Fessakis G., Dimitracopoulou A., Komis, V. (2005). Improving database design teaching in secondary education: action research implementation for documentation of didactic requirements and strategies. Computers in Human Behavior. Vol, 21,(2), pp. 159-194. This is the authors version of the journal article, published in *Computers in Human Behavior* © Elsevier Ltd. 2005 available at http://dx.doi.org/10.1016/j.chb.2004.06.006

IMPROVING DATABASE DESIGN TEACHING IN SECONDARY EDUCATION:
ACTION RESEARCH IMPLEMENTATION FOR DOCUMENTATION OF DIDACTIC REQUIREMENTS AND STRATEGIES

George Fessakis[a,*], Angélique Dimitracopoulou[b], Vassilis Komis[c]

[a,b] Learning Technology and Educational Engineering Laboratory,
Department of Education, University of Aegean, 1 Dimokratias Av., Rhodes, 85100, GREECE
[c] Department of Early Childhood Education, University of Patras, 26500, Patras, GREECE.

Abstract

Database design and use has educational interest for utilitarian and learning reasons. Database technology has significant economic impact and the demand for database design can not be covered by the existent educated experts. Furthermore the database management systems available at schools could be used for the design and implementation of high quality learning activities. Databases are general purpose modeling environments that enable problem solving using conceptual frameworks closer to the solver and the problem than the machine architecture. Databases design introduction in the curricula of secondary education programs raises educational research questions. Research questions concern the didactics of the subject as well as the value of database design based learning activities. In this paper we present some of the more significant findings of an action research concerning the database design in secondary education. Research questions concern the ideas of students about databases and their difficulties during database design. Data, collected using a variety of research activities, are analyzed and discussed and teaching strategies are proposed.

Keywords: Database modeling and design, Cognitive difficulties, Secondary Education

1. Introduction

Database technology has significant economic impact. Labor statistics organizations predict that database management is likely to experience of the faster growth in jobs for the period 1998-2008 (Antony & Batra, 2002). This is mainly because administration systems and many internet applications are based on databases. The increasing demand for database construction results on a significant percentage of databases developed by employees without formal related training. The data base design from uneducated people (e.g. end users) raises issues concerning effectiveness of the produced solutions (Batra, Hoffer, & Bostrom, 1990). The above facts state utilitarian reasons for large scale database design education.

Furthermore Data Base Management Systems (DBMS's) as general purpose modeling environments (Hancock, & Kaput, 1990) are cognitive tools (Jonassen, 2000) that enable their users to exploit computational resources for problem solving providing conceptual frameworks closer to the user and the problem than the computer architecture. Databases represent the structural characteristics of physical systems. In contrast with other structural modeling techniques (e.g. conceptual maps), databases are interactive and executable models facilitating learning activities in

* Corresponding author
E-mail addresses: [gfesakis, adimitr]@rhodes.aegean.gr (Fessakis G., Dimitracopoulou A.), komis@upatras.gr (Komis V).

which the learner actively reflects his own perception of the physical system under study. Database design can be used for the development of learning activities consistent to modern learning theories and didactic approaches. The above observations combined with the wide availability of user friendly DBMS's at schools states a strong interest for database design learning in K12 education and opens didactic research questions.

Databases introduction in secondary education curricula is not accompanied usually by thorough research. In the Greek educational system, database design consists an obligatory subject for information technology vocational schools and an elective subject in general education schools. The official curriculum is quite similar with that of a typical university level database design subject. In other words, due to lack of related research there is not any didactic transformation of the database design subject for the design of a corresponding curriculum proper for the secondary education.

The above arguments clarify the educational research interest about databases design in secondary education. In this paper, we present key findings of an action research aiming to explore learners' difficulties and therefore to improve database design instruction.

2. Survey of related researches

In order to support the feasibility of our research a brief survey of researches about the human factors affecting data modelling and database didactics is presented.

2.1. Researches comparing basic logical data models

There is a first period in the research of human factors affecting data modelling where the researchers were comparing the three competitive logical data models (relational, network, and hierarchical data models (Date, 1990)). Logical data models are representation systems for the specification of how data should be stored using abstract data structures independent of the physical storage medium (disk, tape etc). Each of the above mentioned data models adopts a different basic data structure (relation, network and tree respectively) that enables the designer to view data in terms closer to the problem than the machine. The comparison of data models was based on:

i. Query formulation by users (Lochovsky & Tsichritzis, 1977): Comparison on query formulation shows, in general, that relational data model is easier to use successfully by non expert.
ii. Understanding of the produced schemata (Brosey & Shneiderman, 1978): Researchers supplied subjects with hierarchical and relational schemata for the accomplishment of problem solving activities. Research results claim that hierarchical schemata are considered more understandable by non expert than the corresponding relational one.
iii. Observation of the data structures that people use impulsively (Durding, Becker, & Gould, 1977): In this research, subjects designed data for given problems without the commitment to a specific data model. Results show that humans organize data using structures indicated by the semantic relationships in the problem description. Data models are based on a single kind of data structure while humans would like to use a variety of structures according to

the problem needs. In other words in terms of human usability there is not a clearly superior simple data model.

These researches are methodologically interesting, but the dominance of the relational data model nowadays makes them rather obsolete for the purposes of didactics. In addition the adoption of conceptual level database design reduces the significance of the demand for a variety of data structures to the logical level. Designers can design databases independently of the logical data model using more abstract representation systems (Batini, Ceri, & Navathe, 1992) like Entity Relationship (ER) (Chen, 1976), Integration DEFinition for Information Modeling (IDEF1X) (Federal Information Processing Standards Publication 184, 1993), etc.

2.2 Researches about conceptual models effectiveness

Researches of this kind compare:

- Database design using conceptual modeling to design using only logical one (e.g. (Batra & Davis, 1989); (Juhn & Naumann, 1985))
- Different conceptual models effectiveness (Batra, Hoffer, & Bostrom, 1990).

These interesting researches provide at least the information that the use of conceptual models facilitates the understanding of relationships and their cardinality while the relational model facilitates the primary key definition (Juhn, & Naumann, 1985). This means that is purposeful to use them both in instruction.

The conceptual models usage during database design is widely adopted in industry and academia so there is no an initially open question whether we should use them or not. The appropriateness of the specific conceptual models for young students is still an open question that we will face too. In this research field there is an on-going interest for the object oriented data modelling, but this domain is out of the present paper's scope.

2.3 Researches about the human factors during conceptual modelling using ER

There are some researches that explore the difficulties that designers face using ER conceptual model. In (Goldstein, & Storey, 1989) and (Hall, & Gordon, 1998) there is evidence that designers confuse entities with attributes and despite ER simplicity, users need methodological support to apply it. In (Antony, & Batra, 2002) is mentioned that novice designers express redundant relationships and it is proposed that the difficulties with relationships are related to their combinatorial semantics.

In (Mcintyre, Pu, & Wolff, 1995) authors propose the use of an expert system (named RA) in the teaching of relational database design. Students used RA to produce relational database schemata from business forms and asked to compare this design procedure to the traditional (based on normalization theory) from the non-technical end user point of view. Student responses were mixed. One-third was for continuing to use traditional methods, another third were for using RA and the rest were undecided.

In the above researches, specially designed software environments for the database design learning are proposed. These environments have been designed for use in higher than the secondary education levels; they are not widely available and are usually in a rather prototype state.

2.4 Researches about learning value and use of databases

All the above researches concern undergraduate and postgraduate students or professionals of the IT industry. One of the most widely known database related work for primary and secondary education is the *"Tabletop"* software (Hancock, & Kaput, 1990), (Hancock, Kaput, & Goldsmith, 1992), (Bagnall, 1994), which mainly aims at data analysis rather than database design in the level of typical conceptual and logical design.

In (Jonassen, 2000) there is extensive reference concerning learning activities using databases for secondary education as well as description of general kinds of such activities. Research references concerning databases in secondary education are facing mainly issues about data analysis rather than database design using typical methods and techniques of computer science.

From the above researches' survey and analysis, it is obvious that there is a lack of educational research concerning the teaching of database design in secondary education and its didactic implications. Furthermore, there is no research for the human factors affecting database design learning for secondary education students in the authors' knowledge. This paper would like to contribute in this direction.

3. Methodological framework

In order to formulate the methodological framework of the research the hermeneutic (interpretive) epistemological view is adopted (Hiley, 1991). According to hermeneutics' view there are physical and social phenomena. Physical phenomena evolve independently of the possible human observer and/or participator. Physical phenomena are usually described by scientific models and it is possible to be reproduced in lab conditions by independent observers. Social phenomena, in contrast, are mainly subjective and evolve dependently on how the involved humans deal with them. In other words, social phenomena's evolution is affected by the thoughts, emotional condition, values, and perceptions of the involved humans including the observer-researcher. In general, it is not possible to reproduce social phenomena in lab conditions or to describe those using deterministic scientific models.

The goal of hermeneutic research is mainly to advance the understanding of social phenomena, collecting detailed information, formulating interpretations, even stating axiological arguments. Observer neutrality is not a requirement in hermeneutics. Learning and teaching are considered social phenomena and are going to be studied using "action research" methodology.

3.1. Research methodology

For the determination and exploration of research questions we adopted "action research" methodology. Action research is the study of a social phenomenon in order to improve the quality of action in the framework of this phenomenon (Altrichter, Posch, & Somekh, 2001). This definition indicates that the basic motivation of the involvement in an educational action research is the improvement of teaching and learning quality. Educational action research is implemented usually by in service teachers who wish to face the challenges and problems of educational practice or to implement innovations after thorough speculation.

The theoretical foundation of action research is based on reflective rationalism and can be briefed in the following concessions (Altrichter, Posch, & Somekh, 2001):

- Complex practical problems require specific solutions
- Specific solutions are possible to be developed in the context that the problems appear. In this context the in-service teacher has a determinant role.
- Solutions may not have general applicability but they constitute suppositions for test by other in service teachers.

Educational action research aims at the development of autonomous, professional improvement ability for teachers using systematic self-observation, other teachers' work study and testing of ideas using research procedures in the class.

