Software Qual J (2008) 16:79–106
DOI 10.1007/s11219-007-9030-7

Empirical studies to assess the understandability of data warehouse schemas using structural metrics

Manuel Angel Serrano · Coral Calero · Houari A. Sahraoui · Mario Piattini

Published online: 11 July 2007

Abstract Data warehouses are powerful tools for making better and faster decisions in organizations where information is an asset of primary importance. Due to the complexity of data warehouses, metrics and procedures are required to continuously assure their quality. This article describes an empirical study and a replication aimed at investigating the use of structural metrics as indicators of the understandability, and by extension, the cognitive complexity of data warehouse schemas. More specifically, a four-step analysis is conducted: (1) check if individually and collectively, the considered metrics can be correlated with schema understandability using classical statistical techniques, (2) evaluate whether understandability can be predicted by case similarity using the case-based reasoning technique, (3) determine, for each level of understandability, the subsets of metrics that are important by means of a classification technique, and assess, by means of a probabilistic technique, the degree of participation of each metric in the understandability prediction. The results obtained show that although a linear model is a good approximation of the relation between structure and understandability, the associated coefficients are not significant enough. Additionally, classification analyses reveal respectively that prediction can be achieved by considering structure similarity, that extracted classification rules can

M. A. Serrano (✉) · C. Calero · H. A. Sahraoui · M. Piattini
Alarcos Research Group – Department of Information Technologies and Systems,
Universidad de Castilla – La Mancha, Paseo de la Universidad, 4, 13071 Ciudad Real, Spain
e-mail: Manuel.Serrano@uclm.es

C. Calero
e-mail: Coral.Calero@uclm.es

H. A. Sahraoui
e-mail: sahraouh@iro.umontreal.ca

M. Piattini
e-mail: Mario.Piattini@uclm.es

H. A. Sahraoui
Dep. d'Informatique et de Recherche Opérationnelle, Université de Montréal,
CP 6128 succ. Centre Ville, Montreal, QC, Canada H3C 3J7
e-mail: sahraouh@iro.umontreal.ca

be used to estimate the magnitude of understandability, and that some metrics such as the number of fact tables have more impact than others.

Keywords Data warehouse · Quality · Metrics · Empirical studies

1 Introduction

Companies musSQJO11219-9030Nagendran32t manage information as a product of primary importance. They must capitalize knowledge, treating it as a principal asset. By doing this, they will be able to survive and prosper in the digital economy (Huang et al. 1999). It is in this context that data warehouses emerged as an efficient technology some years ago.

Data warehouses are large repositories created to hold data drawn from several data sources and maintained by different operating units together with historical and summary transformations. A data warehouse can be treated as a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions.

Due to the increasing complexity of data warehouses (Inmon 1997), continuous attention must be given to the evaluation of their quality throughout the development process. As stated in Bouzeghoub and Kedad (2002), quality in data warehouses, as in other software products, is crucial. Following the standard ISO 9126 (ISO 2001), quality can be defined as the extent to which a product satisfies stated and implied needs when used under specified conditions. Bad data warehouse design may lead to the rejection of the decision support system or may result in non-productive decisions.

At this time, there are no well-established and complete methodologies for designing data warehouses. Moreover, as with any software product, using a design methodology cannot be a unique guarantee for obtaining a good product. Hence, the final goal of our work is to define a set of structural metrics for assuring data warehouse quality. The aim of these metrics is to help designers choose among alternative schemas that are semantically equivalent. They can also be used to improve the quality of the resulting products.

Obtaining a valid set of metrics, however, is not only a matter of definition. Two types of validation are necessary, theoretical and empirical. Theoretical validation is used to verify analytically that the metrics are proper numerical characterizations of the measured attribute (conforming to a set of formal properties). We conducted such a validation on data warehouse metrics in a previous work (Serrano et al. 2005).

Empirical validation is useful to show that the metrics can be used in practice to predict/assess a quality attribute, and this kind of validation is crucial for the success of any software measurement project, as it helps us confirm and understand the implications of the product metrics (Basili et al. 1999; Fenton and Pfleeger 1997; Kitchenham et al. 2002; Schneidewind 2002). A proposal of metrics has no value if their practical usefulness is not established through empirical validation.

