ELSEVIER

Information and Software Technology 43 (2001) 949–957

INFORMATION AND SOFTWARE TECHNOLOGY

www.elsevier.com/locate/infsof

Empirical validation of referential integrity metrics

Coral Calero*, Mario Piattini, Marcela Genero

ALARCOS Research Group, University of Castilla-La Mancha, Ronda de Calatrava, 5, 13071 CiudadReal, Spain

Accepted 9 October 2001

Abstract

Databases are the core of Information Systems (IS). It is, therefore, necessary to ensure the quality of the databases in order to ensure the quality of the IS. Metrics are useful mechanisms for controlling database quality. This paper presents two metrics related to referential integrity, number of foreign keys (NFK) and depth of the referential tree (DRT) for controlling the quality of a relational database. However, to ascertain the practical utility of the metrics, experimental validation is necessary. This validation can be carried out through controlled experiments or through case studies. The controlled experiments must also be replicated in order to obtain firm conclusions. With this objective in mind, we have undertaken different empirical work with metrics for relational databases. As a part of this empirical work, we have conducted a case study with some metrics for relational databases and a controlled experiment with two metrics presented in this paper. The detailed experiment described in this paper is a replication of the later one. The experiment was replicated in order to confirm the results obtained from the first experiment.

As a result of all the experimental works, we can conclude that the NFK metric is a good indicator of relational database complexity. However, we cannot draw such firm conclusions regarding the DRT metric.

Keywords: Empirical validation; Referential integrity metrics; Database

1. Introduction

Over recent years, software engineers have proposed large quantities of metrics for software products, processes and resources [1–3]. Metrics are useful mechanisms for improving the quality of software products and also for determining the best ways to help professionals and researchers [4]. Unfortunately, almost all the metrics put forward so far focus on program characteristics (e.g. cyclomatic number [5]) disregarding databases [6]. As far as databases are concerned, metrics have been used for comparing data models rather than schemas. However, metrics for comparing schemas are what are most needed for practical purposes.

Relational databases are the most common databases not only in large industries but also in medium and small ones. In fact, they are the most used and widespread at present [7]. However, no metrics have been proposed for this kind of databases and most of the design decisions are taken based only on subjective experience of the databases designers. The only objective indicator that is commonly used is the normalization theory, upon which Gray et al. [8] proposed the normalization ratio. Although this ratio was proposed for conceptual models, it can also be applied and used in the logical (relational) model.

Nevertheless, we do not think that the normalization theory is sufficient for obtaining good relational database designs. Additional objective indicators must be used to assure their quality. These indicators (the metrics) can provide really useful information to the designers allowing them to take the best design decisions.

Hence, our objective is to provide metrics for controlling relational database quality. However, quality is a multidimensional concept and we decided to work on maintainability. This decision was taken not only because it is one of the factors that affect quality [9], but also because we believe it to be the most important factor, as maintenance incurs between 60 and 80% of the costs of the life cycle. Maintainability is achieved, among other factors, through analyzability [9], which is an external attribute. The most important internal attribute which influences analyzability is complexity [10] and we, therefore, define metrics for complexity, in order to assure relational database analyzability that affects maintainability as a quality factor.

In this paper, we present two metrics for relational databases, which can characterize their complexity (internal attribute) and in doing so help to assess relational database maintainability (external attribute). The presented metrics

* Corresponding author.

E-mail addresses: ccalero@inf-cr.uclm.es (C. Calero), mpiattini@inf-cr.uclm.es (M. Piattini), mgenero@inf-cr.uclm.es (M. Genero).