3.1.1. Action research basic schema

The basic schema of an action research can be described briefly using the following four step iterative and adaptive procedure:

i. Selecting a starting point: Every action research starts from a problematic state which is called starting point. In general, every phenomenon that teachers wish to understand better or to modify could be an action research starting point.

ii. Clarification of the starting point: In this stage the researcher employs several information collection and analysis methods in order to promote the starting point's understanding.

iii. Development and implementation of action strategies: Starting point's clarification enables the development of action strategies for the improvement of the problematic state. Action strategies that are not immediately effective trigger a new cycle of action strategies' formulation.

iv. Analysis and theory development: The research data analysis and the improvement of action strategies can be used by the researcher to formulate a theory. The action research ends with the diffusion of the professional knowledge obtained by the researchers.

3.1.2. Data analysis methods

For the clarification of the starting point we mainly categorize data and analyze the produced frequency distributions. To analyze in a complementary manner some of our research findings we carry on also a Multiple Correspondence Analysis (MCA). Multivariate analysis methods have progressed significantly the last years, and their applications have expanded in various disciplines including educational research studies (Benzécri, 1992). MCA constitutes a tool suitable to explore relationships between qualitative variables, especially when research data concern simultaneous measurements of many parameters. Analytically, MCA include possibilities like sorting and grouping variables (in order to investigate similarities and dissimilarities between groups), exploration of the dependence and/or interdependence relations among variables and prediction of relationships between variables. It offers efficient tools that can help us to overcome the intrinsic limitations of the descriptive statistics. This method is also known as Homogeneity Analysis and Dual Scaling. It aims at the graphical representation of the structure of non-numerical multivariate data. The central principle of MCA method is that complex multivariate data can be accessible by displaying their main regularities and patterns in graphs and diagrams.

3.2. Research purpose

As mentioned above, the increasing importance of database technology, the availability of desktop DBMS in schools, and the pedagogical interest of them, rationalises the demand of large scale education in use and development of databases. Students should be familiarized with database design not only to utilize computational resources in problem solving using databases but also to be able to participate in general learning activities. The general problem (starting point) of the research is the interest to teach effectively database design in secondary education students so they can:

- Exploit the related technology in every day problem solving
- Participate in general learning activities with database design in the context of other teaching subjects.

The problems that authors are interested in can be analyzed in two main categories. The first category concerns the didactics implications of database use and design, while the second concerns the involvement of students in learning activities using database design in the context of general knowledge subjects (Fessakis, G., & Dimitracopoulou, A., 2003). The second category will not be analyzed farther in this paper. For the first category, the main interest is concentrated to the data modelling phase rather than the data analysis using a ready database. Data analysis and information retrieval concerns this research only to the extend they help in design review and feedback circuitry construction.

3.3. Research questions

For the didactics of database design the main initial research questions of interest are:

- What are students' ideas for manual and digital databases?
- What are the students' ability and difficulties on designing databases?
- What are the difficulties that students face during formal conceptual and logical database design?

In the following sections we present research data collection activities, as well as, analysis of research data. The analyses aim to answer research questions, as well as, to document specific teachers' action strategies propositions.

3.4. Research implementation description

For the clarification of the starting point mentioned previously, we formulated a database curriculum and a series of learning activities has been developed to implement it. The implementation of the curriculum and the research lasted for a school year (2001-2002). Learning activities have the form of lectures and group based, hands on lab activities concerning a corresponding sequence of problems. The proposed set of lectures and activities is one of the action strategies under evaluation in the context of the research. The detailed presentation of the proposed curriculum implementation is out of the current paper's scope. For the needs of the research questions we have designed and implemented a series of research data collection activities. The research data collection activities are described below.

3.4.1. Research activities for the investigation of students' ideas about databases and manual database design difficulties

Two of the research activities where implemented before the final design of the learning activities and concern the investigation of:

- students' ideas about databases, both manual and digital
- the ability and difficulties of students designing manual databases

Students were introduced in the notion of database and information systems using authentic documents from the school's manual database in the context of a short discussion (15 minutes) and then asked to fill a questionnaire with open questions on their ideas about databases. During the next two sessions (2x45 min each) students were asked to design manual databases for three increasing complexity problems familiar to them. The research data that were analyzed are of two kinds: (a) students' questionnaires, and (b) students' paper designs.

3.4.2. Research activities for the investigation of students' difficulties during typical digital database design

During the curriculum implementation students were introduced to Chen's ER model (Chen, 1976) and Codd's Relational model (Codd, 1970) for conceptual and logical design respectively. The instruction was based on the presentation of the two models and their use through problem solving examples and in lab activities, where students designed data bases on paper and implemented those using desktop RDBMS. After instruction, students were grouped and assigned small size projects where they appeared to have difficulties with the relationships' understanding and representation. In order to analyze and understand better these difficulties, two more research data collection activities were implemented in which students were asked to produce conceptual from given logical schemata and vice versa. The research does not aim to reproduce some of the many well-known critics for ER (Hay, 1995), but to propose improvements for the instruction of data base design in secondary education using educational research.

3.4.3. Participants

In the research 11th class students participated from two public schools of Rhodes Greece. Forty one (41) students were from a vocational school named 2nd TEE of Rhodes and seventeen (17) were from the 4th Lyceum of Rhodes. Students from the vocational school were assigned the obligatory subject titled *"Databases"* and participated in all the research activities, while students from the 4th Lyceum were assigned an optional subject named *"Computer applications"* and participated only in the research activities for the investigation of students' ideas about databases The researcher was their normal teacher for both subjects. Research was implemented in real classes under realistic conditions, thus students' number may vary through sessions.

In the following sections, the research data analysis is presented, along with proposed teachers' action strategies for instruction are discussed..

4. Research data analysis

4.1. Students' initial ideas about databases

Modern constructivist learning theory and didactics suggests the study of students' ideas and preconceptions about the learning subject before the design of learning approach (Bransford, Brown, & Cocking, 2000). In order to obtain information about students' ideas and mental representations of databases a questionnaire of open

questions had been designed. Forty-eight (48) students of the eleventh grade have answered the questionnaire.

In the case of databases, most students do not use the typical concepts in their everyday life. Thus, it was considered that it is useful to introduce them in the information systems and database concepts using authentic documents that are familiar to them, like the school manual database. Students were involved in such an activity and they had a short discussion of school manual database and its use. After the accomplishment of the activity, students have filled the above-mentioned questionnaire.

4.1.1. Analysis of students' initial ideas about databases

The most interesting findings from the analysis of students' answers are presented in this section.

Q2.1. A manual database looks alike or resembles …

Answers in this question analysed in five categories as presented in the Table 1:

Table 1. Students' ideas about manual databases.

NO	CATEGORY	STUDENTS
1	CUBBYHOLE WITH FOLDERS AND RECORDS	11
2	SPECIFIC EXAMPLE (restaurant menu etc)	10
3	TABLE	10
4	BOOK	4
5	UNSEASONABLE-AMBIGUOUS-NO ANSWER	13
	TOTAL:	48

Observing Table 1 data it is possible to gain the following arguments:

The majority of students are almost equally distributed to the first three representations. The first two categories are considered more realistic or accurate than the third one, which is driven by students' experiences with spreadsheet software. *"Table"* representation is interesting because it is used as a basic data structure in manual as well as in digital databases. In other words, tables could be vehicles to transfer knowledge and experiences from the concrete world of manual databases to the highly abstract of digital ones. Among the first two ideas, the "cubby-hole with folders and records" (Fig. 1) should be considered more general and appropriate to introduce the basic abstract concepts of information systems and data bases to students.

The fifth category shows that a significant percentage of students (≈27%) face difficulties to express an idea or mental representation of manual databases.

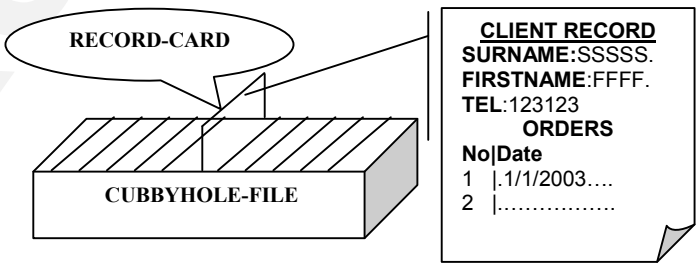


Fig. 1. Manual database as "Cubby-hole with folders and records".

Q2.9. A digital database looks alike or resembles …

Answers in this question were analysed in seven categories as shown in Table 2. The main percentage (50%) of students faces difficulties to express a representation of digital databases. Students in this category are approximately double in comparison to these who do not have a mental representation for manual databases (Table 1). The lack of mental representation for digital databases is the main characteristic of students in the sample. Among the rest of the students *"table"* representations is the most popular. This is due to the experiences that students have with spreadsheet software. This is very interesting for teachers, because relational databases could be introduced in order to overcome inadequacies of spreadsheet software (as the normalized representation of relationships, or the calculation of frequency distribution etc).

Table 2. Students' ideas of digital databases.

No	CATEGORY	STUDENTS
1	TABLE	16
2	TABLE OF MS EXCEL OR MS WORD	3
3	SPECIFIC EXAMPLE (e-addressograph, etc)	2
4	BULLETIN BOARD	1
5	ELECTRONIC BOOK	1
6	CD-ROM	1
7	UNSEASONABLE-AMBIGUOUS-NO ANSWER	24
	TOTAL:	48

Q2.17 A manual database is better than a digital because ….
Q2.18 A manual database is worse than a digital because ….

The above questions estimate students' evaluation of manual and digital databases through their comparison. It is interesting to see students' motivation to use manual or digital databases. Students could state more than one argument. Table 3 shows the main advantages of manual databases over digital according to students (Q2.17). Observing their answers it is obvious that there are

- Some strong arguments as in cases no.: 2, 5 and 9.
- Some opinions that uncover technophobia as in cases no.: 1, 3, 4, 6, 7 and 8.
- Many students that can not evaluate manual databases probably because of lack of experience in use of real world information systems.

Table 3. Advantages of manual over digital databases for students.