In this article, we present an empirical validation of the use of metrics for data warehouses in their quality assessment. This study consists of an initial experiment and a replication, using several techniques for analyzing the experimental data. Our aim, when using several techniques, is to extract as much information as possible from the experimental data. In the next section, we present the metrics used in the experiment. The experimental setting is described in Sect. 3. Section 4 is devoted to the analyses and results

of the experiments. Finally, discussions and conclusions are given in Sects. 5 and 6, respectively.

2 Data warehouse concepts

2.1 Data warehouse

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process (Inmon 1997). A data warehouse is

- subject-oriented, as data provides information regarding a particular subject rather than a company's ongoing operations,
- integrated because data is gathered for the data warehouse from a variety of sources and merged into a coherent whole,
- time-variant, as all data in the data warehouse is identified with a particular time period,
- and non-volatile because data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Usually, data warehouses are structured in the form of star schemas (see Fig. 1). Each star schema consists of one or more fact tables and several dimensional tables. The facts of interest are stored in the fact table (e.g., sales, inventory). Figure 1 shows two fact tables, Orders and Deliveries. Dimensional tables store information about the context of the facts (Jarke et al. 2000). Figure 1 contains four dimensional tables (Product, Date, Client and Salesman).

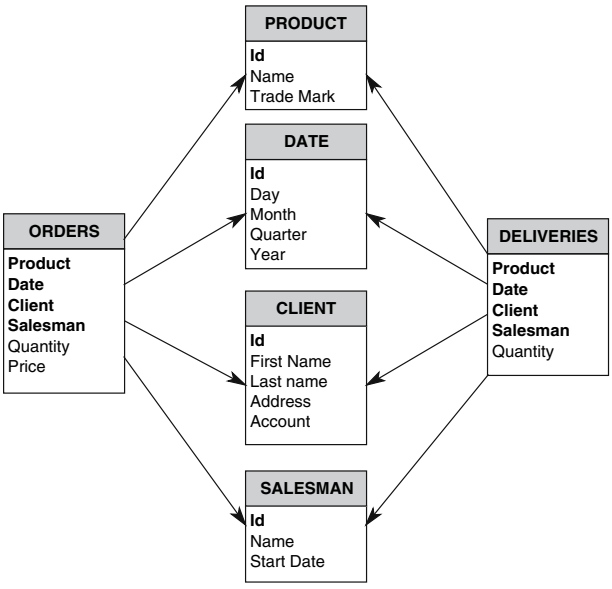

Fig. 1 Example of a star schema

2.2 Data warehouse quality

A first step towards obtaining quality data warehouses has been the definition of development methodologies (Anahory and Murray 1997; Deveboise 1999; Kimball et al. 1998). However, as stated earlier, a methodology, though necessary, may not be sufficient to guarantee the quality of a data warehouse. Indeed, a *good* methodology could lead to *good* products. However, many other factors could influence the quality of the products, such as human decisions. It is thus necessary to complete specific methodologies with metrics and techniques for product quality assessment.

Structural properties (such as structural complexity) of a software artefact have an impact on its cognitive complexity as shown in Fig. 2. Cognitive complexity means the mental burden on those who have to deal with the artefact (e.g. developers, testers, maintainers). High cognitive complexity of an artefact reduces its understandability and leads to undesirable external quality attributes as defined in the standard ISO9126 (ISO 2001). The model presented in Fig. 2 is an adaptation of the general model for software artefacts proposed in Briand et al. (1998). Indeed, as data warehouse schemas are software artefacts, it is reasonable to consider that they follow the same pattern. It is thus important to investigate the potential relationships that can exist between the structural properties of these schemas and their quality factors.

2.3 Data warehouse metrics

In a previous study, a set of metrics for logical data warehouse schemas was defined (Calero et al. 2001). A preliminary validation of theses metrics, although performed on only six schemas, revealed that some of them were potential predictors (Serrano et al. 2005). Our current study uses four of these candidate metrics. In addition to the initial validation motivation, the reasons behind this selection are provided in Sect. 3.2. The chosen metrics are defined in Table 1.

To illustrate the definitions of the proposed metrics, Table 2 provides results for the calculation of the four metrics for the schema in Fig. 1.

