Identifying the Weaknesses of UML Class Diagrams during Data Model Comprehension

Gabriele Bavota[1], Carmine Gravino[1], Rocco Oliveto[2], Andrea De Lucia[1], Genoveffa Tortora[1], Marcela Genero[3], and José A. Cruz-Lemus[3]

[1] Software Engineering Lab, University of Salerno, Fisciano (SA), Italy
{gbavota, gravino, adelucia, tortora}@unisa.it
[2] STAT Departement, University of Molise, Pesche (IS), Italy
rocco.oliveto@unimol.it
[3] Dep. of Technologies and Information Systems, University of Castilla, La Mancha
{marcela.genero, joseantonio.cruz}@uclm.es

Abstract. In this paper we present an experiment and two replications aimed at comparing the support provided by ER and UML class diagrams during comprehension activities by focusing on the single building blocks of the two notations. This kind of analysis can be used to identify weakness in a notation and/or justify the need of preferring ER or UML for data modeling. The results reveal that UML class diagrams are generally more comprehensible than ER diagrams, even if the former has some weaknesses related to three building blocks, i.e., multi-value attribute, composite attribute, and weak entity. These findings suggest that a UML class diagram extension should be considered to overcome these weaknesses and improve the comprehensibility of the notation.

1 Introduction

A data model is a set of concepts that can be used to describe both the structure of and the operations on a database [1]. It represents the output of data modeling (or conceptual design), an activity that aims at creating a conceptual schema in a diagrammatic form and facilitating the communication between developers and users [1]. Understanding and interpreting data models represents a fundamental activity from the earliest stages of software development, e.g., requirement analysis. Thus, a comprehensive notation is really desirable to avoid misunderstanding that can lead to the introduction of errors very expensive to remove in the later phases of the software development. A comprehensive notation is also desirable during software maintenance, since it facilitates the comprehension activities that have to be performed to understand the design of the system before the analysis and the implementation of a change request.

Entity-Relationship (ER) and its extensions are the most used notations for database conceptual modeling and still remains the *de facto* standard [1]. The success of the Object-Oriented (OO) approach for software development has encouraged the use of this approach also for database modeling [2]. In particular, UML class diagrams can be used to represent the conceptual schema of the whole software system, so the same notation can be used to model the functionality of the system as well as to represent its data. The structural constructs of the UML class diagram which represents the data structure is somewhat equivalent to Extended ER (EER) representation (e.g., object classes considered equivalent to entity and relationship types). The functionality is represented through "methods" that are attached to the object classes. However, while UML is becoming a *de facto* standard for the analysis and design of software systems, it is not exploited with the same success for modeling databases. Indeed, nowadays ER remains the most used notation to model databases and in some cases it complements UML in the design of software systems. A recent survey also indicated that in some cases both ER and UML class diagrams are employed to represent the same database [3]. Such behaviors might be the trigger for possible problems during the evolution of the data models. More effort is required to maintain the models and their implementation up-to-date, since out-of-data models can generate inconsistency and misunderstanding during software maintenance and evolution. All these considerations lead researchers to empirically compare the ER and UML diagrams to show the actual benefits given by one notation as compared to the other [3, 4]. The results achieved in all these studies indicate that the support given by UML class diagrams in comprehension tasks is at least equal (and in some cases higher than) the support given by ER diagrams. However, a qualitative and quantitative analysis concerning the identification of the graphical elements of one notation that are more comprehensible than the corresponding element in the other notation is still missing (this kind of analysis is quantitatively performed in [2] during the comparison of EER and OO models). Such an analysis is vital to provide insight on why UML class diagrams are better than ER diagrams or *vice versa* and highlight strengths and limitations of the two notations. This kind of analysis can be used to (i) justify the need of preferring ER or UML class diagrams for data modeling; or (ii) identify weakness in a notation that could be overcome to improve its comprehensibility.

In this paper we aim at bridging this gap presenting the results of a controlled experiment and two replications to deeply analyze the support given by ER and UML class diagrams during the comprehension of data models. The experiments aimed at performing a fine-grained analysis to (quantitatively and qualitatively) compare the single building blocks, i.e., Entity, Primary Key/ID, Composite Attribute, Multi-value Attribute, Recursive relationship, Relationship cardinality, Ternary relationship, Generalization IS-A, Weak entity, M:N relationship, of the two notations. The experiment and its replications involved 156 students of the university of Salerno (Italy) with different academic background represented by fresher, bachelor, and master students.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 provides details of the design of the experiment and presents the results achieved while Section 4 discuss the possible threats to validity. Concluding remarks and directions for future work are given in Section 5.

2 Related work

In the last two decades some papers have analysed, through controlled experiments, empirical studies, or surveys, graphical notations supporting the software development process.