A. Arguments of security and integrity		
1	It is not easy to lose data	9
2	They do not need electricity	3
3	Mistakes are fixed more easily	4
4	they are protected easily from unauthorized access	3
B. Usability		
5	They do not need computer use knowledge to be used	5
6	They can be constructed more easily without computer	2
7	Access is easier	3
8	They can be transferred more easily	2
9	From small amounts of data they may be preferable	1
C. Other		
10	They do not have any advantage over digital databases	4
11	Unseasonable-ambiguous-no answer	15
	TOTAL:	52

Table 4 shows the main disadvantages of manual over digital databases according to students (Q2.18). The largest number of total arguments (75) in comparison to Table 3 (52) gives an indication that students may find digital databases better than manual, in general. Most students (37) report the easiest accessibility (insertion, update, deletion, and search) of data as the advantage of digital databases. There are some misconceptions as in cases no.: 2, 4, and 9 and a number of students that can not evaluate the advantages of digital databases. Given that most students lacked experience on digital database use, it is interesting to mention that the evaluation is based on general knowledge of digital technology.

Table 4. Disadvantages of manual over digital databases for students.

A. Arguments of security and integrity		
1	Manual DBs are ruined more easily	9
2	Manual DBs lack security	1
3	Manual DBs do not have automatic spelling correction	1
4	In manual DBs is easier to make mistakes	1
B. Usability		
5	In manual DBs data access is more difficult and time consuming	37
6	It is easier to construct a digital DB	7
7	It is easier to make copies for digital databases	1
C. Capacity		
8	Digital DBs need less space	7
9	Digital DBs have larger storage capacity	1
D. Other		
10	Ecological reasons (Paper use)	2
11	Unseasonable-ambiguous-no answer	8
	TOTAL:	75

4.1.2. Discussion of students' ideas about databases and teachers' strategies

Most students develop a realistic mental representation of manual databases using authentic documents from information systems. Introduction of digital databases abstract concepts is possible to be based on students' ideas for manual databases and spreadsheets that use the notion of *"table"*. Students seem to be familiar with the concept of *"table"*. There are some misconceptions of students concerning databases but the main problem is the lack of experience and ideas. It is purposeful to get students in contact with real information systems in order to estimate the problems' domain they apply, the role of searching and sorting problems in design evaluation, as well as, the problems their use is facing concerning space and time, information retrieval potentialities, reality representation accuracy, etc.

4.2. Difficulties in manual database design for students without previous training.

A DBMS are more than simple productivity tool (e.g. text editor) because it is based on representation systems that students are not usually familiarized with. This is often underestimated when a didactic approach introduces DBMS to students by enumeration of the interface menus. DBMS's are modeling environments and cognitive tools that require familiarization with their representation system in order to be used effectively. Furthermore there is not a straightforward analogy in the design of manual to the digital databases. Manual databases are not restricted to specific data

structures. The designer is restricted only by the paper end his/her imagination. In contrast, in digital databases the designer is restricted to a specific basic data structure (tree, network, relation). With the adoption of relational data model from the industry most of the available DBMS are relational and exploit the table as a basic data structure.

Questions that rise before the introduction of students to a data model include:

- What are the structures that are used by designers without special training?
- What are the main difficulties that they face during design?

In order to collect information for the above questions we use three every day problems of progressively advanced complexity that students had 4x45 minutes sessions to solve. The progression of complexity is based on the number of the entities and the number and kind of relationships among them. The first problem (P1) concerned the design of a manual database for the contact data (phones, addresses etc) of students' friends. The problem concerns one entity and no relationships. The second problem (P2) requests the design of a manual database for the class cashier in order to store data about the students' contributions and class expenses for several activities. The second problem contains two 1-N binary relationships "student-contributes-cashier" and "cashier-finances-class_activity". The third problem (P3) concerned the design of a more complex database for the storage of school's data such as classes, students, teachers etc. The third problem concerns many entities as well as binary and ternary relationships. In order to facilitate the construction of feedback for students designing process each problem had specific questions and reports that should be supported from the design.

Students produced 41 designs for the first problem 39 for the second and 17 for the third. Most students completed designs for problems P1 and P2 in the first 45 minutes. The analysis of the design aims to identify the data structures used by students and the design mistakes.

4.2.1. Data structures used by students in manual database design

Students participated the research did not have any previous instruction on abstract data structures so the variety of the structures they use is determined from their experience. More specifically students proposed the following kinds of structures:

i) Tables:

The majority of designs use table data structure in multi entry form. In problem P1, 25 designs propose a table with a column for each data field of the contact and a row for each contact. In P2, 32 designs propose a table with one column for each month that students must contribute to the cashier and a column for each student. In many cases there are extra columns for computed fields like each student total or the grand total of contributions. Table data structures are used also in the few proposed solutions for P3 but none could be considered as a complete solution.

ii) Records:

Students used the record data structure in many cases. Related to problem P1, 11 designs propose an independent card for each friend and for P2, 3 designs propose a record for each student of the class.

iii) Other:

In one case there is a tree data structure for the representation of class activities that need finance for P2. Finally, there is an indexed table (like the usual address books) in one solution of P1.

It is obvious that students use the structures that they are familiar with, preferring tables and records. Students are familiar with tables from mathematics, every day life (e.g. books), software tools like spreadsheets and/or text editors. It could be interesting to examine if students preserve their perforation to tables and records if they are introduced to other data structures.

4.2.2. Manual database design mistakes and difficulties

The majority of solutions could not be considered correct for any problem. The most correct solutions proposed are for the simplest problem (P1), while only one correct solution was provided for P2, and there is none for P3. Analysis of students' solutions reveals two general categories of difficulties:

i) Difficulties with entities: Students tend to make mistakes as:

- Invention of redundant attributes that are not mentioned in the problem statement
- Drop of attributes that are explicitly mentioned in the problem statement
- Create synonyms using new names for attributes that are named differently in problem statement
- Drop parts of aggregated attributes (e.g. using only street in address attribute)
- Confuse generalized attributes with an instance of them (e.g. using phone attribute to store business and personal phone numbers)
- Confuse attributes with entities (e.g. P1 problem stated that a comment attribute is needed for each contact but some students design a comment store place for the database)

ii) Difficulties with relationships

Students find it difficult to solve problems with multiple relationships like P3. In P2 most students ignored the problems' demand for the design of a monthly report for the state of the cashier. Furthermore, even if most of students faced the relationship "student-contributes-cashier", many of them ignored or failed to represent the relationship "cashier-finances-activities". This supports the hypotheses that students:

- Do not check their proposed designs (fail to construct feedback circuitry)
- Are influenced in relationship understanding by the context in which it is found.

4.2.3. Discussion on students' difficulties designing manual databases and teachers strategies

a) Difficulties with entities: The students' difficulties with entities could be facilitated by using a formal data dictionary during problem analysis and database design. The errors about aggregated and generalized attributes dictate the need for explicit instruction of their treatment, as well as the basic abstraction mechanisms behind the design.

b) Difficulties with relationships: As far as the difficulties about relationships are concerned students seem to be confused when they try to understand the problem in hand and design the proper database in parallel. This could be improved if the problem understanding is separated from the database design phase. Students could produce a more informal description of the problem information content that is needed to be stored producing for example a concept map and then use this description in order to decide how the information is going to be stored in a database. A concept map helps the designer to enumerate the relationships and entities of the

problem, and thus to verify that will not just forget any one. Furthermore, a concept map functions as a cognitive support during the problem understanding and representation, as well as a communication mean among the students..

c) Solution design review and verification: Moreover, in order to facilitate the feedback circuitry construction, students could give their designs to other students for evaluation and/or use software design tools that reduce the time delay between the design and test of the proposed databases. Students are expected to evaluate better their designs if they become conscious of the general searching and sorting problems.

4.3. Students' difficulties during typical digital database design

In the following sections we present the research activities that were implemented in order to clarify the students' difficulties with the relationship concept during database design. Students were asked to transform given relational logical models to conceptual and vice versa. For each activity, the problems assigned to students are presented first; the categories of solutions are presented consequently while at last, the solution categories distribution and MCA are discussed. Analyzing the errors during transformations, we discuss some difficulties' sources and we propose related teachers' actions.

4.3.1. Logical level "Relationships" interpretation

Students were instructed explicitly how to transform conceptual to logical schemata. The reverse process is not a teaching subject, usually. The transformation of logical to conceptual schemata is expected to activate students' understanding of the subject and to produce rich information about their mental models of the related concepts. In this first research activity students were asked to produce ER schemata for given relational ones. Students worked alone for 90 minutes maximum.

Table 5. Logical schemata given to students to produce corresponding conceptual ones.

C1. Single entity schema.	
C1S1	SHOP(**Name**, Address, Telephone, BossName)
C2. Three relations schema for to entities and a binary relationship.	
C2S1	WAREHOUSE(**wCode**, Address) PRODUCT(**pCode**, Description) EXIST_IN(**wCode**, pCode, Quantity, Position)
C2S2	NEWSPAPER(**Name**, Owner, Telephone) ANNOUNCEMENT(**aCode**, Client, Text, Category) PUBLISH(**Name, aCode, Date**, Page)
C2S3	CAR(**cCode**, Model) SPARE_PART(**pCode**, Description) USES(**cCode**, **pCode**)
C2S4	STUDENT(**sCode**, Name) SUBJECT(**Title**, Kind) EXAM(**sCode**, **Title**, **Date**, Time)
C3. Two relations schema for a recursive relationship.	
C3S1	EMPLOY(**ID**, FirstName, SurName, Telephone, Position) MARRIED(**Hasband_ID**, **Wife_ID**)
C4. Four relations schema for three entities and a ternary relationship.	
C4S1	REFEREE(**ID**, Name) TEAM(**Name**, Home) STADIUM(**Stadium_Name**, Address) GAME(**ID**, **HomeName**, **GuestName**, **Date**, Time)

Four categories of logical schemata were provided to students as presented in Table 5. In Table 5, 'primary keys' are formatted bold and underline and 'external keys' have the same name with the corresponding primary keys. Short verbal descriptions specified schemata meaning to students. All binary relationships were given using three tables in order to "hide" the cardinality from students. Students attended instruction on binary relationship translation from conceptual to logical level according to cardinality. The data schemata used are in the students' familiar problems sphere.

4.3.1.1. Solutions analysis

Students' solutions analysis comprised an interesting puzzle. Finally solutions were organized in groups according to students' problem solving strategies. The erroneous solving strategies are analyzed in order to investigate the students' misconceptions about relationship concept.
We have distinguished five Solution categories:

i) Category 1 (C1): «Correct»

C1 contains all the correct solutions. The correctness of cardinality is not evaluated in this research because of the basic difficulties found in the understanding of relationships. Students gave correct solutions for all problem categories except the recursive relationship (C3S1) although they met examples during instruction.

ii) Category 2 (C2): «Attaching relationship properties to entities»

This group of solutions concerns ER schemata with relationship properties attached to entities (Fig. 2). The percentage of these solutions is rather small but it is interesting to mention because corresponding students recognize relationships but they do not find it "normal" to assign properties to them.