Before using the selected metrics, we validated them theoretically, in order to ensure that the intuitive idea of the measured attributes is captured by the metrics. There are two main approaches to theoretical validation of metrics: frameworks based on measurement theory and those based on axioms. The first helps us to identify the scale of a metric and thus determine which operations can be applied to it. Frameworks based on axiomatic

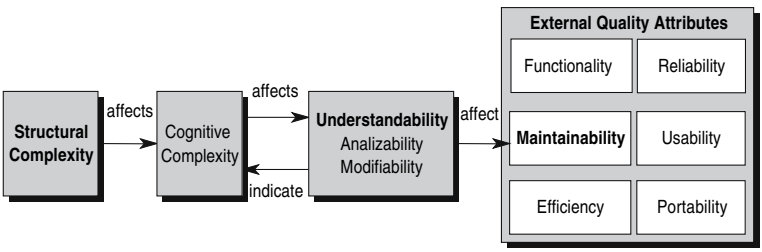

Fig. 2 Relationship between structural properties, cognitive complexity, understandability and external quality attributes - based on the work described in Briand et al. (1998)

Table 1 Metrics at the schema level

NFT(Sc). Number of fact tables in the schema
NDT(Sc). Number of dimension tables in the schema
NFK(Sc). Number of foreign keys in all the fact tables of the schema $NFK(Sc) = \sum_{i=1}^{NFT} NFK(FT_i)$ Where $NFK(FT_i)$ is the number of foreign keys in the fact table i of the schema Sc
NMFT(Sc). Number of facts in the fact tables; Number of attributes in the fact tables that are not foreign keys NMFT(Sc)=NA(Sc)–NFK(Sc) Where NA(Sc) is the number to attributes in the fact tables of the schema Sc

Table 2 Metrics values

Metric	Value
NFT	2
NDT	4
NFK	8
NMFT	3

approaches are used to classify the metric. Our theoretical validation is based on measurement theory: Zuse's formal framework (Zuse 1998) and the DISTANCE formal framework (Poels and Dedene 1999), and the axiomatic approach of Briand's formal framework (Briand et al. 1996). As a result of this theoretical validation (see Table 3), it appears that all of the metrics are at least in the interval scale (ratio for the DISTANCE framework). This means that they are theoretically valid software metrics according to the frameworks used. Details about the theoretical validation process can be found in Serrano et al. (2005).

Now that the schema metrics are defined and theoretically validated, the next step is to determine through experimentation that these metrics can be used in practice for quality assessment. A first empirical validation using basic statistical techniques was conducted and discussed in Serrano et al. (2002) and Serrano et al. (2005). This validation indicated that structural metrics are related to complexity. In these previous studies, we performed the theoretical and empirical validation of the proposed metrics. The empirical validation of the metrics was performed exclusively by means of classical statistical analysis. Subsequent to these studies, it is worthwhile to strengthen the results obtained and to use advanced techniques that support the first findings. The remainder of this article is dedicated to study this relationship in depth through two experiments using some advanced analysis techniques.

Table 3 Theoretical validation of the metrics

Metric	Scale (Zuse 1998)	Scale (DISTANCE)	Briand et al. (1996)
NFT	Ratio	Ratio	Size
NDT	Above ordinal	Ratio	Size
NFK	Above ordinal	Ratio	Complexity
NMFT	Above ordinal	Ratio	Size

3 Experimental design

3.1 Hypotheses

The first step in our empirical study is to define the goals of our experiment and state the related hypotheses. As stated earlier, our main goal is to "define a set of metrics to assess and control the quality of logical data warehouse star schemas". The focus is specifically put on finding the set of valid structural complexity metrics that affects the understandability of logical data warehouse schemas. This is done by studying the relationship between structural metrics and understandability.

The main hypothesis of our empirical work is: "Star schema metrics can be used as indicators for schema understandability".

This main hypothesis is refined into four specific hypotheses:

1. The relationship between structure measures of a schema and its understandability is linear.
2. Similar schema structure leads to similar understandability.
3. Different level of understandability can be predicted by different subsets of metrics.
4. Structural metrics have different impacts on understandability prediction.

For assessing the relationship between the metrics and the understandability of the schemas, a correlational study will be carried out, followed by a linear regressional study to see if this relationship is linear. A Case Base Reasoning (CBR) technique will be used to determine if similar schemas (defined in terms of the metrics) exhibit similar behaviour in terms of complexity and understanding. Another way of viewing the problem is to see if specific subsets of metrics are better indicators of specific understandability levels. Formal Concept Analysis (FCA) is known to be a good tool to achieve this goal. Finally, a Bayesian classifier (BC) will be used to determine the degree of participation of the metrics in the decision about understandability classification.