To the best of our knowledge only four papers compare the ER notation, or its extensions, and Objected-Oriented (OO) models [5], [2], [6], [7]. In particular, Shoval and Shiran [5] compare Extended ER (EER) and OO data models from the point of view of design quality, where quality is measured in terms of correctness of the produced models, time to completely perform the design task, and designers' opinions. The goal of our empirical investigation is different, since we compare ER and UML diagrams from a maintainer perspective in order to verify whether the use of UML diagrams provides better supports during comprehension activities on data models. The comparison performed by Shoval and Shiran reveals that there are no significant differences between Extended ER (EER) and OO data models, except for the use of ternary and unary relationships since in this case EER models provide better results. Furthermore, the designers preferred to work with the EER models.

Shoval and Frumermann [2] also perform a comparison of EER and OO diagrams taking into account the user comprehension. As done by Shoval and Shiran [5], they separately examine the comprehension of various constructs of the analysed models. Their analysis reveals that EER schemas are more comprehensible for ternary relationships while for the other constructs no significant difference is found.

Bock and Ryan [6] also examine the correctness of the design for several constructs of the considered diagram types in an empirical analysis comparing EER and OO models from a designer perspective. The analysis reveals significant difference only in four cases (i.e., representation of attribute identifiers, unary 1:1 and binary m:n relationships) and no difference is found concerning the time to complete the tasks.

A comparison between OO and ER models from an end-user perspective is also carried out by Palvia *et al.* [7], whose aim is to establish which is more comprehensible. Differently from previous reported studies, they measure comprehension on overall terms, not considering specific constructs, and the results of their investigation suggested that OO schemas are superior in this respect.

3 Empirical evaluation

This section describes in detail the design of the controlled experiment we performed and the analysis and interpretation of the achieved results. A discussion of the threats to validity is also presented at the end of the section.

3.1 Goal, Definition, and Context

The *goal* of our experimentation was to analyse whether UML class diagrams are more comprehensible than ER diagrams during the comprehension of data models. Moreover, we are interested in performing a fine grained analysis to compare the single building blocks B_i of the two notations to identify possible weaknesses of the UML class diagrams with respect to the ER diagrams, where $B_i \in$ { *Entity, Primary Key/ID, Composite Attribute, Multi-value Attribute, Recursive relationship, Relationship cardinality, Ternary relationship, Generalization IS-A, Weak entity, M:N relationship*}.

The performed experiments involved students of the University of Salerno (Italy) having different academic backgrounds and, consequently, different levels of experience on ER and UML diagrams:

- *fresher students*, i.e., 1st year B.Sc. students that were starting their academic career when the experiment was performed;
- *bachelor students*, i.e., 2nd year B.Sc. students that attended Programming and Databases courses in the past and were attending the Software Engineering course when the experimentation was performed;
- *master students*, i.e., 1st year M.Sc. students that attended advanced courses of Programming and Software Engineering in the past and were attending an advanced Databases course when the experimentation was performed;

Note that in the Software Engineering course the design notation used is UML while for the Databases course the design notation is ER. The number of subjects involved in the original experiment were 37 bachelor students, while the first and second replications involved 52 master students and 67 fresher students subjects, respectively. We employed the data models of the following systems:

- Company, a software system implementing all the operations required to manage the projects conducted by a company;
- EasyClinic, a software system implementing all the operations required to manage a medical doctor's office.

In particular, we exploited two different data models represented in terms of ER and UML class diagrams. Table 1 shows the characteristics of the data models we employed in the experiments. The selection of the objects for each experiment was performed ensuring that the data models had a comparable level of complexity. For this reason, we extracted sub-diagrams of comparable size from the original data models according to the *"the rule of seven"* given by Miller [8] to build comprehensible graphical diagrams[4]. In the context of our experimentation we applied such a rule to select data models easy to comprehend. This was necessary because (i) each experiment was designed to be performed in a limited amount of time and (ii) a simple data model is preferred to a more complex data model since the latter might influence the comprehension activities.

3.2 Design

Each experiment was organised in two laboratory sessions. In particular, in the context of the experiment subjects had to perform two comprehension activities

[4] The rule of seven is the generally accepted claim that people can hold approximately seven chunks or units of information in their short-term memory at a time [8].

Table 1. Data models used in each controlled experiment

System	# entities	# attributes	# relationships
Company	7	17	5
EasyClinic-BookingManagement	6	18	5

Table 2. Experimental design

Group	**Treatment**	
	ER	**UML**
A	EasyClinic, Lab1	Company, Lab2
B	Company, Lab2	EasyClinic, Lab1
C	Company, Lab1	EasyClinic, Lab2
D	EasyClinic, Lab2	Company, Lab1

on the data models of two different software systems. Each subject analysed the UML diagram (or ER diagram) of one system in one laboratory session and the ER diagram (or UML diagram) of the other system in the other laboratory session. The organisation of each group of subjects[5] in each experimental lab session (*Lab1* and *Lab2*) followed the design shown in Table 2. In particular, the rows represent the four experimental groups, whereas the columns refers to the design notation used to represent the data model (i.e., ER and CD).