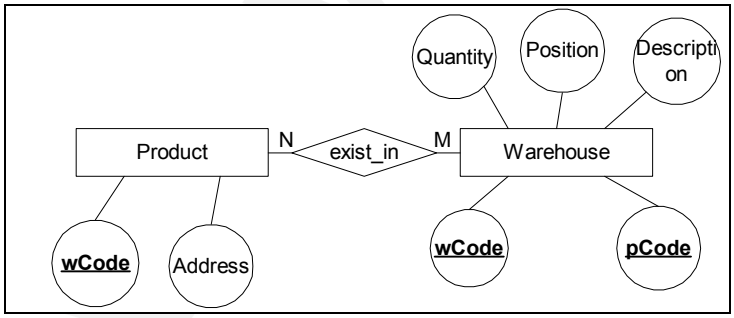


Fig. 2. TA14 student's solution for C2S1 problem as a typical C2 solution. 'Position' and 'Quantity' properties are attached to entity 'Warehouse' instead of the relationship.

iii) Category 3 (C3): Syntactical solutions- 'Making an entity for each relation and device relationships to connect them in order to make readable sentences'

Solutions of this kind are quite descriptive for the students' difficulties with relationships. In this category students propose an entity for each relation of the logical schema and connect them with 'artificial' relationships in order to get the conceptual schema readable as a natural language sentence (Fig. 3).

We call this kind of solution "syntactical". Students giving this kind of solutions are in a good relation to ER model syntax, but they obviously mistake relationship concept. Students treating relationships syntactically use ER as a conceptual map. Furthermore the same student can give a correct solution in a problem and a syntactic in another. This fact denotes that students giving syntactic solutions could be in a

transient level of understanding relationships. The group of syntactic solutions is the most populated.

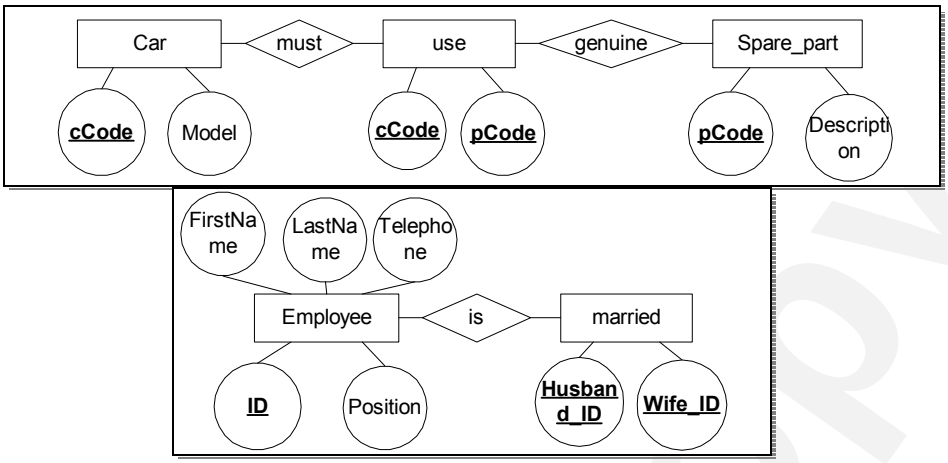


Fig. 3. TA07 and TA11 student's solutions for problem C2S3 and C3S1 correspondingly as typical C3 solutions. There is an entity for each relation and artificial relationships connecting entities constructing a readable sentence.

iv) Category 4 (C4): «Ignoring relationships»

Solutions of C4 ignore relationships. Students in this category produce ER schemata designing an entity for each relation from the logical schema. The solutions contain only entity symbols which are sometimes connected through unlabeled links. Students utilizing this problem solving strategy appear to be unaware of the relationship concept and its representation. They are considered at a lower level of ability to produce sound ER schemata.

iv) Category 5 (C5): «Unclassified»

This group contains solutions that could not join any of C1 through C4 groups. Solutions of this kind use arbitrary entity and/or characteristic names etc. The number of unclassified solutions is from 0 to 3 for each problem totalling 7 solutions (3,74%) so they do not represent a significant percentage. Unclassified solutions where given mostly by irregular attendance students.

4.3.1.2. Summary of solutions' analysis

Table 6 presents the categorical distribution of solutions for each problem and in total. Column labelled 'N.S' presents the number of students that did not give a solution.

Table 6. Categorical Distribution of solutions for each problem and totally.

	C1	%	C2	%	C3	%	C4	%	C5	%	N.S	%
C1S1	24	**88,89**	0	**0,00**	0	**0,00**	0	**0,00**	3	**11,11**	0	**0,00**
C2S1	6	**22,22**	2	**7,41**	15	**55,56**	2	**7,41**	2	**7,41**	0	**0,00**
C2S2	13	**48,15**	0	**0,00**	11	**40,74**	2	**7,41**	0	**0,00**	1	**3,70**
C2S3	13	**48,15**	2	**7,41**	8	**29,63**	2	**7,41**	1	**3,70**	1	**3,70**
C2S4	9	**33,33**	0	**0,00**	13	**48,15**	2	**7,41**	0	**0,00**	3	**11,11**
C3S1	0	**0,00**	0	**0,00**	22	**81,48**	4	**14,81**	1	**3,70**	0	**0,00**
C4S1	1	**3,70**	1	**3,70**	20	**74,07**	1	**3,70**	0	**0,00**	4	**14,81**
TOTAL	66	**34,92**	5	**2,65**	89	**47,09**	13	**6,88**	7	**3,70**	9	**4,76**

Some remarkable points:

a) Observing column C1 (Correct Solutions)

- Students find it more difficult to interpret ternary relationships (C4S1) than usual binary relationships (C2S1- C2S4).
- Problem seems to effect on students' performance for binary relationships since the percentage of correct solutions in this category varies with problem (C2S1- C2S4).
- Recursive binary relationships (C3S1) is a serious problem for students since none gave a correct solution despite the examples during instruction.

b) Observing column C3 (Syntactical solutions)

- Most students do not understand the representation of relationships using foreign keys in relational schemata and treat relationships syntactically.
- When the difficulty increases and the percentage of correct solution decreases students give more syntactical solutions. This finding supports the hypothesis that students producing syntactical solutions may be in a transient level of relationships understanding and backtrack when difficulties increase.
- Students giving syntactical solutions are concerning ER schemata as concept maps where relationships are more informal and arbitrary.

c) Observing column C4 (Solutions ignoring relationships)

- From problems C2S1-4 this kind of solutions where given from the same two students but for problem C3S1 this kind of solution proposed from two more students. It seems that ignoring relationships is a first level for relationships understanding where students backtracked when the difficult problem of recursive relationship occurred.

From the above analysis it seems that students could be classified in the following ***three levels of understanding*** the "relationship" concept and the 'foreign key' representation:

Level A. Ignoring relationships

Level B. Syntactical treatment of relationships.

Level C. Representing relationships correctly or almost correctly.

Students of a certain level is possible to backtrack to a smaller ability level and utilize a less sound problem solving strategy depending on the problem difficulty.

The above descriptive analysis can not describe the behaviour of students among different problems. For example the question "is there a group of students that systematically produces syntactical solutions?" is not answered. In the next section a MCA will bring in light student behaviour patterns and evidence for the rationalisation of the descriptive analysis.

4.3.1.3. Multiple Correspondence Analysis

We have applied Multiple Correspondence Analysis (MCA) method on the data related to logical relationship interpretation. In this MCA, we use CiSj as active variables. Each CiSj variable contains as value the category of solution (*C*) for the corresponding problem (Table 5) for each student (*S*).

Interesting findings come up observing the first three axes (factors) resulting from MCA containing 60 % of total information (Table 7).

Table 7. MCA parameters' values (the first five factors)

FACTOR	EIGENVALUE	COEFFICIENT OF INERTIA	CUMULATIVE PERCENTAGE
1	0,7464	23,75	23,75
2	0,6975	22,19	45,94
3	0,4431	14,10	60,04
4	0,2620	8,34	68,38
5	0,2045	6,51	74,88

First axis (23.75% of total information) illustrates the contrast between two groups of students. First group answers have been categorized as *"No Solution -NS"* (for problems C2S4, C2S2, C4S1) and *"Unclassified – C5"* (for problems C1S1, C2S3, C3S1, C2S1) while the other group *ignores relationships* (C4 – class of solutions) for problems C4S1, C2S2, C2S4, C2S1 and C2S3 and answers *correct* (C1-class of solutions) for problem C1S1. Given that C1S1 is a trivial problem without relationships it seems that first axis contrasts students that do not give answers to them produce answers ignoring relationships representation. In other words, first axis illustrates a boundary between level A of ability and the students with significant problems with database design.

Second axis (22.19% of total information) brings in light two more groups of students in contraposition. The first group produces *correct* (C1) solutions for problems C1S1, C2S3, and C2S2 while treats *syntactically* (C3) the ternary relationship in problem C4S1. The other group *does not produce solutions* or *ignore*s relationships for problems C2S2, and C2S4, or produces *unclassified* solutions for problems C1S1, C2S3 and C3S1. In general, the second axis discriminates students that produce some correct solutions or "syntactical" solutions (for the "difficult" case of ternary relationship in problem C4S1) from students that face more or less significant problems with database design.

The **third axis** (14.1 % of the total information) uncovers one more significant contrast in students' solutions. The axis discriminates one group of students that produce *correct* solutions for all problems except C3S1 which contains the recursive relationship from another group that produces *"Syntactical"* solutions for problems C2S1, C2S2, C2S3, C2S3, C2S4 and C3S1. This axis illustrates a boundary between student of level B and level C ability.

From the above analysis four main students groups arise with specific cognitive behaviour in the solution of CiSj problems. These groups appear in Fig. 4.

First group (2nd quadrant) contains the students that produce unclassified solutions or they do not produce any solution. Students of this group face serious cognitive problems concerning database design.