0950-5849/01/$ - see front matter
PII: S0950-5849(01)00202-6

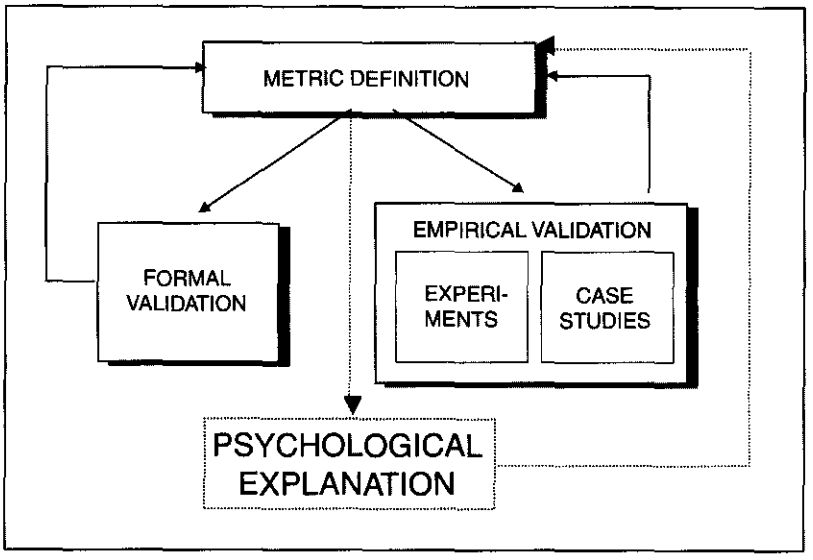


Fig. 1. Steps followed in the definition and validation of database metrics.

are related to referential integrity. We think these metrics are important because they deal with one of the most important concepts of the relational model, which is the referential integrity.

The two metrics presented have been developed in a methodological way that is explained in the following paragraphs. By applying this method, we obtain correct metrics and we know the main characteristics of the metrics defined.

1.1. Method for defining valid database metrics

The method we follow consists of a number of steps which ensure the reliability of the obtained metrics (see Fig. 1).

In Fig. 1, four main activities are represented:

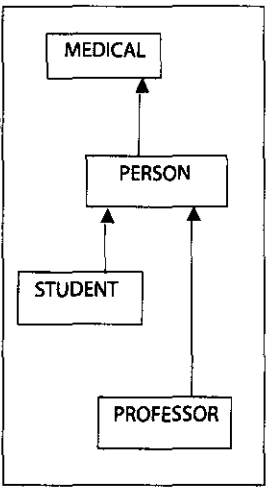
- *Metrics definition.* The definition must be made taking into account the specific characteristics of the database. After consulting relational database experts, we have identified two metrics related to referential integrity:
  Depth of the referential tree (DRT). DRT is defined as the length of the longest referential path in the database schema. To calculate this metric, we can consider the schema database as a graph where tables are the nodes of the graph and arcs represent the referential integrity relationships among tables. Once the graph has been defined, the value of this metric is calculated by following all the possible paths from each table and adding one to the value of the metric for each arc we go through. The maximum value of those obtained for all the possible paths in the graph is the value that the metric takes. The only indication that must be taken into account when the metric is calculated is that cycles are considered only once. In Fig. 2, an example of the calculation of this metric is shown.
  Number of foreign keys (NFK). The NFK metric is defined as the number of foreign keys in the schema. Fig. 2 presents an example of this metric.
- *Formal validation.* The second step is the formal validation of the metrics. Formal validation helps to know when and how to apply metrics. There are two main tendencies in metrics validation: the property-based frameworks [11–13] and those based on measurement theory [3,14]. Formal validation of the presented metrics in the Briand et al. formal framework can be found in [15] and in the Zuse's formal framework ascertained in [16]. As a result of this formal validation, we have obtained that the NFK metric is a complexity metric (using the framework of Briand et al. [12]) and is above the ordinal scale (following the Zuse's [3] formal framework) and that the DRT metric is a length metric (applying the Briand et al. [12] formal framework) and is above the ordinal scale (using the framework of Zuse [3]).
- *Empirical validation.* Here, the objective is to prove the practical utility of the proposed metrics. Basically, we can divide the empirical validation into two parts: experimentation and case studies. Replication of the experiments is also necessary because with the isolated results of one experiment only, it is difficult to appreciate how widely applicable the results are and, thus, to assess to what extent they really contribute to the field [17]. In this paper, the complete replica of an experiment conducted with the two metrics presented is explained in detail.
- *Psychological explanation.* Ideally, we should also be able to explain the influence of the values of the metrics from a psychological point of view. Some authors, such as Siau [18], propose the use of cognitive psychology as a reference discipline in method engineering and the study of information modeling. In this way, cognitive psychology theories such as the Adaptive Control of Thought

```
CREATE TABLE PERSON(
ID_NUMBER NUMBER,
NAME VARCHAR(10),
SURNAME VARCHAR(30),
ADDRESSVARCHAR(20),
MED_NR NUMBER,
PRIMARY KEY(ID)
FOREIGN KEY(MED_NR) REFERENCES MEDICAL);

CREATE TABLE STUDENT(
ID_STU NUMBER,
COURSE CHAR,
START_DATE DATE
ID_PER NUMBER,
PRIMARY KEY(ID_STU),
FOREIGN KEY (ID_PER) REFERENCES PERSON);

CREATE TABLE PROFESSOR(
ID_PROF NUMBER,
NR_OFFICE NUMBER,
TEL VACHAR(13),
E-MAIL VARCHAR(20),
ID_PER NUMBER,
PRIMARY KEY (ID_PROF),
FOREIGN KEY (ID_PER) REFERENCES PERSON);

CREATE TABLE MEDICAL(
DOCTOR_NR NUMBER,
DOCTOR_NAME VARCHAR(40),
PRIMARY KEY(DOCTOR_NR));
```