3.3 Comprehension Questionnaires

The main outcome observed in the three experiments was the comprehension level. To evaluate it, we asked the subjects to answer a questionnaire (similar to [9]) consisting of 10 multiple choice questions where each question has one or more correct answers. The number of answers is the same for each question (i.e., three answers), while the number of correct answers is different. The questions cover all the building blocks B_i of the two notations exploited to model a database. Figure 1 shows a sample question of the comprehension questionnaire regarding the system Company.

The same building blocks were qualitatively analysed through a questionnaire where subjects specified their preferences between the two considered notations. In particular, for each building block B_i they manifested a preference between ER diagram, No preference, and UML class diagram.

Moreover, at the end of each laboratory session a survey questionnaire was proposed to the subjects. This survey aimed at assessing the overall quality of the provided material as well as the clearness and difficulty of the comprehension tasks. In particular, the subjects provided answers to the following questions (one choice for each question):

S1 : I had enough time to perform the tasks
S2 : The task objectives were perfectly clear to me
S3 : The tasks I performed was perfectly clear to me
S4 : Judging the difficulty of the comprehension task

[5] The students were assigned to the four groups in a randomly balanced way.

where S1, S2, and S3 expected closed answers according to the Likert scale [10] from 1 (strongly disagree) to 5 (strongly agree), while S4 from 1 (very low) to 5 (very high).

3.4 Variable selection

We performed a single factor within-subjects design, where the independent variable (main factor) is represented by the design notation used to represent a data model. This variable is denoted as **Method**, that can be ER diagram (*ER*) or UML class diagram (*CD*).

The dependent variable is **comprehension level**, which denotes the comprehension level achieved by the subjects using the two notations. To measure it we use two well known Information Retrieval metrics, namely recall and precision [11]. Indeed, since the questionnaire is composed of multiple-choice questions, we define recall and precision as follow:

$$recall_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |correct_i|}\% \quad precision_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |answer_{s,i}|}\%$$

where $answer_{s,i}$ is the set of answers given by the subjects s to the question i and $correct_i$ is the set of correct answers expected for the question i. Note that the measures defined above represent aggregations of the precision and recall values that have been obtained considering each question of the questionnaire. Differently from aggregate measures based on the mean of precision and recall values the adopted measures also consider the fact that subjects do not provide any answer for a given question [12].

Finally, it is worth noting that recall and precision measure two different concepts. Thus, we decided to use their harmonic mean (i.e., F-measure [11]) to obtain a balance between them and compute the comprehension level.

However, to better assess the effect of **Method** it was necessary to control other factors (called co-factors) that may impact the results achieved by the subjects and be confounded with the effect of the main factor. In the context of our study, we identify the following co-factors:

– **ER and UML experience**: fresher students did not know the ER and UML diagrams, while bachelor and master students had a fairly good knowledge

Q4 Let us focus on the classes Project and Company.
Which of the following statements is true:
[] A company has a unique office
[] A project has a unique office
[] A company may have multiple offices

Fig. 1. A question example

of these notations and master students were more trained than bachelor students on the design methods. We were also interested in analysing the effect of the ER and UML experience since the different levels of education (and, consequently, the different levels of UML and ER experience) may impact the results achieved by subjects.
- **System**: even if we tried to select two software systems of a comparable size and tried to balance the complexity of the data models by using as heuristic the Miller's rule, there is still the risk that the system complexity may have a confounding effect with **Method**. For this reason we also considered the modeled system as an experimental co-factor.
- **Lab**: the experiments were organised in two laboratory sessions. In the first session subjects performed the task using UML class diagrams (or ER diagrams) and in the other session they performed the task using ER diagrams (or UML class diagrams). Although the experimental design limits the learning effect, it is still important to analyse whether subjects perform differently across subsequent lab sessions.

3.5 Procedure and data analysis

Subjects performed the assigned tasks individually. Before the experiments, subjects were trained on both ER and UML class diagrams. To avoid bias (i) the training was performed on a data model not related to the systems selected for the experimentation and (ii) its duration was exactly the same for the experiment and the replications. Right before the experiments, the students attended a 30 minutes presentation where detailed instructions concerning the tasks to be performed were illustrated. The design, the material[6] and the procedure were exactly the same for the experiment and its replications. Subjects represented the only substantial difference among the experiment and the two replications.