Second group (1st quadrant) contains students that ignore relationships representation. Students of this group face understanding problems concerning mainly the relationship concept and are classified in the level A of database design ability.

Third group (3rd quadrant) consists of students that are conscious of the relationships but they treat them "syntactically" (C3-category). Students of this group face problems with understanding of relationships' representation in the logical level of database design. Some students of this group produce C2 category solutions for problem C2S3. These transitions of students between solution categories are the main

reason of the ability levels proposal. Students of this group correspond to the ability level B.

Fourth group (4th quadrant) concerns students that in general produce correct solutions. Students of this group have a good degree of understanding relationships and their representation during database design and correspond to the ability level A. The figure shows that even these students found it difficult to represent correctly the recursive relationship of problem C3S1. Students of this group backtracked to level B in the case of the "difficult" problem and produced syntactical solutions.

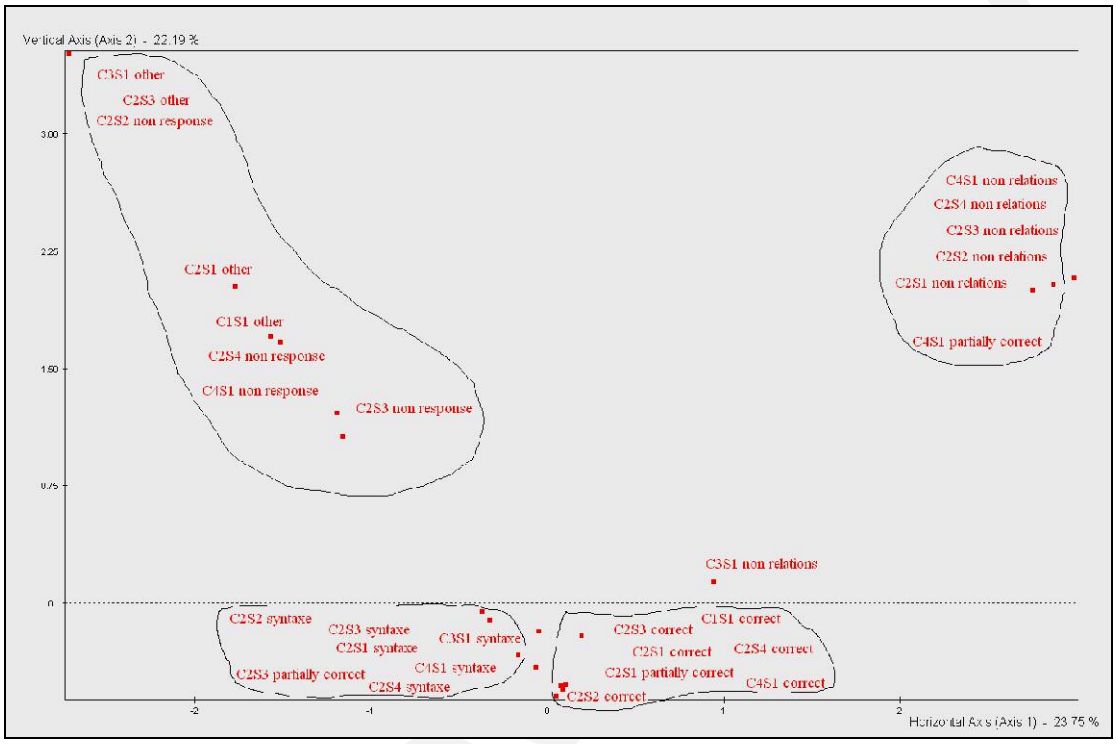


Fig. 4. Multiple Correspondence Analysis with the CiSj variables (the 2 first factors)

Concluding the MCA for this activity gives evidence for the claim that students' problem solving ability can be described by a three transition level hierarchy in accordance to their understanding of relationship concept and representation. The systematic ignoring, syntactical treatment, and correct representation by corresponding students group shows that these kinds of solutions are not accidental but originate by corresponding levels of relationship understanding. MCA analysis rationalizes the interpretation of descriptive analysis presented in the previous section.

4.3.2. Conceptual level "Relationships" interpretation

In the second activity students were asked to produce Relational from ER schemata. Table 8 shows the ER schemata given to students categorized according to the kind of their relationships. In order to understand the analysis it is useful to summarize the methodology that was instructed to students. Students were instructed to produce relational schemata using the following rules:

1. *"1-1" binary relationships could be represented using three alternative solutions:*
   a. One table solution (1T): Create one table using all attributes from the related entities along with the relationship's attributes if any. Use any from the entities' keys as primary key (alternative keys).

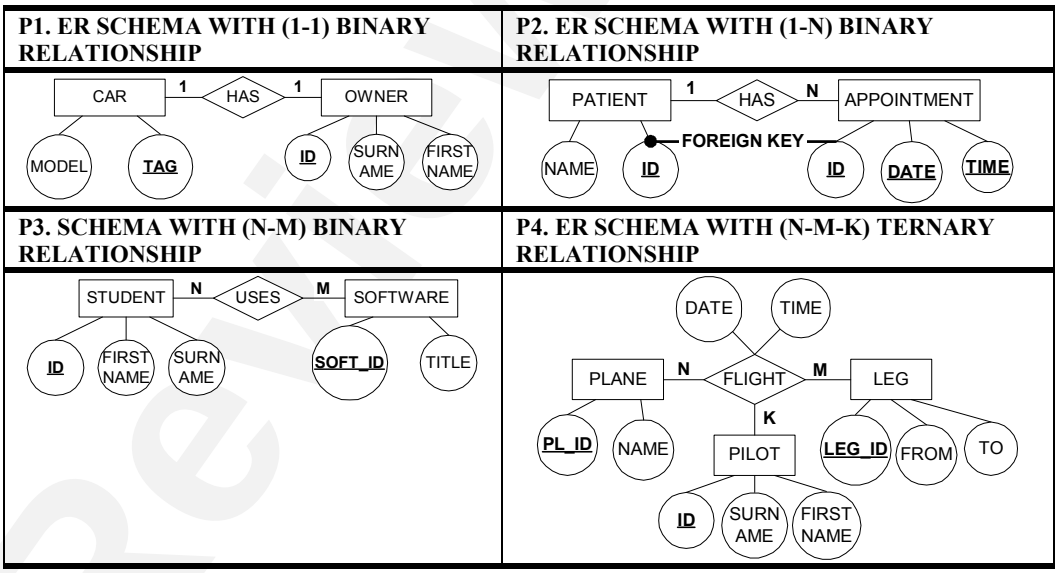
b. Two tables' solution (2T): Create a table for each entity preserving their primary keys. Choose one table and add as foreign key the primary key of the other along with the relationship's attributes if any.

c. Three tables' solution (3T): Create one table for each entity preserving their primary keys. For the relationship create a table using the primary keys of the tables created so far along with the relationship's attributes if any. The last table has a composite primary key consisting of the primary keys of the entities; furthermore any primary key of the entities is a foreign key.

2. *"1-N" binary relationships could be represented using two alternative solutions:*
   a. Two tables' solution (2T): Create a table for each entity preserving their primary keys. To the table corresponding to the entity at the relationship side labeled (N) add the primary key of the entity at the relationship side labeled with cardinality (1) as foreign key along with the relationship's attributes if any.
   b. Three tables' solution (3T): Same as rule 1.c. above.
3. *"N-M" binary relationships could be represented using the following solution:*
   Three tables' solution (3T): Same as rule 1.c. above.
4. *Many-member relationships could be represented* using the rule 1.c. above with the difference that the number of tables produced is equal to the number of entities plus one for their relationship.

The set of the above rules has been created during instruction using examples and alternative solutions. The alternative solutions were rejected using normalization criteria. Using the above rules the designer is able to produce relational schemata with a good degree of normalization.

Table 8. ER schemata given to students for transformation to relational one



The analysis of students' solutions is presented in the following paragraphs. The analysis of the students' solutions uncovers the problem solving strategies that students adapted after the instruction. Students' problem solving strategies are described and evaluated. It will be shown that students are mainly ignoring relationships producing a table for each entity or applying in a rote manner the 1.c. rule because of their difficulties with the relationship concept.

4.3.2.1. Solutions analysis

Students' solutions are of the following categories of decreasing ability to relationships' representation:

i) Category 1 (C1). Correct

Solutions of this category contain tables for the representation of entities as well as relationships and mention foreign keys that implement them. It is interesting to mention that according to the methodology instructed to students, problem P1 could be treated with three kinds of solutions each containing from 1 to 3 tables. Table 9 shows the distribution of correct solutions to the number of tables in the solution. The representation of a 1-1 binary relation using three tables could be characterized in general as unusual or impractical. It is possible that those students choose to memorize the rule 1.c because it seems to solve all cases. The solutions volume (13) containing 3 tables for problem P2 that could be solved using 2 tables support the above hypothesis.

Table 9. Number of tables in correct solutions for P1.

TABLES	SOLUTIONS
1	1
2	1
3	11

ii) Category 2 (C2). Inadequate relationship representation

Solutions of this kind propose relationship representation with minor or more significant errors. Some typical errors concern addition of arbitrary fields and/or elimination of others. Some students assign relationship attributes to entities because as in the previous activity. There are two interesting cases to mention in more detail:

- Problem P2 states a case where the application of rule 1.c. creates a redundant table (Frame 2) because of the explicit representation of foreign keys. Some students instead of eliminating the redundant table proceed to the invention of a simple key for the "APPOINTMENT" entity (Frame 1). Students often choose to change the problem than to adapt its solution.

Frame 1. Typical solution with the invention of a simple key for P2

Table for the 'Patient' entity: T1. PATIENT(#ID, NAME)
Table for the 'Appointment' entity: T2. APPOINTMENT(#AP_ID, DATE, TIME)
Table for the relationship: T3. HAS(#ID, #AP_ID)

Frame 2. The relational schema produced by rote application of rule 1.c.

Table for the 'Patient' entity: T1. PATIENT(#ID, NAME)
Table for the 'Appointment' entity: T2. APPOINTMENT(#ID, #DATE, #TIME)
Table for the relationship: T3. HAS(#ID, #ID, #DATE, #TIME)

- For the P4 (N-M-K relationship) problem some students' solutions contain 5 tables. The fifth table appears because students include relationship's attributes in a different table or because they apply rule 1.c. for each pair of related entities.

iii) Category 3 (C3). Ignoring relationship

Solutions in this category contain a table for each entity without any foreign key and no representation for the relationship. Solutions of this kind represent a significant percent.