The combination of these four analysis techniques will provide enough information to draw a conclusion on the possible use of a subset of metrics as indicators of schema understandability.

3.2 Variables

To be evaluated experimentally, the hypothesis is mapped to a set of independent and dependent variables (metrics).

3.2.1 Independent variables

Metrics shown in Table 1 are the independent variables. This decision is motivated by the fact that structural complexity of a schema can be seen as the combination of its size in terms of elements (tables and facts) and the density of relationships that link these elements. Two schemas having the same size (number of elements) may have different structural complexities when considering the relationships. Consequently, NFT and NDT are selected because they count respectively, the numbers of fact and dimensional tables in the schema. The number of Foreign Keys (NFK) represents the number of relationships in the schema and, consequently, the complexity of the "navigation". Finally, as fact tables are the starting points for navigation when using a data warehouse, it is important to

measure their complexity. This is done by counting the attributes (which are not foreign keys) stored in the data warehouse (NMFT). Moreover, these four metrics were found to be good indicators for cognitive complexity in a previous study (Serrano et al. 2005).

3.2.2 Dependent variables

It is difficult to directly measure the understandability of a schema. A subject needs to understand a schema for a particular purpose and not in an absolute manner. Thus, one way to capture understandability is to measure the time subjects spend accomplishing tasks. In this way, difficulty is based exclusively on the effort required to understand the schema. In other words, the more complex the schema, the more effort required to understand it, and the more time spent working with it.

Accordingly, in our study, the average time (in seconds) that a group of subjects spends performing the experimental tasks on a particular schema is the dependent variable. Subjects record the starting and finishing time of each task which allows deriving the time.

3.3 Analysis techniques

One of the main difficulties when conducting an empirical study is to define a representative data sample. Purposing sampling rules (schema size, application domain, etc.) result in 13 data warehouse schemas extracted from data warehouse textbooks. Considering the nature of the objects of our study (data warehouses) and the sampling rules, this number is fairly interesting because schemas cover more than 10 different application domains. Moreover, the schemas are different in terms of metrics values, and thus provide us with a good set of objects for our study. To take advantage of this sample, we decided to perform different types of analyses to study different properties. In the remainder of this section, a brief description is provided for each of them. A summary of the study process and purposes of the techniques used are provided in Sect. 3.3.5.

3.3.1 Correlation and linear regression

As a first step in the analysis, we hypothesized that the relationship between the metrics and understandability (time) is linear. To study this relationship, Spearman's Rho correlation (individual correlations) and multivariate linear regression (with ANOVA analysis) are the tools of choice.

3.3.2 Case base reasoning

The goal of the first technique is to abstract a predictive model (linear regression). However, due to the relatively small sample size, we explored a similarity-based technique to verify if each particular case could be assessed using its similarity to the others.

For this kind of analysis, CBR is known to be an appropriate technique (Grosser et al. 2003). CBR emerged as an approach to problem solving in *weak theory* domains (domains where little is known about key processes and their dependencies). It is usually used as a prediction mechanism, since when a new case arises the CBR-based algorithms try to find the most similar past cases in order to solve the new problem. In our context, this theory is based on the belief that data warehouses that have similar structures will have a similar understandability level.

More specifically, each schema s is mapped to an n-tuple ($m_1(s)$, $m_2(s)$, $m_3(s)$,..., $m_n(s)$, t) where $m_i(s)$ is the value for the structural metric m_j and t is a value of the time measure. Following the principle of the leave-one-out cross validation technique, the time value is predicted for each schema and compared with the actual value.

Similarity between cases can be assessed using classical distance measures such as Euclidean and Manhattan distances (Wilson and Martinez 1997). Indeed, a case (schema) can be seen as a point in an n-dimensional Euclidean space, where coordinates correspond to metrics m_j. The measure used in this analysis is derived from the Manhattan distance. For a given metric, the Manhattan distance is adapted through the division by the definition domain to obtain relative distances; this allows the combination of metrics from different magnitudes.