Since in our experiments each subject performed a task on two different models (i.e., *Company*, or *EasyClinic*) with the two possible treatments (i.e., ER, and CD), it was possible to use a paired Wilcoxon one-tailed test [14] to analyse the differences exhibited by each subject for the two treatments. A one-tailed paired t-test [14] can be used as alternative to the Wilcoxon test. However, we decided to use the Wilcoxon test since it is resilient to strong departures from the t-test assumptions [15]. The achieved results were intended as statistically significant at $\alpha = 0.05$. This means that if the derived p-value is less than 0.05, it can be concluded that there is significant difference between the support given by the treatments when performing comprehension tasks on data models. Furthermore, we analysed the students preferences about the single building blocks of the two notations using histograms, while the answers provided by subjects to the survey questionnaire were analysed using boxplots. The chosen design also permitted to analyse the effects of co-factors and their interaction with the main factor. To this aim we used the two-way Analysis of Variance (ANOVA) [14].

[6] See [13] for the complete material used in the experiments.

Table 3. Descriptive statistics of comprehension by method and subjects group

Subjects	ER			CD		
	Mean	Median	St. Dev.	Mean	Median	St. Dev.
Fresher	0.801	1.000	0.307	0.816	1.000	0.280
Bachelor	0.849	1.000	0.242	0.845	1.000	0.278
Master	0.849	1.000	0.277	0.838	1.000	0.272

Table 4. Wilcoxon Test results of comprehension by method and subjects group

Subjects	CD$FM - ER$FM			p-value	effect size
	Mean	Median	St. Dev.		
Fresher	0.014	0.000	0.404	0.343	0.037
Bachelor	0.003	0.000	0.330	0.420	-0.011
Master	-0.012	0.000	0.383	0.817	-0.030

3.6 Analysis and Interpretation of the Results

Table 3 reports the descriptive statistics of the F-measure, i.e., comprehension level, achieved by the subjects in our experimentation. The results highlighted that the two notations provided comparable support when performing comprehension activities on data models. In particular, the higher difference between the two notations in terms of F-measure is just 1% (see Table 3). As designed, to analyse if the difference between the results obtained using the two notations is statistically significant, we performed the Wilcoxon test. Table 4 reports the achieved results that highlight no significant difference between the two notations when used to comprehend data models (p-value always higher than 0.05).

Our finding contrasts with the results achieved in [4] where the authors demonstrated the benefits provided by the UML class diagrams with respect to the ER diagrams during the comprehension of data models. To further investigate this discrepancy, we analysed the support given by the two notations at a fine-grained level, i.e., on each building block used in the definitions of data models. Table 5 reports the descriptive statistics of the results achieved in terms of F-measure (considering the subjects answers to questions related to each building block). The achieved results confirmed an overall "performance equilibrium" between the two notations. In particular, there are some building blocks that represent strengths of CD, e.g., Entity and Ternary Relationship, as well as building blocks that represent weaknesses of CD, e.g., Composite and Multi-value attributes. In order to statistically analyse the weaknesses of CD, Table 6 shows the results of the Wilcoxon test executed for each building block to verify where the ER performances are statistically better than those of CD. The achieved results revealed that ER has a comprehension level significantly higher than the comprehension level of CD for three building blocks, i.e., Composite attribute, Multi-value attribute, and Weak entity. These results held for all the subjects involved in the experimentation. The only exception is given by Bachelor students when analysing the Multi-value attribute building block. However, Table 5 shows that Bachelor students also achieved better results in terms of descriptive statistics with ER when answering the questions related

Table 5. Descriptive statistics of the results (F-measure).

Method	Element	Fresher			Bachelor			Master		
		Mean	Median	St. Dev.	Mean	Median	St. Dev.	Mean	Median	St. Dev.
ER	Entity	0.887	1.000	0.260	0.936	1.000	0.125	0.872	1.000	0.281
	Primary Key/ID	0.784	1.000	0.406	0.955	1.000	0.179	0.907	1.000	0.277
	Composite attribute	0.883	1.000	0.159	0.897	1.000	0.146	0.920	1.000	0.140
	Multi-value attribute	0.859	1.000	0.195	0.847	1.000	0.168	0.862	1.000	0.213
	Recursive relationship	0.779	1.000	0.301	0.757	0.667	0.224	0.817	1.000	0.243
	Relationship cardinality	0.875	1.000	0.240	0.892	1.000	0.158	0.929	1.000	0.179
	Ternary relationship	0.741	1.000	0.347	0.828	1.000	0.220	0.804	1.000	0.321
	Generalization IS-A	0.684	0.667	0.369	0.734	1.000	0.363	0.712	1.000	0.379
	Weak entity	0.725	0.800	0.266	0.767	1.000	0.305	0.747	0.900	0.329
	M:N relationship	0.789	1.000	0.368	0.865	1.000	0.319	0.923	1.000	0.244
CD	Entity	0.961	1.000	0.108	0.937	1.000	0.234	0.926	1.000	0.145
	Primary Key/ID	0.875	1.000	0.296	0.937	1.000	0.234	0.926	1.000	0.246
	Composite attribute	0.742	0.667	0.255	0.781	0.800	0.251	0.815	1.000	0.308
	Multi-value attribute	0.775	0.667	0.259	0.788	0.667	0.257	0.801	0.667	0.209
	Recursive relationship	0.767	1.000	0.323	0.856	1.000	0.226	0.806	0.800	0.210
	Relationship cardinality	0.865	1.000	0.261	0.856	1.000	0.320	0.906	1.000	0.150
	Ternary relationship	0.827	1.000	0.265	0.888	1.000	0.150	0.855	1.000	0.162
	Generalization IS-A	0.828	1.000	0.225	0.838	1.000	0.290	0.804	1.000	0.328
	Weak entity	0.629	0.667	0.407	0.611	0.667	0.407	0.608	0.733	0.447
	M:N relationship	0.890	1.000	0.162	0.955	1.000	0.179	0.929	1.000	0.212