4.3.2.2. Summary of solutions analysis

Table 10 presents the categorical distribution of solutions for each problem and in total. The column labeled 'N.S' presents the number of students that did not give a solution.

Table 10. Categorical Distribution of solutions for each problem and in total.

	C1	%	C2	%	C3	%	N.S	%
P1	13	**48.15**	2	**7.41**	12	**44.44**	0	**0**
P2	14	**51.85**	13	**48.15**	0	**0**	0	**0**
P3	13	**48.15**	5	**18.52**	9	**33.33**	0	**0**
P4	4	**14.81**	11	**40.74**	9	**33.33**	3	**11.11**
TOTAL	44	**40.74**	31	**28.70**	30	**27.78**	3	**2.78**

Some remarkable points:

- Observing C3 (Ignoring relationships) column it is interesting to analyze the 0% for P2. In P2 the ER schema has an explicit representation for the foreign key. Most correct solutions for P2 belong to students that systematically ignore relationships! Students that ignore relationships come up with correct solutions just by accident because of the explicit representation of the foreign keys and the cardinality of the binary relationship that is possible to be represented using a table for each entity. Students that utilize the solution with three tables face a surprise because of the redundant table they produce (Frame 2). Students that apply methodology rules in a rote manner do not review their solutions and propose inadequate relationship representation.
- Observing C1 column it is obvious that students face increasing difficulties with ternary relationships. Although the same methodology rule applies as well as in problem P3 only four students come up with correct solutions. Correct solutions are significantly reduced and uncover a small core of students that understand the relationship concept and the meaning of the logical design. Most students either ignore relationships or apply in a rote manner the rule 1.c.

Taking into account the students' problem solving technique, we could classified them in increasing levels of relationship concept understanding as follows:

Level A. Ignoring relationships
Students of this level ignore relationships and represent only entities.

Level B. Inadequate relationship representation
Students in this level are applying the methodology rules in a rote manner, preferring the rule 1.c. and/or do not like to give attributes to the relationship table adding them to entity tables. Finally, some students of this level eliminate or add arbitrary fields.

Level C. Sound understanding of relationships
Students in the third level understand the semantics of the relationships and represent them correctly during the logical design of the database.

The three levels of this activity are in correspondence with the levels proposed in the previous activity where students asked to transform relational to ER schemata. In the following section we use MCA in order to identify students groups with coherent problem solving strategies in accordance with the above analysis.

4.3.2.3. Multiple Correspondence Analysis

In this case, we use Pi as active variables for MCA. For each student Pi contains the category of answer for the corresponding problem (Table 8). Interesting findings can be concluded analyzing the first two axes (factors) (Table 10) that results by MCA and represent 58.43% of total information.

First axis (39.71% of information) brings in light two main groups of students with contraposition in behavior. The first group of students ignore relationships in P1, P3 and P4 while produce correct solution for P2. Students of this group ignore systematically the representation of relationships designing a relation for each entity in the ER and produce correct solutions for P2 by accident because of the explicit representation of foreign keys. In contrast, the second group of students applies the transformation rules in a rote manner and produce correct solutions for problems P1 and P3 while they fail in the special case of P2 as explained in the previous analysis. Most students of the second group fail in P4 because it contains a ternary relationship, which has been already classified "difficult" for most students.

Second axis (18.72% of total information) discriminates the group of students that produce correct solutions for P1, P3 and P4 and the group of students that produce unclassified solutions for P1. This axis lights a boundary between students that produce correct solutions applying in a rote manner the rules and those who do not produce solutions.

Table 10. MCA parameters' values (the first five factors).

FACTOR	EIGENVALUE	COEFFICIENT OF INERTIA	CUMULATIVE PERCENTAGE
1	0,7943	39,71	39,71
2	0,3743	18,72	58,43
3	0,3203	16,02	74,45
4	0,2500	12,50	86,95
5	0,1596	7,98	94,93

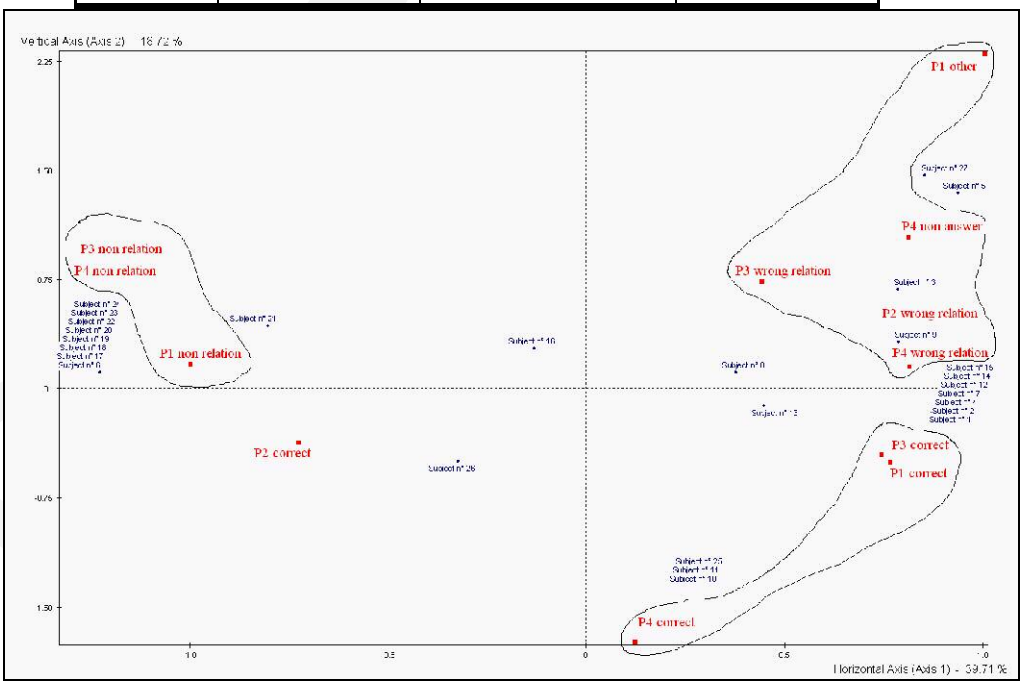


Fig. 5: Multiple Correspondence Analysis with the Pi variables (the 2 first factors)

Examining Fig. 5 produced using the first two factors we perceive three main students groups. **First group (2nd quadrant)** ignores relationships to P1, P3, and P4 while produce correct solution for P2. This group presents A' level ability in relationship representation. **Second group students (4th quadrant)** produces correct solutions for P1, P3 and P4 applying in a rote manner the rules and presents B level of relationships representation. **Third group (1st quadrant)** contains students that do not produce solutions for P4, represents incorrectly relationships for P2, P3 and P4, or produce unclassified solution for P1. In other words the third group of students did not adapt a specific pattern of relationships representation. **Finally, as a fourth group (3rd quadrant)** we have students that produce correct solution for P2. A small percentage of these students belong to the small group of students that produce correct solutions to all problems (ability level C) but most of them ignore systematically relationships.

4.3.3. Interrelating analysis of students' behavior in relationship interpretation to the logical and conceptual level

It is actually interesting to combine the students' behavior in the two research activities concerning the transformation of relational to ER schemata and vice versa in order to identify possible coherent patterns of behavior according to the understanding of relationship concept and representation. In order to obtain the above goal, we apply MCA, using as active variables both CiSj and Pi. This analysis helps to describe relations between students' cognitive difficulties for problems CiSj with difficulties for Pi problems.

Table 11. MCA parameters' values (the first five factors).

FACTOR	EIGENVALUE	COEFFICIENT OF INERTIA	CUMULATIVE PERCENTAGE
1	0,5022	18,42	18,42
2	0,4765	17,47	35,89
3	0,4255	15,60	51,49
4	0,2643	9,69	61,18
5	0,2083	7,64	68,82

First axis (18.42% of total information) discriminates two groups of students. The students in the first group do not produce solutions (C2S2, C2S4, C4S1), produce unclassified solutions (C3S1, C2S3, C2S1, C1S1), do not represent correctly the relationship of P4 and represent correctly relationship of P3. It is obvious that this group concerns the students that face significant cognitive difficulties with database design. It is interesting to mention that students who rather fail for problems CiSj it is possible to produce occasionally even correct solutions for Pi. This is because of the instructed methodology for Pi problems in contrast to CiSj. This fact justifies the selection of CiSj problems as research data collection activity.

The second group of students ignores relationships (C4S1, C2S2, C2S4, C2S1, C2S3) while produces correct solution for the trivial problem C1S1 which do not concern any relationship. First axis draws a boundary between students that face most serious problems with database design and students that ignore relationship representation.

Second axis (17.47% of total information) illustrates the contrast between the following groups. In the first group students treat syntactical the relationships in C2S4

and the difficult problems C4S1 and C3S1 while they produce correct solutions for C1S1, C2S3 and C2S2. Students of the first group can solve some "easy" problems while backtrack to the syntactical solutions for more "difficult" problems. These students do not present any specific pattern of behavior for Pi problems because of the methodology instruction. The second group of students, in general, does not produce solutions for problems C4S1, C2S2 C2S4 and P4, produce unclassified solutions for C2S3 and C3S1 while they ignore relationships for C2S2, C2S1, C2S4, C2S3. The second group of students has more significant problems with relationships than the first. Second axis draws a second boundary of ability between students' relationship understanding while it does not present any significant interrelation between students' behavior for CiSj and Pi.

The **third axis** is interesting because it identifies a pattern of students' behavior for the two kinds of problems. Third axis uncovers a first group of students that treat syntactical relationships for KiSj (C2S2, C2S3, C2S4, C3S1, C2S1) and ignore or represent inadequate relationships for Pi (P1, P3, P4). In other words most students that treat syntactical relationships in CiSj fail to represent relationships for Pi. This is evidence that a significant percentage of students do not understand relationships semantics in database design and treat them informally like in concept maps. The second group that the third axis uncovers concerns students that produce correct solutions in P1, P3, P4, C2S1, C2S4, C2S3 and C2S2 while fail in the difficult cases of P4 (ternary relationship) and C3S1 (recursive relationship). This group consists of a small number of students that understand the notion of relationships and its representation in the context of database design. This group is evidence for the reasonable rule that students who solve CiSj solve also Pi with exception of the difficult cases of ternary and recursive relationships. The difficulties with ternary and recursive relationship have been mentioned also in the descriptive analysis and MCA validates them.