NFK = 1 + 1 + 1 = 3
DRT = MAX(2,2) = 2

Fig. 2. Example of the NFK and DRT metrics.

(ACT, [19]) could justify the influence of certain metrics in database comprehension. Knowledge of the limitations of the human information processing capacity could also be helpful to establish metrics threshold values for assuring database quality.

As we can see in Fig. 1, the process of defining and validating database metrics is evolutionary and iterative and as a result of the feedback, these metrics could be redefined or discarded depending on their formal and empirical validation, or psychological explanation.

The following sections will present all the information related to the replica of the controlled experiment. The statement of the problem is given in Section 2, and in Section 3, the experiment planning is detailed. Analysis and interpretation of the results are presented in Section 4, and in Section 5, some conclusions are presented together with the future work. All the experiments presented have been prepared following Ref. [20].

2. Problem statement

The objective of our empirical study is to determine if the two metrics presented can be used as a mechanism for controlling the maintainability of relational databases from a practical point of view (because metrics must be used in the real world and it is there, where they must be useful).

The experimentation presented in this paper is a replica of another controlled experiment carried out with the same objective. We have also conducted a case study with the metrics used in the controlled experiments and others (not related with the referential integrity). However, this paper concentrates only on the replica. In Section 2.1, we will briefly present the results obtained from the original experiment and the case study. At the end of this paper, we will discuss the conclusions we can draw from the experimentation process as a whole.

2.1. Previous work

Prior to this study, we conducted another controlled experiment with the aim of proving if the metrics presented were good indicators of the maintainability of relational databases [15].

As the controlled experiment presented later in this paper is the replica of this earlier one, most of their characteristics are similar. Here, we will only mention some specific characteristics and the results obtained in the original experiment and where necessary, during the detailed presentation of the replica, we will point out specific differences between the original and the replica.

The participants were computer science students at the University of Castilla-La Mancha (Spain), who were enrolled in the final-year. When the experiment was done, they were following a course on databases lasting for two semesters. Sixty students performed the experiment, but

Table 1
Characteristics of the case study databases

Database	Number of tables	Number of attributes	Number of foreign keys
One	912	11,683	3621
Two	7	81	9
Three	8	80	10
Four	6	83	24

only 59 were finally accepted because one of them gave the answers in an incorrect way.

The dependent variable was measured as the number of correct answers obtained in a fixed time.

With the data obtained, we applied the *F*-statistic, concluding that the NFKs in a relational database schema could be a solid indicator of its complexity, that the length of the referential tree could not be relevant and that interaction exists between both metrics.

We also conducted a case study with the relational database metrics. Our study was centered on four different databases from the CIEMAT (Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas), which is one of the Spanish organizations possessing very large ORACLE databases [21].

The databases we worked with were used by this organization and were developed by their own designers. Some of the characteristics of each of these databases are presented in Table 1. The names of the databases have been concealed for reasons of confidentiality.

For each one of the databases, we collected two different kinds of information. Firstly, we needed the characteristics of the schema database (which allowed us to calculate the metrics values). To obtain this information, we required the schema database. However, we had some problems obtaining the schemas and finally they gave us the value of some characteristics that we used to calculate the metrics. It was impossible, for example, to calculate the value of the DRT metric based on the information given and we had to work with the metrics we were able to calculate.

Furthermore, we asked the designers about different aspects of the development process, all of them related to the maintainability aspects. The data resulting from both sources of information were used for obtaining the statistical results of our case study.