Table 6. Wilcoxon Test by Questions

Element	Fresher					Bachelor					Master				
	ER$FM - CD$FM			p-value	effect size	ER$FM - CD$FM			p-value	effect size	ER$FM - CD$FM			p-value	effect size
	Mean	Median	St. Dev.			Mean	Median	St. Dev.			Mean	Median	St. Dev.		
Entity	-0.059	0.000	0.262	0.983	-0.257	-0.036	0.000	0.153	0.599	-0.032	-0.054	0.000	0.309	0.796	-0.161
Primary Key/ID	-0.091	0.000	0.517	0.927	-0.166	-0.027	0.000	0.198	0.415	0.059	-0.019	0.000	0.388	0.660	-0.049
Composite attribute	0.141	0.000	0.303	**0.000**	0.490	0.116	0.000	0.306	**0.022**	0.380	0.105	0.000	0.304	**0.012**	0.343
Multi-value attribute	0.085	0.000	0.316	**0.014**	0.269	0.059	0.000	0.324	0.141	0.180	0.061	0.000	0.311	**0.080**	0.196
Recursive relationship	0.012	0.000	0.401	0.455	0.024	-0.010	0.000	0.287	0.983	-0.345	0.011	0.000	0.308	0.536	0.037
Relationship cardinality	0.009	0.000	0.358	0.439	0.028	-0.009	0.000	0.200	0.446	0.094	0.023	0.000	0.224	0.258	0.103
Ternary relationship	-0.086	0.000	0.471	0.897	-0.184	-0.042	0.000	0.266	0.869	-0.221	-0.050	0.000	0.368	0.720	-0.135
Generalization IS-A	-0.145	0.000	0.421	0.999	-0.388	-0.104	0.000	0.476	0.905	-0.217	-0.093	0.000	0.526	0.903	-0.177
Weak entity	0.096	0.000	0.457	**0.027**	0.211	0.156	0.000	0.504	**0.045**	0.309	0.139	0.000	0.590	**0.049**	0.234
M:N relationship	-0.105	0.000	0.379	0.972	-0.249	-0.045	0.000	0.334	0.942	-0.252	-0.006	0.000	0.313	0.562	-0.020

bold if ER comprehension level statistically higher than CD comprehension level

to the Multi-value attribute. It is worth noting that the controlled experiments and replications reported in [4] did not consider these three building blocks to determine comprehension level provided by the two notations, i.e., the questionnaires used by the authors did not include questions related to Composite attribute, Multi-value attribute, and Weak entity. To verify whether the different findings between our experimentation and the results achieved in [4] was due to these three building blocks we also performed the comparison between ER and UML class diagrams without considering the answers of the students related to Composite attribute, Multi-value attribute, and Weak entity. In particular, we re-executed the Wilcoxon test to analyse if CD provided a significant higher comprehension level than ER. The results in Table 7 highlight that CD achieved statistically significant higher comprehension level than ER for the Fresher and Bachelor students. Moreover, CD provided better results than ER also for Master students even if this is not statistically significant (p-value 0.096).

Besides a quantitative analysis, we also conducted a qualitative comparison of the support given by the building blocks of the two notations. Figures 2, 3, and 4 report the preferences expressed by the Fresher, Bachelor, and Master stu-

Table 7. Wilcoxon Test results of comprehension support by method and subjects' group without the identified weaknesses

Subjects	CD$FM - ER$FM			p-value	effect size
	Mean	Median	St. Dev.		
Fresher	0.066	0.000	0.410	**0.000**	0.161
Bachelor	0.052	0.000	0.290	**0.010**	0.120
Master	0.027	0.000	0.358	0.096	0.074