Our distance is basically a linear combination of the point-wise differences (absolute values) between the vectors representing a pair of schemas. Formally:

$$\mathrm{Dis}(s, s') = \sum_{j=1}^{n} \beta_j \mathrm{dis}(m_j(s_i), m_j(s'_i))$$

where $\beta_j \geq 0$ is the weight of the metric m_j and $dis(m_j(s), m_j(s'))$ is the dissimilarity with respect to the metric m_j that can be calculated as follows:

$$\mathrm{dis}(m_j(s), m_j(s')) = \frac{\left|m_j(s) - m_j(s')\right|}{\left|\mathrm{dom}(m_j)\right|}$$

$|\,\mathrm{dom}(m_j)|$, used to normalize the distances, stands for the maximal difference of two values v_1, v_2 in $\mathrm{dom}(m_j)$.

As there is no a priori knowledge about the weight of each structural metrics, all the weights β_j's are considered equal to 1, i.e., all the metrics contribute at the same level to the distance measure.

3.3.3 Formal concept analysis

The third step in our analysis is to study the subsets of the structural metrics that could be the best indicators of the understandability of the schemas (a feature selection problem). An interesting technique used to reach this goal is the FCA (Godin et al. 1995). Formal concept analysis provides a way of identifying groupings of elements (referred to as objects in FCA literature) that have common properties (referred to as attributes in FCA literature). All the possible groups (called concepts) are organized in a concept lattice.

In our problem, the objects are the schemas and the attributes are the metrics. A concept is therefore a group of schemas that share similar values for the metrics (the understandability level or time is not used to build the lattice). In the obtained lattice, groups that share the same understandability level indicate which (level of) metrics are also involved.

3.3.4 Bayesian classifiers

The final step in the analysis aims to determine the degree of participation of the metrics in the decision about understandability classification. Bayesian classifiers offer a very interesting approach for this kind of analysis (Ramoni and Sebastiani 1999). A BC is trained by estimating the conditional probability distributions for each attribute from a sample. The

classification of each case is made using the Bayes' Theorem. Indeed, for a schema with the values (v_1, ..., v_n) for respectively the metrics (m_1,...,m_n), the probability that it has the level of understandability c is defined as:

$$P(c/v_1, \ldots, v_n) = \frac{P(v_1, \ldots, v_n/c)P(c)}{P(v_1, \ldots, v_n)}$$

with Bayes' assumption that the values taken on by the different metrics are conditionally independent given the understandability level (time). For a given case, the predicted understandability level is the one that maximizes the calculated probability (Flach and Lachiche 1999)

$$P(c)P(v_1/c) \ldots P(v_n/c)$$

For our experiment, ROC classifier tool described in Ramoni and Sebastiani (1999) is used.

3.3.5 *Summary of the analyses*

Figure 3 summarizes the experiment analysis process and Table 4 shows the summary of the analyses used along with the underlying assumptions.

3.4 Data collection

Some 13 logical data warehouse schemas were used in performing the experiments. Although the domain of the schemas was different, we tried to select examples that represent real cases, in such a way that the understandability measures obtained (time) were due to the structure of the schema and not to the complexity of the domain problem. The

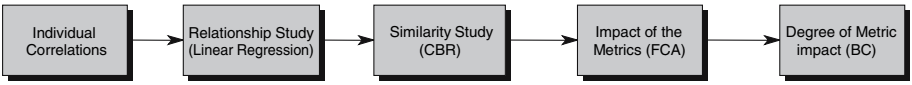

Fig. 3 Steps in the analysis process

Table 4 Summary of analyses

Analyses	Assumption	Technique
Individual correlations	Individual correlation between understandability and the metrics	Spearman correlation
Relationship study	Linear relationship between structural metrics and time	Multivariate linear regression
Similarity study	Similar structures lead to similar understandability level	CBR
Impact of the metrics	Which metrics are joint factors for understandability	FCA
Degree of the metric impact	Metric values conditionally independent given time	BC

Table 5 Metrics values for the experiment schemas

Schema	NFT	NDT	NMFT	NFK
S01	1	2	2	2
S02	2	4	3	8
S03	1	3	3	3
S04	1	4	2	4
S05	1	2	2	2
S06	1	3	2	3
S07	1	7	4	7
S08	1	7	5	7
S09	2	8	5	12
S10	2	5	4	7
S11	1	4	2	4
S12	2	3	12	2
S13	2	5	3	9

distributions of metrics are also important criteria, i.e., schemas with different complexity and metric values (see Table 5).