Using Spearman's coefficient statistic, we obtained a correlation coefficient of 0.775 between the NFK metric and the analyzability of the database.

2.2. Current work

As previously indicated, the main goal of this paper is to explain the replica of the previous controlled experiment. The hypotheses did not vary in the replication of the experiment. However, we did change the way in which the dependent variable was measured (we wanted to capture the analyzability in another way in order to confirm if the previous results were independent of the way the analyzability was captured) and of the subjects (due to the limitations related to the experiments performed by students). By carrying out this kind of replication in which the same hypothesis is studied, but some details of the experiment are changed, our aim is to make the results of the experiment more reliable.

The goal definition of our experiment can be summarized as follows:

To analyze the metrics for relational databases
for the purpose of evaluating if they can be used as a useful mechanisms
with respect of the relational databases analyzability
from the designer *point of view*
in the context of experienced professionals in relational databases.

3. Experiment planning

Eleven professionals from Cronos S.A., a Spanish consulting software firm, who on average had 3 years experience working on relational databases, carried out the experiment. We tried to define the tests involved in the experiment in such a way that they were representative of real cases.

The use of a simple experiment allows the other experimenters to replicate it as the number of examples is not too large. All the experimental materials (laboratory package) of this experiment can be found in http://alarcos.inf-cr.uclm.es.

3.1. Hypotheses formulation

The hypotheses of our experiment are:

Null hypothesis: Different values of DRT and NFK metrics do not affect the analyzability of the database schema.

Alternative hypothesis H1: The value of the DRT metric affects the analyzability of the database schema.

Alternative hypothesis H2: The value of the NFK metric affects the analyzability of the database schema.

Alternative hypothesis H3: The interaction of the DRT and NFK metrics affects the analyzability of the database schema.

Hypotheses H1 and H2 are stated on the basis that when

Table 2
Crossed design for the experiment

Factor A (DRT)	Factor B (NFK)	
	Low	High
Low	2,5	2,8
High	5,5	5,8

the referential integrity among tables increases (or decreases), the maintainability will be affected in some way.

Hypothesis H3 is stated to determine if there is any kind of interaction between both metrics, based on the fact that they are defined among the referential integrity characteristics of relational databases.

3.2. Variables in the study

Independent variables. The independent variables are the variables for which the effects should be evaluated. In our experiment, these variables correspond with the two metrics under study: NFK and DRT. Each of these metrics (factors) can take two different values (levels): five and eight for the NFK metric and two and five for the DRT metric.

Dependent variables. As we have previously stated, the analyzability of the tests was measured as the time each subject used to perform all the tasks of each experimental test and the accuracy with which they completed the tasks included in each test (delete, update and insert). This decision (as opposed to that of the original experiment where we worked with the correct number of answers) was adopted because we wanted to measure analyzability in a different manner. Taking into account the experience of the subjects in relational database design and the fact that the tasks were not too difficult, we thought that all of them would give the right answers. We were proved right as on correcting the tests, all the subjects were found to have answered correctly and we were therefore able to work with the results of all 11 subjects.

Regarding the time, it is necessary to point out that this time included, the time to analyze the schema and the time to answer the three questions about it.

3.3. Design

Taking the hypotheses into account, the experiment must consider two factors: DRT and NFK, so we have a factorial model, where all the values a factor can take are combined with all the values of the other factor (in our case, both metrics can take two values, two and five for the DRT metric and five and eight for the NFK metric). A crossed design as the one described before produces the matrix shown in Table 2 where each value of the matrix is a pair (DRT, NFK).

Table 3
Some characteristics of the four cases

	Case 1	Case 2	Case 3	Case 4
Number of tables	6	6	6	6
DRT	2	2	5	5
NFK	8	5	8	5
Number of attributes	26	30	29	27

START TIME:

1. What tables and how many rows in each table are affected if we delete in the Table 5 the row with code1=210?

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

2. What tables and how many rows in each table are affected if we updatethe column X of the row with code2=11 in the table 3?

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

3. What tables and how many rows and columns are necessary to add if we want add a new row in the table 4? (Suppose that all the necessary data are news in the database)

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

END TIME:

Fig. 3. Question/answer paper.

3.4. Data used in the study

Four relational databases were used for performing the experiment. The documentation for each design was approximately seven pages long and included, in addition to the database schema, a general description and a requirements document. To make the designs comparable, we tried to make them as similar as possible. In Table 3, we present the metrics values of the four designs and the number of tables and the number of attributes for each design.