bold if CD comprehension level statistically higher than ER comprehension level

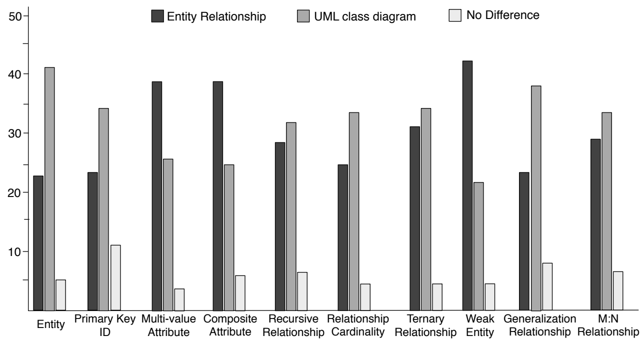

Fig. 2. Subject's preferences - Fresher

dents, respectively. It is worth noting that the results of the quantitative analysis are confirmed by the preferences expressed by the students. In particular, the students preferred ER diagrams to represent the three building blocks identified as weaknesses of the UML class diagrams during the quantitative analysis, i.e., Multi-value attribute, Composite attribute, and Weak entity. Concerning the remaining building blocks, the students preferred UML class diagrams to represent the Entity, the Relationship cardinality, and the Generalization relationship, while they did not provide a clear preference for the Primary key/ID, Recursive relationship, Ternary relationship, and M:N relationship.

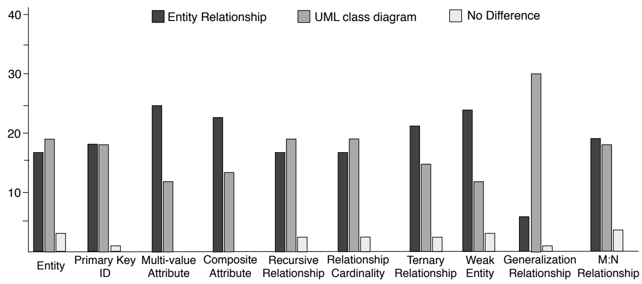

Fig. 3. Subject's preferences - Bachelor

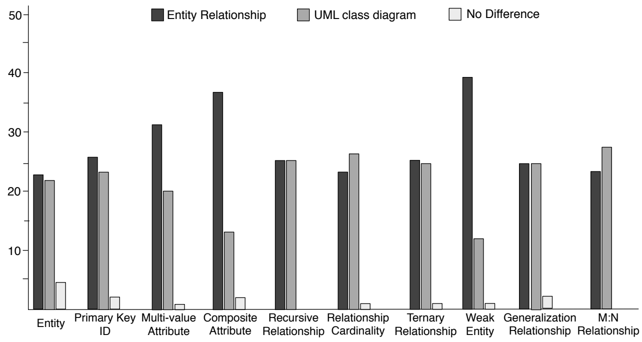

Fig. 4. Subject's preferences - Master

4 Discussion and Threats to Validity

Summarising, the achieved (quantitative and qualitative) results highlighted that the UML notation is characterized by three weaknesses related to the representation of Composite attribute, Multi-value attribute, and Weak entity, with respect to the ER notation, when performing comprehension activity on data models. However, except for the three identified weaknesses, the UML notation is generally more comprehensible than the ER notation, confirming the findings of previous experiments [4]. These findings suggest that a UML class diagram extension focused on these three building blocks should be considered to overcome these weaknesses and improve the comprehensibility of data models given in terms of UML notation. All these findings could be affected by many threats to validity [16] discussed in the following.

Goal, Design, and Statistical analysis. Ease of comprehension was the only criterion examined, because comprehension is a key issue for a graphical notation. However, especially where the design of performance-critical, data-intensive software like databases is concerned, there are other key considerations as well, e.g., analysability. One may choose to sacrifice expressiveness for analysability or other properties. For this reason, future work will be devoted to evaluate other properties of the two notations.

As explained in Section 3 we captured the students' opinion about the quality of the provided material, the clearness of the comprehension tasks and the laboratory goals, and the difficulty in performing the comprehension tasks, to verify if the results of our experimentation could be influenced by these threats. Figure 5 shows boxplots of answers for (a) fresher, (b) bachelor, and (c) master students. The analysis suggested that students had enough time to carry out the tasks (S1) and the objectives and the tasks to perform were clear (S2 and S3), since the median of boxplots of answers was 4 (i.e., I agree). Furthermore, they experienced no particular difficulties when performing the comprehension tasks (S4) since the median of the answers was 3.

Table 8. ANOVA: analysis of the Lab and System co-factors.

Factor	Fresher	Bachelor	Master	All
Lab	No (0.787)	No (0.163)	No (0.175)	No (0.216)
System	No (0.793)	No (0.636)	No (0.113)	No (0.229)
Method *vs* Lab	No (0.817)	No (0.833)	No (0.305)	No (0.439)
Method *vs* System	No (0.793)	No (0.817)	No (0.618)	No (0.679)

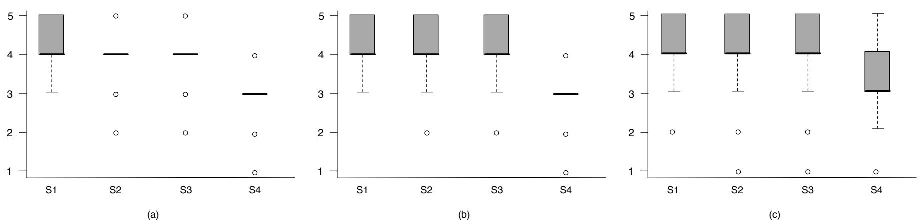

Fig. 5. Answers of subejcts to survey questionnaire

The metric used to assess the subjects' performance (comprehension) is an aggregate measure of precision and recall that well reflects the results achieved by the subjects. We are also confident that the used tool (multiple-choice questions) actually measures the comprehensibility of the data models. This is also confirmed by the fact that previous empirical studies also used similar approaches to measure the same attributes (see for instance [2], [5], [6], [7], [9]).

