985).

Figure 6. Distribution of solutions - company B



The students who provided three binary relationships solutions asserted that:

I modeled every entity and relationship I found in the text, the way we were taught. I did not understand that one of the relationships is redundant, and I believe that it is not problematic anyway.

All three binary relationships are of equal importance, and I don't know how to choose among them. I can't understand how one is redundant, since each captures unique information.

Again, we see from the first excerpt that a naive heuristic (Tversky & Kahnemann, 1974) is used by novice data modelers. From both the first and second excerpts we can see that novices follow the strict rules of design, and are not sensitive to nuances. Without being immediately faced with the consequences of their bad design, they lack the ability to avoid errors and understand them. Data model design is not always straightforward, and a profound understanding of the inter-relationships between the model constituents is required.

OVERALL DISCUSSION AND IMPLICATIONS FOR INSTRUCTION

The process of solving a data-modeling problem is a complex task that requires cognitive activities such as translation, mapping objects from reality into abstract entities in the model, as well as continual validation that the result obeys the problem requirements. The difficulty increases with ternary relationships. According to Batra (2007), there are four major sources for cognitive complexity: problem-solving principles, design principles, information overload, and systems theory. As was mentioned before, we believe there is an additional source, which may affect some other sources: professional experience.

Gaining professional experience in data modeling means, among other things, confronting the consequences of erroneous models. This process usually leads to the understanding that continual reflection on the built model must be carried out, and that the obtained model should be examined against the data and queries. Professional experience also means the developing of problem solving and critical thinking abilities, as well as being sensitive to the nuances embedded in various situations.

In this study, we found that many students, who are considered novice data modelers, do not reflect on their solution process. They approach modeling problems making prior assumptions without examining them against the problem requirements. Part of the students assume that, as in other related areas (e.g. programming, mathematics) a problem can have more than one correct solution (Batra, 2007).

Hence, we suggest the following recommendations for data modeling instruction:

1. We recommend extending the process of learning data modeling principles to reach its implementation. That is, immediately after the study of modeling principles, students will be asked to build the relational model and run queries on the obtained model. This way, they can be exposed to the consequences of incorrect modeling and avoid the problem of 'time delay' (Batra, 2007);
2. The students should engage with different types of modeling problems. Namely, not only the standard problems they usually get, which involve the modeling of a given situation, but also different kinds of tasks. For example, to provide them with an erroneous model, and ask them to check the model's 'quality' through examining its derived relational model and the execution of different queries;
3. The students should carry out comparison tasks in which they will have to examine two data models, one correct and the other erroneous, against the problem's requirements in order to develop their awareness of the meaning of each requirement in the problem;

4. The students should be provided with tasks involving peer work. That is, one student has to design the data model and hand it to a classmate. The classmate has to construct a relational model and run queries on it using that data model. In case of encountering difficulties due to erroneous modeling, that classmate has to provide feedback to the original student;
5. All the above recommendations should be followed by class discussions in which students are exposed to problems their classmates encountered, so as to avoid them in the future.

REFERENCES

Batra, D. (2007). Cognitive complexity in data modeling: Causes and recommendations. *Requirements Engineering*, *12*(4), 231–244. doi:10.1007/s00766-006-0040-y

Batra, D., & Antony, S. (1994). Novice errors in database design. *European Journal of Information Systems*, *3*(1), 57–69. doi:10.1057/ejis.1994.7

Batra, D., Hoffer, J. A., & Bostrom, R. P. (1990). Comparing representations with the relational and extended entity relationship models. *Communications of the ACM*, *33*, 126–139. doi:10.1145/75577.75579

Chen, P. P. S. (1976). The entity-relationship model—Toward a unified view of data. *ACM Transactions on Database Systems*, *1*(1), 9–36. doi:10.1145/320434.320440

Dey, D., Storey, V. C., & Barron, T. M. (1999). Improving database design through the analysis of relationships. *ACM Transactions on Database Systems*, *24*(4), 453–486. doi:10.1145/331983.331984

Hitchman, S. (2003). An interpretive study of how practitioners use entity-relationship modelling in a ternary relationship situation. *Communications of the Association for Information Systems*, *11*(1), 451–485.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Krippendorff, K. (2004). *Content Analysis: An introduction to its methodology*. Beverley Hills, CA: Sage Publications.

Lenzerini, M., & Nobili, P. (1990). On the satisfiability of dependency constraints in entity-relationship schemata. *Information Systems*, *15*(4), 453–461. doi:10.1016/0306-4379(90)90048-T

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: John Wiley & Sons.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

Polya, G. (1985). *How to Solve It*. Princeton University Press.

Purao, S., Storey, V., & Han, T. (2003). Improving reuse-based design: Augmenting analysis patterns reuse with learning. *Information Systems Research*, *14*, 269–290. doi:10.1287/isre.14.3.269.16559

Taylor, S. J., & Bogdan, R. (1998). *Introduction to Qualitative Research Methods*. New York: John Wiley & Sons.

Topi, H., & Ramesh, V. (2002). Human factors research on data modeling: A review of prior research, an extended framework and future research directions. *Journal of Database Management*, *13*(2), 3–15. doi:10.4018/jdm.2002040101

Tversky, A., & Kahnemann, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124 PMID:17835457

Rami Rashkovits is a senior lecturer at the Academic College of Yezreel Valley in Israel since 2000 in the department of Management Information Systems. His PhD dissertation (in the Technion – Israel Institute of Technology) focused on content management in wide-area networks using profiles concerning users' expectations for the time they are willing to wait, and the level of obsolescence they are willing to tolerate. His research interests are in the fields of distributed systems as well as computer sciences education. He has published over thirty papers and research reports.

Ilana Lavy is an associate professor with tenure at the Academic College of Emek Yezreel and is the department head of Management Information Systems. Her PhD dissertation (in the Technion) focused on the understanding of basic concepts in elementary number theory. After finishing doctorate, she was a post-Doctoral research fellow at the Education faculty of Haifa University. Her research interests are in the field of pre service and mathematics teachers' professional development as well as the acquisition and understanding of mathematical and computer science concepts. She has published over seventy papers and research reports (part of them is in Hebrew).