For each design, three operations were performed. The subjects had to insert, delete and to update in each one of the four schemas and, after, answer how many columns of each table would be changed (were affected) as a result of each one of these operations. With the results, they had to complete a questionnaire (Fig. 3). In this questionnaire, the result of each operation in each table of the schema had to be recorded. For example, if as a result of a deletion of a row in one table of the schema another table was also changed (that is to say, the deletion of a row of this other table was required due to a referential integrity relationship) in the questionnaire, the subjects must note down one row

for the first table and one row for the second table. The tables not affected by the operation were left blank.

Before the subjects took the test, the experiment was conducted with a small group of people in order to improve it and ensure that both the experiment and the documentation, were well designed. As a result of this pre-test, the only change required was in the questionnaire paper, that was changed to make it easier to use (we made it in a tabular form).

Tests were performed over 1 h. Before starting the experiment, it was explained to the subjects, what kind of exercises they had to do, the material they would be given, what kind of answers they had to provide and how they had to record the time spent solving the problem.

Before starting each test, the subjects had to record the start time, and when they had completed it, they had to record the end time. In this way, when a subject finished a test, he was able to go on to the next one without waiting for the rest of the participants in the experiment.

Tests were performed in a different order by the different subjects in order to avoid learning effects. As all the tests were correctly answered to obtain the results of the experiment, we used the number of minutes needed by each subject.

3.5. Validity of results

As we know, different threats to the validity of the results of an experiment exist. In this section, we are going to discuss threats to construct, internal and external validity.

3.5.1. Construct validity

Construct validity is concerned with the relationship between theory and observation [20]. We propose, as a reasonable measure of analyzability, the time for determining the final state of the database (after performing the tasks).

We must point out that the time recorded was the time used in answering the questions of the tests (in executing the operations) but also the time needed by the subject to analyze and to understand the initial state of the database. To assure construct validity, it would be necessary to perform more experiments, varying the operations to be developed.

3.5.2. Internal validity

Internal validity is related to the assurance that the relationship observed between the treatment and the outcome is a causal relationship, and that it is not a result of a factor over which we have no control or have not measured [20]. Regarding internal validity, the following issues must be considered:

- *Differences among subjects.* Within-subject experiments reduce variability among subjects. In the experiments, all the subjects had approximately, the same experience working with relational databases.
- *Differences among schemas.* We designed the schemas with six tables, which were related with more or less foreign keys depending on the metrics values. The domains of the schemas were different and this could influence in some way, the results obtained.
- *Precision in the time values.* The subjects were responsible for recording the start and finish times of each test. We think that this method is more effective than having a supervisor who records the time of each subject. However, we are aware that the subject could introduce some imprecision.
- *Learning effects.* The tests were put in different order for different subjects in order to prevent learning effects. So, each subject answered the tests in the given order.
- *Fatigue effects.* The average time for completing the experiment was an hour, so fatigue effects, hardly exist at all. Also, the different order of the tests helped to avoid these fatigue effects.
- *Persistence effects.* We must be sure that, when a set of subjects perform an experiment, the effects of previous similar experiments do not persist. In our case, persistence effects are not present because the subjects had never performed a similar experiment.
- *Subject motivation.* We told the subjects that the results of the experiment were anonymous and that the information obtained would never be commented on with other workers or with the managers. We think, though, motivation is greater when working with students than with professionals.
- *Other factors.* Plagiarism and influence among subjects were controlled. Subjects were informed that they should not talk or share answers with other subjects. Nevertheless, the subjects were not watched or controlled during the experiment, so we cannot assure that the mentioned effects do not appear, although we think it improbable.