The dependent variable (average time of understandability) was collected by asking groups of subjects to perform tasks on the selected schemas as explained in Sect. 3.2.2. The documentation, for each data warehouse, was composed of a schema, a task to perform (one question to answer) and a space for writing the results of the task. For each schema, the subjects had to select some information by "navigating" through the tables of the schema. Subjects had to indicate (in natural language) which information had to be recovered from which table in order to obtain a specific result. Figure 4 shows an example of the question/answer form used in our paper-based experiment. The other tasks in the experiment were similar to the one shown in Fig. 4, i.e., same kind of question.

We selected a within-subject design experiment (i.e. all the tasks had to be performed by each of the subjects). Schemas were given in different orders to the subjects, to minimize the effects of learning and fatigue. The data collected in the experiment consisted in the number of seconds required by each subject to perform a task on each schema.

Before starting the experiment, the subjects took a mini-tutorial that explains the kind of tasks they had to perform, the material they would be given, what kind of answers they had to provide and how they had to record the time spent performing the tasks (time in hours, minutes and seconds before and after working on a particular task).

Some 24 volunteer subjects participated in the first experiment. Nine of them were graduate students from the University of Montreal (Canada), seven were graduate students from the University of Castilla – La Mancha (Spain) and eight were teachers at that same university.

The students from the University of Montreal had a good knowledge of software engineering and relational databases. However, they had not taken a data warehouse-related course. The graduate students from the University of Castilla – La Mancha had a similar profile, but several of them had conceptual knowledge of data warehouses. The eight teachers from the University of Castilla – La Mancha had experience in the three fields (software engineering, databases and data warehouses), as they are working on these topics.

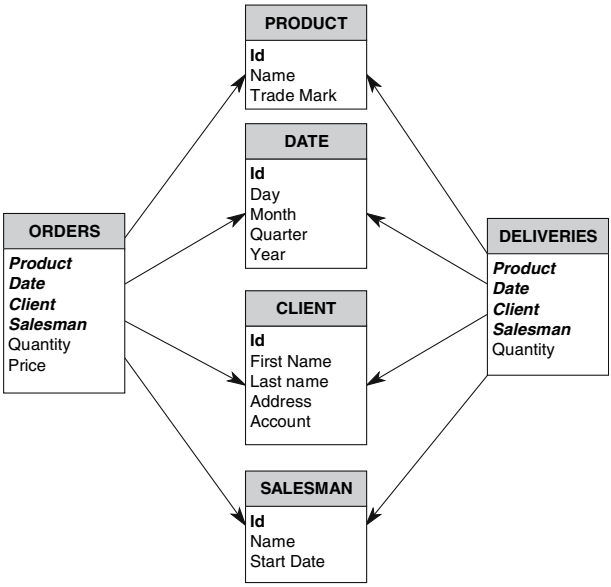

Fig. 4 Question/answer sheet

Some 85 students from the University of Castilla – La Mancha participated in the replication. These subjects were also volunteers. They were enrolled in an information retrieval course, offered in the fifth year of computer engineering studies. All of the students had considerable knowledge of the field of data warehouses because they were been involved in a complete course about data warehouses where all the concepts related to data warehouses design were explained and another about databases where many of the concepts related with the relational model (used for the star design of data warehouses) were also explained.

4 Analyses and results

4.1 Collected data validation

After running the first experiment, data was validated in order to avoid noise. The first step of the data validation consisted of checking the subjects' answers and eliminating incorrect ones. Answers were considered correct if they were complete, i.e., subjects answered the questions in full. Incorrect answers were not considered to avoid noise in the data. Indeed, incorrect answer cannot be explained only by the complexity. Other possible reasons are lack of motivation and subject ability to perform the task. For the remaining schema-subject pairs, we time values in seconds were calculated.