Even if the chosen design mitigates the learning (or tiring) effect, there is still the risk that, during labs, subjects might have learned how to improve their comprehension performances. We tried to limit this effect by means of a preliminary training phase. In addition, as highlighted in [15], one possible issue related to the chosen experiment design concerns the possible information exchange among the subjects between the laboratories. To mitigate such a threat the experimenters monitored all the students during the experiment execution to avoid collaboration and communication between them. Finally, subjects worked on three different diagrams and, even if we tried to select diagrams having comparable size, there is still the risk that one diagram might be easier than another.

All these considerations suggest to account **Lab** and **System** as co-factors in the analysis of results. Indeed, the chosen design permitted to analyse the effect of co-factors and their interaction with the main factor. Table 8 shows the results of the ANOVA test by **Method** and **Lab**. The analysis did not reveal any significant influence of the two co-factors nor any significant interaction between the main factor and the two co-factors.

Since the assigned task had to be performed in a limited amount of time, the time pressure could represent another threat to validity. However, we decided the duration of each experiment taking into account previous laboratory exercises performed by the students involved in the experimentations during their courses. Furthermore, we also exploited our experience in performing similar controlled experiments in the past [4]. However, all the subjects completed the assigned task and they declared (in the post-experiment questionnaire) that the available time was enough to complete the task. For these reasons we are confident that time pressure did not condition the results and thus we did not consider it as a confounding factor.

Proper tests were performed to statistically analyse the difference in the performance achieved employing the two experimented notations, i.e., ER and UML class diagram. Survey questionnaires, mainly intended to get qualitative insights, were designed using standard ways and scales [10] allowing us to use statistical analysis to analyse differences in the feedback provided by subjects.

Subjects and objects. The three controlled experiments involved students having different backgrounds, i.e., fresher, bachelor, and master students. Concerning the undergraduate and graduate students, they had an acceptable analysis, development, and programming experience. In particular, in the context of the Software Engineering courses, both master and bachelor students had participated to software projects, where they experienced software development and documentation production, including database design documents. Moreover, as highlighted by Arisholm and Sjoberg [17] the difference between students and professionals is not always easy to identify. Nevertheless, there are several differences between industrial and academic contexts. For these reasons, we plan to replicate the experiment with industrial subjects to corroborate our findings. We also plan in the future to conduct a survey involving people from database and software engineering communities aiming at obtaining opinions on why weak entity, multi-value and composite attributes are (or might be) problematic in the UML notation. In this way, we can perform a more notation-oriented discussion about the identified weaknesses.

The different backgrounds of the students involved in the experiments have been accounted as a co-factor to analyse its influence on and interaction with the main factor. As expected the ANOVA test revealed a statistically significant effect of **ER and UML Experience** (p-value < 0.001); bachelor and master students achieved statistically significant better performances than fresher students, while the performances achieved by bachelor and master are almost comparable. In addition, ANOVA did not reveal any interaction between **ER and UML Experience** and the main factor (p-value $= 0.486$).

To avoid social threats due to evaluation apprehension, students were not evaluated on the performances they achieved in the experiments. During the experiment, we monitored the subjects to verify whether they were motivated and paid attention in performing the assigned task. We observed that students performed the required task with dedication and there was no abandonment. Moreover, students were aware that our goal was to evaluate the impact of using ER or UML class diagrams during modelling activities, but they were not aware of the exact hypotheses tested and of the considered dependent variables.

Finally, the size of the data models is small compared to industrial cases, but it is comparable with the size of models used in other related experimentations (see, for instance, [5], [15], [9]). Future work will be devoted to assess the usefulness of the notations on realistically sized artefacts. However, we believe that the comparison of the two notations on small/medium artefacts is still a worthy contribution.