3.5.3. External validity

External validity is concerned with generalization of the results [20]. Regarding external validity, the following issues must be considered:

- *Materials and tasks used.* We tried to use schemas and operations representative of real cases in the experiments although more experiments with larger and more complex schemas are necessary.
- *Subjects.* Due to the difficulty of getting professionals to perform the experiments, the original experiment was done by students. In general, more experiments with a larger number of subjects, students and professionals, and with a greater difference between the values of each metric are necessary to obtain more conclusive results regarding the relationship between referential integrity and the analyzability of the relational databases, and, hence, their maintainability.
- We tried to increase external validity by performing the

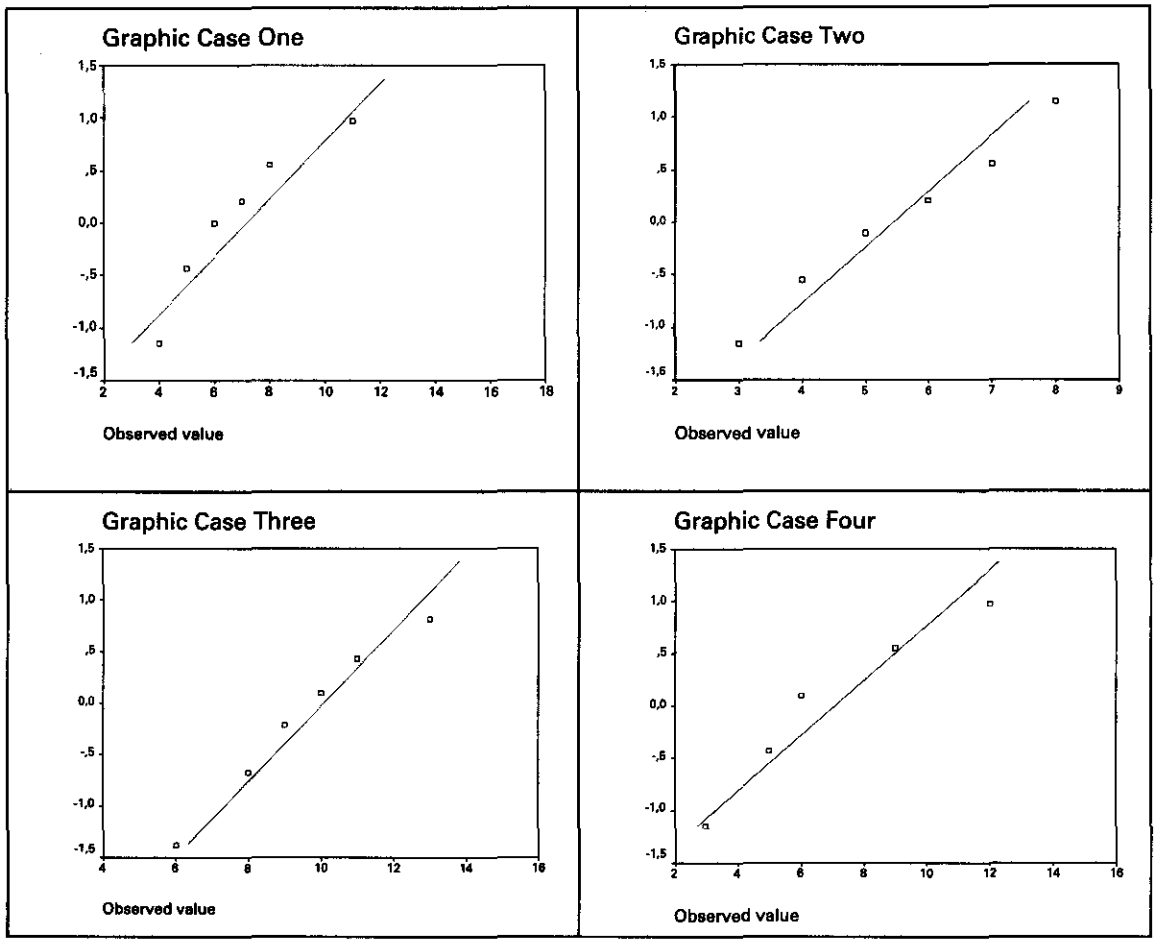


Fig. 4. Normality tests for the data of the four cases.

replica with professionals, so the results could be more generalized.

4. Analysis and interpretation

As all the answers for the four tests were correctly answered, we had 11 data sets for each case, one of which corresponded to the answer of each subject of the experiment.

First, we had to check the normality of the data obtained. If the data were normal, the best option in our case was to use parametric tests because they are more efficient than non-parametric tests and because they require fewer data points (this point is very important in our case because we only have 11 data points), and therefore smaller experiments than do non-parametric tests.