Fig. 6 deals with the first and third factor and displays four main students groups that loosely correspond to the groups produced by the three first axes analysis.

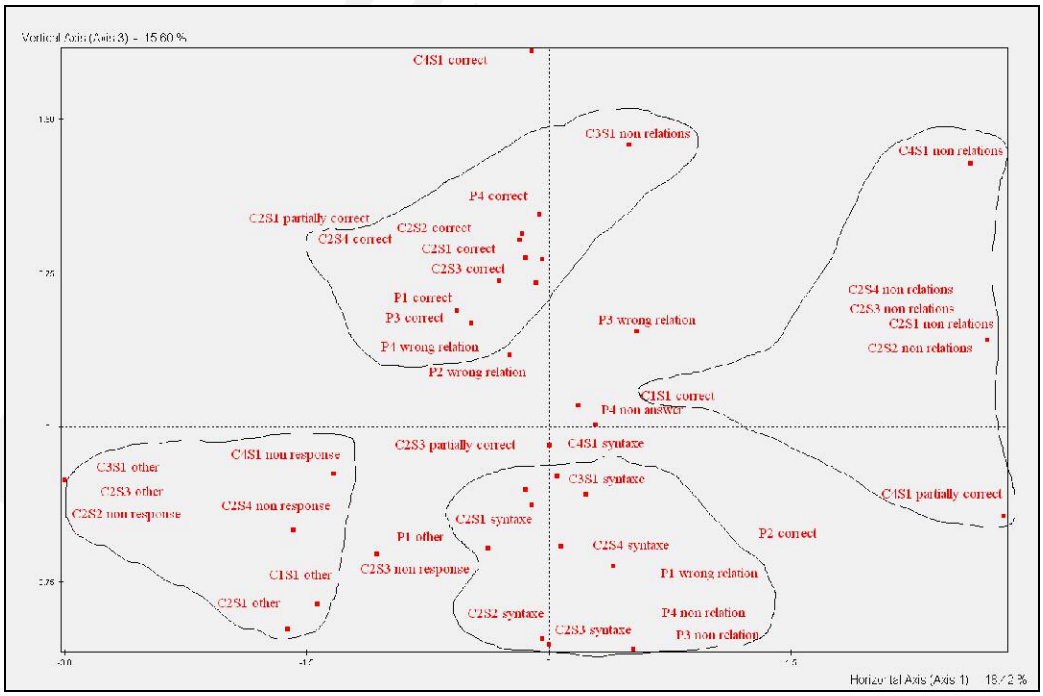


Fig. 6: Multiple Correspondence Analysis with the CiSj & Pi variables (plan formed by the first and the third factors)

First group (3rd quadrant) concerns students that do not produce solutions or produce unclassified ones while solve P3 and fail to represent correctly ternary relationship in P4.

Second group (4th quadrant mainly) concerns students that produce "syntactical" solutions for CiSj while fail to solve Pi (except for P2 which is a special case).

Third group (1st quadrant mainly) concerns students ignoring relationships.

Fourth group (2nd quadrant mainly) concerns students that produce correct solutions for CiSj as well as for Pi except for the difficult cases of P4 and C3S1.

The MCA demonstrates interrelations between students' behavior in the Relational to ER schema interpretation and vice versa activities. MCA gives evidence for claims of the descriptive analysis as well as new facts.

More specifically, there is evidence that:

- Students *face increased difficulty with ternary and recursive relationships*.
- Students seem to *belong in hierarchy of ability levels* for both kinds of activities according to their understanding of relationships. The first level of the hierarchy concerns students that ignore relationships' representation, the second level concerns students that face more or less serious difficulties in relationships understanding while the last level contains students that understand relationships and their representation in the framework of database design. Students of one specific level of ability backtrack to a lower ability level in case of difficult problems.

New claims by the last MCA concern that:

- A small number of students present a good understanding of relationships since they solve both kinds of problems except the difficult ones.
- A significant percentage of students that produce syntactical solutions for CiSj ignore relationships representation for Pi. These students have a real cognitive difficulty with relationships since their errors are not incidental. Most of these students produce by accident correct solutions for P2 where foreign keys are explicitly mentioned.
- Some students that ignore relationships for CiSj are producing solutions for Pi applying the methodology they were instructed. The success of their tries varies depending on student and problem and does not seem to follow any specific pattern. This means that the understanding of relationships can be replaced by a simple methodology. Database design is a cognitive assuming modeling process.

4.3.4. Conclusions on students' difficulties during digital database design and proposed teachers' action strategies

Combining the findings from both activities presented previously it is possible to rationalize a set of action strategies for the improvement of database design teaching for secondary education students:

4.3.4.1. Students' difficulties during digital database design

According the research activities data there is evidence that:

A). Most students treat relationships syntactically and use ER as a kind of conceptual map.

During the initial phase of database design the problem is under ontological analysis (Perakath, 1994). The database design process, as usually, presented to students merges ontological analysis with conceptual design using ER model. In other words, during conceptual design students are faced with two mental challenges:

i). The problem domain understanding (recognition of the concepts-entities, their characteristics, decision about the appropriate detail level, synonyms clarification etc).

ii). The detailed and formal specification of the information needs of the problem. That is the specification of what information is going to be stored in the data base.

The confrontation of two tasks (problem understanding and information content specification using ER model) simultaneously is considered a heavy duty for young students. Thus, it is reasonable to propose the separation of the two problems using concept maps for problem understanding and a conceptual model for database design.

B). Students ignore that relationships' representation produce correct relational schemata from ER ones that explicit mention foreign keys. In relational model, relationships are implemented using foreign keys that are fields working as references between tables. The representation of foreign keys in ER model is practically optional. This confuses the relational schema production. If the ER schema represents explicitly the foreign keys and there are only binary relationships without attributes, most students could produce a correct logical schema. The production of relational schemata for given ER ones is important in order for students to obtain feedback and review their designs. A didactically proper conceptual model should impose the foreign key representation.

C). Students face difficulties in understanding relationships' semantics and representation especially in the cases of recursive and ternary relationships. Students need a more tangible relationship representation, for this purpose it is reasonable to propose the relationship concept introduction using a lower level representation as the tuple sets (Fig. 7). In addition, the understanding of relationship misconceptions could be based on feedback from the logical level according to normalization criteria. This observation recommends the automation of logical schema production for the conceptual (and vice versa) in order to get feedback as soon as possible for the meaning of their designs.

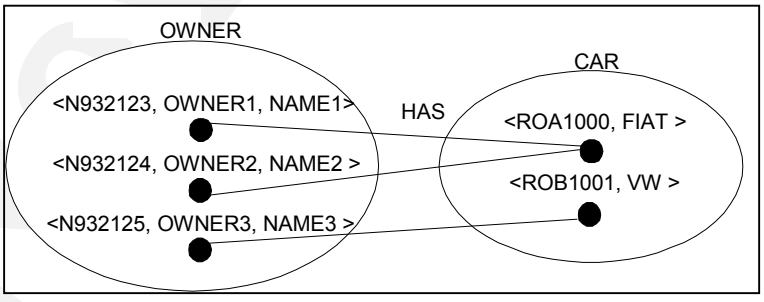


Fig. 7. Tuple set representation of the relation and relationships concepts.

4.3.4.2. Proposition of IDEF1X as didactically appropriate conceptual model

The above analysis presents main difficulties of secondary education students during database design using ER conceptual model. Most of the researches mentioned in section 2 that concern the treatment of such difficulties propose actions regarding the learner while they leave the conceptual model as is. At this point, we will adopt a different approach. Instead of considering ER as a constant we will propose the replacement of it by a conceptual model compatible to didactical requirements that can be defined from the research data analysis. This approach is inspired from considerations included in the following quotations:

> *"Confusion and clutter are failures of [drawing] design, not attributes of information. And so the point is to find design strategies that reveal detail and complexity rather than to fault the data for an excess of complication. Or, worse, to fault viewers for a lack of understanding. (Tufte, 1990)."*

> *"Data models are vehicles for describing reality. Designers use data models to build schemata which are representations of reality. The quality of the resulting schemata depends not only on the skill of the database designers, but also on the qualities of the selected data model. (Batini, Ceri, & Navathe, 1992, pp. 15)."*

From the previous analysis we can conclude that a Conceptual Model should fulfil at least the following didactical requirements:

i). Permit the automatic conceptual to relational translation and vice versa in order to facilitate feedback.

ii). Use only binary relationships without attributes

iii). Impose the explicit representation of the foreign keys in the conceptual level and systematize their introduction to the conceptual schema reducing the problem of foreign key definition to a proper relationship selection decision.

Searching for conceptual data models consistent with requirements raised by the above observations could result IDEF1X like models. The properties of IDEF1X that confront the above observations are presented next.

The ER model as proposed by Chen is an informal model that addresses the need of databases design independently from the logical data model. At the time ER was proposed the relational model was not accepted as widely as nowadays. Furthermore, relationship notion of the ER model has been fairly reproved and improved conceptual models have been proposed (Hay, 1995). One such conceptual model that concentrates characteristics compatible to the didactical requirements above is IDEF1X. IDEF1X is widely accepted for relational database design and is an official standard in USA (Federal information Processing Standards Publication 184, 1993). Furthermore there are many software tools available that support IDEF1X notation.

A detailed presentation of IDEF1X is out of the purpose of the paper. In the following section there is a brief description of the main characteristics of the model.

Entities

IDEF1X entities are of two kinds, independent (or parental) and depended (or child). The primary key of the dependent entities include at least a relationship with another entity meaning that it is composite key containing at least a foreign key. Graphically dependent entities are different from independent. The graphical symbol of an entity shows the key attributes separated from the simple ones.

Relationships

IDEF1X relationships are binary and asymmetric and vary according to cardinality. From each relationship the designer can define two tags according to the direction of the reading. The foreign key that implements a relationship is explicitly mentioned on the conceptual schema. IDEF1X relationships can not have attributes. IDEF1X schema transformation to relational and vice versa is trivial.