5 Conclusion and Future Work

We have reported on the results of a controlled experiment and two replications aimed at analysing the support given by ER and UML class diagrams during the coprehension of data models. We have also performed a fine-grained analysis to compare the single building blocks of the two notations (e.g., entity, relationships). The results of the empirical analysis have suggested that UML class diagrams are generally more comprehensible than ER diagrams, confirming the results achieved in a previous study [4]. However, the fine-grained analysis has revealed some weakness of UML class diagrams with respect to ER diagrams. In particular, if we take into account the results about the weak entity, multivalue and composite attributes building blocks, the performances achieved with ER diagrams are superior than those obtained with UML class diagrams. Moreover, the performed qualitative analysis has also highlighted that the subjects preferred ER diagrams for specifying weak entities, multivalue and composite attributes. Taking into account these results, in the future we intend to exploit stereotypes, as done in other studies [9], [18], [19], [20], to extend the UML class digrams and bridge the gap with ER diagrams about the specification of weak entity, multivalue and composite attributes building blocks. The aim is to improve the comprehensibility of UML class diagrams and candidate such notation as a new de facto standard also for data modeling.

As it always happens with empirical studies, replications in different contexts, with different subjects and objects, is the only way to corroborate our findings. It would be interesting to consider alternative experimental settings in several respects, but maybe the most important one is the profile of the involved subjects. Replicating this study with students/professionals having a different background would be extremely important to understand how UML class diagrams influence the results of these different sub-populations.

References

1. S. B. Navathe, "Evolution of data modeling for databases," *Commun. ACM*, vol. 35, pp. 112–123, September 1992.
2. P. Shoval and I. Frumermann, "OO and EER conceptual schemas: A comparison of user comprehension," *Journal of Database Management*, vol. 5, no. 4, pp. 28–38, 1994.
3. A. De Lucia, C. Gravino, R. Oliveto, and G. Tortora, "Assessing the support of ER and UML class diagrams during maintenance activities on data models," in *12th European Conference on Software Maintenance and Reengineering, 2008. CSMR 2008.*, april 2008, pp. 173 –182.
4. A. D. Lucia, C. Gravino, R. Oliveto, and G. Tortora, "An experimental comparison of ER and UML class diagrams for data modelling," *Empirical Software Engineering*, vol. 15, no. 5, pp. 455–492, 2010.

5. P. Shoval and S. Shiran, "Entity-relationship and object-oriented data modeling: an experimental comparison of design quality," *Data Knowledge Engineering*, vol. 21, pp. 297–315, February 1997.
6. D. Bock and T. Ryan, "Accuracy in modeling with extended entity relationship and object oriented data models," *Journal of Database Management*, vol. 4, pp. 30–39, 1993.
7. P. Palvia, C. Lio, and P. To, "The impact of conceptual data models on end-user performance," *Journal of Database Management*, vol. 3, no. 4, pp. 4–15, 1992.
8. G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *The Psychological Review*, vol. 63, pp. 81–97, 1956.
9. F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, and M. Ceccato, "How developers' experience and ability influence web application comprehension tasks supported by UML stereotypes: A series of four experiments," *IEEE Transactions on Software Engineering*, vol. 36, no. 1, pp. 96 –118, jan.-feb. 2010.
10. A. N. Oppenheim, *Questionnaire Design, Interviewing and Attitude Measurement.* Pinter Publishers, 1992.
11. R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.
12. G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, and E. Merlo, "Recovering traceability links between code and documentation," *IEEE Transactions on Software Engineering*, vol. 28, no. 10, pp. 970 – 983, oct 2002.
13. G. B. *et al.*, "UML vs ER - experimental material," 2011. [Online]. Available: http://sesa.dmi.unisa.it/UMLvsER.html
14. W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. Wiley, 1998.
15. L. C. Briand, Y. Labiche, M. D. Penta, and H. D. Yan-Bondoc, "An experimental investigation of formality in UML-based development," *IEEE Transactions on Software Engineering*, vol. 31, pp. 833–849, 2005.
16. C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in Software Engineering - An Introduction.* Kluwer, 2000.
17. E. Arisholm and D. I. K. Sjoberg, "Evaluating the effect of a delegated versus centralized control style on the maintainability of object-oriented software," *IEEE Transactions on Software Engineering*, vol. 30, pp. 521–534, August 2004.
18. J. Cruz-Lemus, M. Genero, M. Manso, and M. Piattini, "Evaluating the effect of composite states on the understandability of UML statechart diagrams," in *Model Driven Engineering Languages and Systems*, ser. Lecture Notes in Computer Science, L. Briand and C. Williams, Eds., vol. 3713. Springer Berlin / Heidelberg, 2005, pp. 113–125.
19. M. Genero, J. A. Cruz-Lemus, D. Caivano, S. Abrahão, E. Insfran, and J. A. Carsí, "Assessing the influence of stereotypes on the comprehension of uml sequence diagrams: A controlled experiment," in *Proceedings of the 11th international conference on Model Driven Engineering Languages and Systems*, ser. MoDELS '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 280–294.
20. M. Staron, L. Kuzniarz, and C. Thurn, "An empirical assessment of using stereotypes to improve reading techniques in software inspections," in *Proceedings of the third workshop on Software quality*, ser. 3-WoSQ. New York, NY, USA: ACM, 2005, pp. 1–7.