Table 4
Descriptive statistics of the cases

	Minimum	Maximum	Mean	Standard deviation
Case 1	4	16	7.1818	3.6005
Case 2	3	8	5.4545	1.8635
Case 3	6	15	10.0909	2.7002
Case 4	3	15	7.0909	3.7803

Among the tests, we can apply to ascertain if a distribution is normal, we have the Shapiro–Wilk test and the Kolmogorov–Smirnov test. Both tests were applied to our data and we found that they were normal. In Fig. 4, the normality test graph is shown. As we can see, the points are near to the line, so the data can be considered normal and the parametric tests can be applied.

In Table 4, the values of some descriptive statistics, as the minimum, the maximum, the mean, and the standard deviation of the dependent variable (the time) of the four cases are presented.

To proceed to the analysis, we have to first preset a level of significance. Several factors have to be considered when setting α, because we can commit a type I error (probability of incorrectly rejecting the null hypothesis). We decided to select $\alpha = 0.1$ which means a 90% level of confidence.

After all these considerations, we can now select the best test to apply to our data for obtaining results susceptible of being interpreted. As we have previously indicated, we have

Table 5
Results of the F-statistic for the experiment replication

Source of variation	Degrees of freedom	F-ratio	Sig
DRT	1	9.137	0.013
NFK	1	12.659	0.005
Interaction	1	0.631	0.445
Error	1		
Total	40		

normal data, so parametric tests are the appropriate ones, and the most appropriate one taking into account our hypothesis is the F-statistic, which is extremely robust.

In Table 5, we present the results obtained from this statistic with the values of the dependent variables in the four cases. In this table, the first column represents the source of variation. The second column represents the degrees of freedom, the third one indicates the results obtained for our experiment, and these values must be compared to the table values. The last column represents the level of significance. In each row of the table, we have the two factors of the experiment, the interaction, the error and the total.

The values which appear in the F-ratio column have to be compared with the appropriate value in the tables. In our case, the value we need is $F_{1,40}$. The numerator is one (in the three cases, DRT, NFK and interaction, we have one degree of freedom) and the value 40 corresponds to the degrees of freedom of the error (sum of number of subjects in each test and the degrees of freedom of the three first rows of the table). Consulting the tables we find that $F_{1,40} = 2.84$.

The way to know if the hypothesis is true or not is by comparing the value in the tables with the value obtained in the experiment. If the value obtained is greater than the value in the tables, the hypothesis can be considered valid if not it can be considered invalid.

H1. The value of the DRT metric affects the analyzability of the database schema.

The first row of Table 3 corresponds to the values obtained for the dependent variable related to hypothesis one. As $9.137 > 2.84$, alternative hypothesis one is valid and the value of the DRT metric affects the results obtained.

H2. The value of the NFK metric affects the analyzability of the database schema.

The second row of Table 3 corresponds to the values obtained for the dependent variable related to hypothesis two. As $12.659 > 2.84$, alternative hypothesis two is valid and the value of the NFK metric affects the results obtained.

H3. The interaction of the DRT and NFK metrics affects the analyzability of the database schema.

The third row of Table 3 corresponds to the values obtained for the dependent variable related to hypothesis three. As $0.631 < 2.84$, alternative hypothesis 3 is invalid and the interaction of the values of the DRT and the NFK metrics do not affect the results obtained.

As a summary of our experiment we can say that both metrics seem to be good indicators of the analyzability of a relational database schema. This means that we can use the value of these metrics for comparing schemas and select the easier one (taking into account that both must be semantically equivalent). We have also demonstrated that there is no interaction between the metrics.

From the experiments as a whole, we can conclude that NFK seems to be a good indicator of the relational databases analyzability.

However, it is difficult to draw a conclusion for the DRT metric because we have obtained different results in each experiment. Perhaps, this difference is due to the different subjects selected for each experiment (novice and experienced designers). It would be necessary to make more experiments with this metric in order to come to a more solid conclusion about it.

In general, more experiments with the two metrics would be necessary in order to obtain more conclusive results.

5. Conclusions and future work

Metrics are useful mechanisms for improving the quality of software products and also for determining the best ways to help professionals and researchers. For conventional databases, such as relational ones, only the normalization theory has been traditionally used. We have presented two metrics for relational databases related to referential integrity (NFK and DRT) with the aim of providing additional objective indicators to the normalization theory for helping database designers to take the best design decisions.