A small example of IDEF1X's use

For the evaluation of IDEF1X consistency with the didactic requirements a simple problem is adequate. Next figure pictures the definition of the IDEF1X schema for the 1-N variation of problem P1 (Table 8). First the two entities are defined and then a relationship is established between them using drag and drop. Software tools usually add automatically the foreign key to the dependent entity which is marked with rounded squares. It is obvious that foreign keys are explicit in IDEF1X and they concern the designer at the conceptual level making relationship semantics clear.

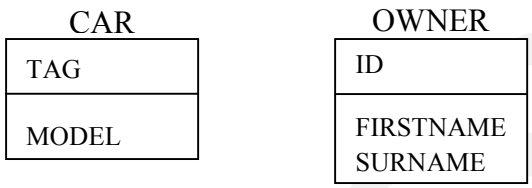


Fig. 8. IDEF1X use. Entities before relationship definition

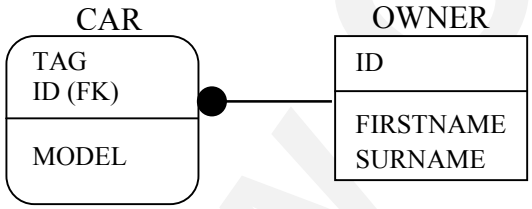


Fig. 9. IDEF1X use. Entities after relationship definition

For the learning value of the above example consider a student that produces syntactical solutions. This student will come up with a correct solution or he/she through the automatic production of the corresponding database (logical level feedback) probably will find out soon that the proposed conceptual schema does not represent the problematic situation.

5. Discussion

Databases design learning in secondary education is interesting because of utilitarian and didactic reasons. The effective introduction of database design in secondary education needs thorough research. In this direction, we designed and implemented educational action research aiming to identify the learning difficulties and then to improve database design instruction. The main findings of this research as well as the corresponding proposed didactic implications are summarized below.

- **Students' ideas about databases**

Many students can not express ideas about databases. Students could formulate a quite realistic and functional mental representation of databases using authentic documents from manual information systems. Manual database systems can make concrete many of the highly abstract concepts and procedures of the digital databases. Students' ideas about databases are affected by the data handling facilities of software they are familiar with. Students are familiar with the concept of table, which they

know from the spreadsheet software. Relational databases introduction to students could be based in the inadequacies of spreadsheet software.

- **Students designing manual databases**

Students designing manual databases use a small number of data structures with main representatives the structures of 'table' and 'record'. This means that despite the freedom that paper gives to designer students are constrained to structures they are familiar by the software they learn to use. Students' difficulties designing manual databases concern usual difficulties related to attribute elimination etc as well as to relationships representation. The use of data dictionary could help to overcome the usual problems while the relationships representation could be improved using concept maps for the problem analysis documentation. Young designers often do not review their designs so it is proposed to organize them in groups, which review each other's designs. Finally, in order to improve students' ability to evaluate database designs we propose the instruction of searching and sorting using manual databases.

- **Students' difficulties with typical digital database design**

Most students designing digital relational databases using ER in the conceptual level face difficulties with the relationship concept understanding and representation. Analyzing their difficulties during conceptual to logical schemata transformation and vice versa we found that most students either ignore relationships representation or treat them "syntactically" as in the case of concept maps while a small fraction of them understands relationships and their representation. These main students groups define three levels of ability and understanding of relationships in the framework of database design.

Furthermore students' difficulties increase with the number of related entities in a relationship as well as in special cases like the recursive binary relations. For the clarification of relationships semantics and their representation we propose:

- The use of concept maps in order to document the problem analysis and specification (what to store) before the database design phase (how should be stored)
- The use of low level representations like tuple sets for the introduction of relationship concept in order to visualize them and their characteristics.
- The use of didactically proper conceptual data model instead of the traditional ER which:
    - Permits the automatic conceptual to relational translation and vice versa in order to facilitate feedback.
    - Uses only binary relationships without attributes
    - Represents explicitly the foreign keys in the conceptual level and systematize their introduction to the conceptual schema reducing the problem of foreign key definition to a proper relationship selection decision.

The presented action research is going to be continued implementing the teachers' action strategies proposed in order to evaluate their effectiveness.

6. References

Altrichter, H., Posch, P., & Somekh, B. (1993). Teachers investigate their work. An introduction to the methods of action research, Routledge

Antony, S., & Batra, D. (2002). CODASYS: A Consulting Tool for Novice Database Designers, ACM SIGMIS Database, 33(3), 54-68

Bagnall, L. (1994). Tabletop and Tabletop Jr: Two tools for hands-on data exploration for kids, ACM CHI-94, 63-64

Batini, C, Ceri, St., & Navathe Sh. (1992). Conceptual database design, Benjamin Cummings PC inc

Batra D., & Davis, G. (1989). Conceptual database design by novice and expert database designers, Proceedings of the Tenth international conference on information Systems, 91-99

Batra D., Hoffer J., & Bostrom R. (1990). Comparing Representations with Relational and EER Models, Communications of the ACM, 33(2)

Benzécri, J. P. (1992). Correspondence analysis handbook. New York: Marcel Dekker.

Bransford, D., J., Brown, L., A., & Cocking, R., R. (Eds) (2000). How people learn. Brain, Mind, Experience, and School, National Academy Press. Washington, D.C.

Brosey, L., & Shneiderman, B. (1978). Two experimental comparisons of relational and hierarchical database models, international Journal of Man Machine Studies,10, 625-237

Chen, P (1976). The Entity-Relationship model - toward a unified view of data, ACM transactions on database systems, 1,(1), 9-36

Codd, E. (1970). A Relational model of data for large shared data banks, Communications of the ACM, 13(6), 377-387

Date C. (1990). An introduction to Database Systems, 5th Ed., Addison-Welsey Inc

Durding, M., Becker, A., & Gould D., (1977). Data organization., Human Factors, 19(1), 1-14

Edward R. Tufte (1990). Envisioning information, Cheshire, Connecticut: Graphics Press

Federal information Processing Standards Publication 184 (1993). INTEGRATION DEFINITION FOR INFORMATION MODELING (IDEF1X), http://www.essentialstrategies.com/publications/modeling/idef1x.htm (recent access, 2-Sep-03)

Fessakis, G., & Dimitracopoulou, A., Exploitation of data modeling for database design in secondary education learning activities: A case study concerning real stories analysis., Interactive Computer Aided Learning (ICL) 2003, Carinthia Tech Institute, 24-26 Sep 2003 Villach, Austria

Fessakis, G., Dimitracopoulou, A., & Halatsis, C., Secondary education students' difficulties on database design and remedial teaching strategies, Second International Conference on Multimedia and ICTs in Educations (m-ICTE 2003), University of Extremadura Spain/Formatex Research Center, 3-6 Dec 2003, Badajoz, Spain

Goldstein, R., & Storey, V. (1989). Some findings on the intuitiveness of entity-relationship constructs, Proceedings of the 8th international Conference on Entity-Relationship Aproach to Database Design and Querying, 9-23

Hall, L., & Gordon, A. (1998). a virtual environment for Entity Relationship modelling, Proceedings of ACM SIGSE, 345-349

Hancock, C., & Kaput, J. (1990). Computerized tools and the process of data modelling in G. Booker, P. Cobb, & T. deMendicutti (Eds.), Proceedings of the 14th Conference of the international Group for the Psychology of Mathematics Education, Mexico, 3, 165-172

Hancock, C., Kaput, J., & Goldsmith, T. (1992). Authentic inquiry with data: Critical barriers to Classroom Implemenation., Educational Psychologist, 27 (3), 337-364

Hay C. D, (1995). A comparison of data modelling techniques, the database newsletter (1999 revision at http://www.essentialstrategies.com), 23(3)

Hilley, D., et al., The Interpretive Turn: Philosophy, Science, Culture. Ithaca, N.Y.: Cornell University Press, 1991

Jonassen, H. D. (2000). Computers as Mindtools for schools. Engaging critical thinking, 2nd Ed., Merrill: Prentice Hall

Juhn, S., & Naumann, D. (1985). The effectiveness of data representation characteristics on user validation., Proceedings of the Sixth international Conference on information Systems, 212-226

Lochovsky, H., & Tsichritzis, C., (1977). User performance considerations in DBMS selection., Proceedings of ACM SIGMOD, 128-134

Mcintyre, D., Pu, H., & Wolff, F., (1995), Use of software tools in teaching relational database design, Pergamon Computers Education, 24(4), 279-286

Perakath C. B. et all (1994). Information integration for Concurrent Engineering. IDEF5 Method report, Report prepared by KBS inc for Armstrong Laboratory AL/HRGA Wright-Patterson Air fromce Base (http://www.essentialstrategies.com)

George Fessakis holds degree (1994) and an MSc (1996) in Informatics from the Department of Informatics and Telecommunications of National and Kapodistrian University of Athens. He has participated in several research projects about the educational use of Information and Communication Technologies while he is Informatics teacher at a vocational public school named 2nd TEE of Rhodes at Greece and a PhD student at University of Aegean. His research interests include data and procedural modeling, algorithms visualization, and Computer Sciences Education.

Angelique Dimitracopoulou is Associate Professor of the University of the Aegean, School of Human Studies, and member of the "Learning Technology and Educational Engineering" Laboratory. She holds degree in Physics Sciences (Univ Patras, 1986), Master and PhD in Information and Communication Technologies in Education (University of Paris 7 - Denis Diderot, 1995). She has an extensive research and teaching experience related to the design of technology-based learning environments in the field of science and mathematics education, the implementation of learning environments in real education contexts and the research on learning and teaching

using ICTs, in a large variety of education levels (from pre-primary school to adults). She is an expert in the design of various kinds of technology based learning environments (modelling systems, intelligent tutoring systems, collaborative systems and tools, web-based systems).

Vassilis Komis holds a degree in Mathematics from the University of Crete (1987), DEA (1989) and doctoral degrees (1993) in Computer Science Education (Didactique de l'Informatique) from the University of Paris 7 - Denis Diderot . He is currently an Assistant Professor in the Department of Early Childhood Education of the University of Patras. His publications and research interests concern the computer science Education, the pupils' representations of the Information and Communications Technologies (ICT) as well as their use in classroom, the integration of ICT in education, the design and the development of educational software.