Performing empirical validation with the metrics is fundamental in order to demonstrate their practical utility. In this line, we have summarized two previous empirical studies made with metrics for relational databases: a controlled experiment and a case study.

A detailed description of the replica of the controlled experiment with the NFK and DRT metrics, has been presented in this paper.

From the experimentation process as a whole, we can conclude that the NFK metric is a good metric for measuring the analyzability of a relational database schema. However, it is difficult to draw a conclusion for the DRT metric because we have obtained different results in each of the controlled experiments. This difference could be due to the different subjects who performed the experiment, the first and the second time. It would be necessary to perform more experiments with the metrics to obtain more solid conclusions about them (mainly about DRT).

Some changes that could be made to improve the presented experiment are:

- *To increase the size of the schemas.* By increasing the size of the schemas, we can have examples that are closer to reality. Also, as the examples are more real, if we are

working with professionals, we can make better use of their potential capability and the results can be more generalized.

- *To increase the difference between the values of the metrics.* This option could lead to more conclusive results about the metrics and their relationship with the factor we are trying to control.
- *To work with real data.* Another way to obtain more conclusive results about metrics is by working with real data in additional case studies.

Acknowledgements

This research is part of the MANTICA project, partially supported by the CICYT and the European Union (CICYT-1FD97-0168).

References

[1] N. Fenton, S.L. Pfleeger, Software Metrics: A Rigorous Approach, second ed, Chapman & Hall, London, 1997.
[2] Software Measurement, in: A. Melton (Ed.), International Thomson Computer Press, London, 1996.
[3] H. Zuse, A Framework of Software Measurement, Walter de Gruyter, Berlin, 1998.
[4] S.L. Pfleeger, Assessing software measurement, IEEE Software (1997) 25–26.
[5] T.J. McCabe, A complexity measure, IEEE Transactions on Software Engineering 2 (5) (1976) 308–320.
[6] H.M. Sneed, O. Foshag, Measuring legacy database structures, in: Coombes, Van Huysduynen and Peeters (Eds.), Proceedings of the European Software Measurement Conference FESMA 98, Antwerp, May 6–8, 1998, pp. 199–211.
[7] N. Leavit, Whatever happened to object-oriented databases? Industry trends, IEEE Computer Society (2000) 16–19.
[8] R.H.M. Gray, B.N. Carey, N.A. McGlynn, A.D. Pengelly, Design metrics for database systems, BT Technology J 9 (1991) 69–79.
[9] ISO, ISO 9126, Software product evaluation—quality characteristics and guidelines for their use. ISO/IEC Standard 9126, Geneva, 1999.
[10] H.F. Li, W.K. Cheng, An empirical study of software metrics, IEEE Transactions on Software Engineering 13 (6) (1987) 679–708.
[11] E.J. Weyuker, Evaluating software complexity measures, IEEE Transactions on Software Engineering 14 (9) (1988) 1357–1365.
[12] L.C. Briand, S. Morasca, V. Basili, Property-based software engineering measurement, IEEE Transactions on Software Engineering 22 (1) (1996) 68–85.
[13] S. Morasca, L.C. Briand, Towards a theoretical framework for measuring software attributes, Proceedings of the Fourth International Software Metrics Symposium, 1997, pp. 119–126.
[14] S.A. Whitmire, Object Oriented Design Measurement, Wiley, New York, 1997.
[15] M. Piattini, C. Calero, M. Genero, Table oriented metrics for relational databases, Software Quality Journal 9 (2001) 79–97.
[16] C. Calero, M. Piattini, M. Genero, M. Serrano, I. Caballero, Metrics for relational databases maintainability, United Kingdom Academy of Information Systems, UKAIS 2000, Cardiff, UK, ISBN 0-07-7097556, McGraw-Hill, New York, 2000, pp. 109–119.
[17] V.R. Basili, F. Shull, F. Lanubille, Building knowledge through families of experiments, IEEE Transactions on Software Engineering (4) (1999) 456–473.
[18] K. Siau, Information modeling and method engineering: a psychological perspective, Journal of Database Management 10 (4) (1999) 44–50.
[19] J.R. Anderson, The Architecture of Cognition, Harvard University Press, Cambridge, 1983.
[20] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, Dordrecht, 2000.
[21] C. Calero, M. Piattini, M. Genero, A case study with relational database metrics, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2001), Beirut, Lebanon, June 26–29, 2001.