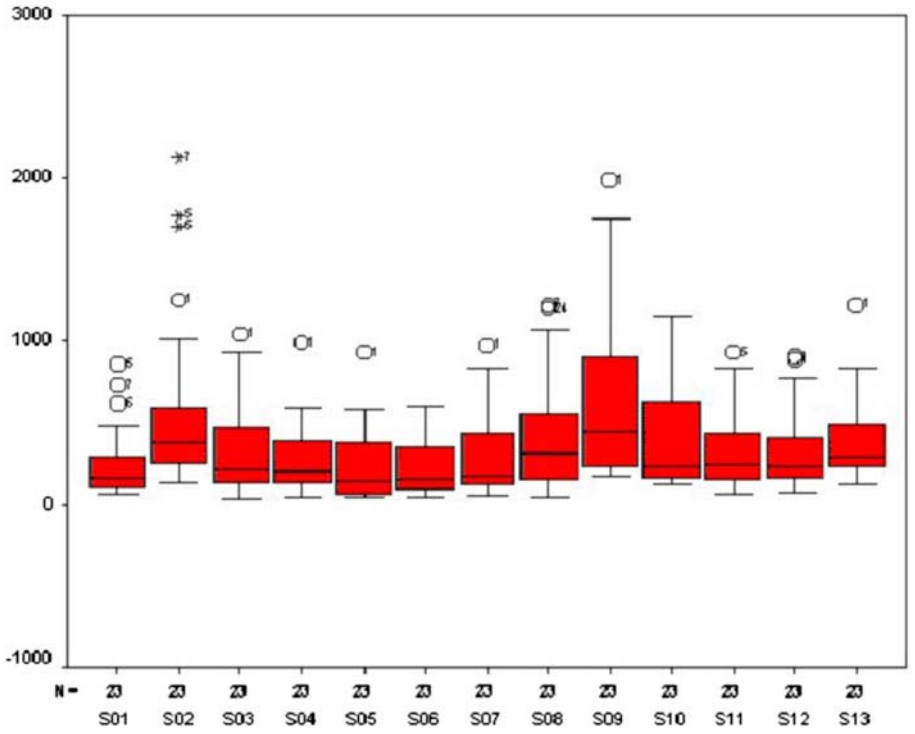

Fig. 5 Box plot of the experiment data

The second step for pre-processing consisted in eliminating outliers. Moreover, when the time of the same subject was considered as outlier for many schemas, his ability to perform the experience without biasing the results was questionable. Therefore, to avoid any subject biases, such a subject was eliminated.

The detection of outliers was carried out using the box plot technique, shown in Figs. 5 and 6. In these figures, the horizontal axis represents the schemas and the vertical axis represents the time spent performing the tasks.

Figure 5 reveals several outliers displayed with O or * followed by the number of the subject. The time for Subject 1 was considered outlier for nine schemas, which causes its elimination from the experiment. The other outlier values were eliminated from the collected data. Table 6 shows the details (descriptive statistics) about the time variable. Descriptive statistics for the final set of data are presented in Table 7. This data set was used in all the analyses.

We performed the same analysis with the collected data of the replication (see box plot in Fig. 6[1]). Few outliers were found and eliminated, but no subject had too many outliers to be eliminated. In addition to the outliers, 21 subjects had missing answers for many questions. Too many missing answers is an indication about the lack of motivation. To avoid motivation biases and the number of subjects is too large, it was reasonable to

[1] Figure 6 does not show schema S07 because it was removed in the replication. See Sect. 4.3.1 for more information.

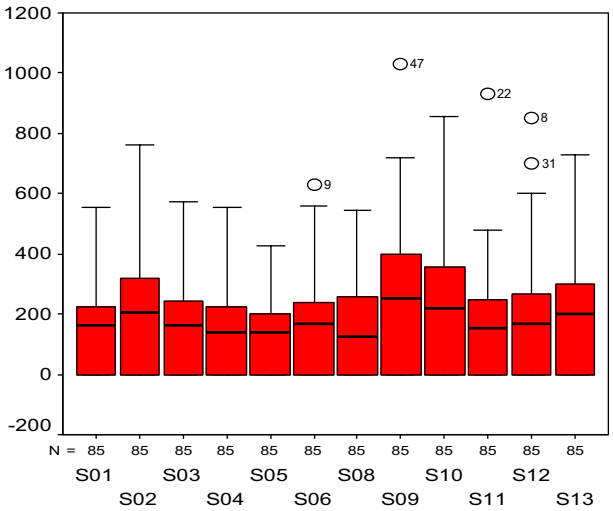

Fig. 6 Box plot of the replication data

Table 6 Descriptive statistics of the collected data

	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13
Avg	186	443	290	253	237	227	301	325	569	390	302	288	363
Min	75	129	38	50	50	52	60	52	173	118	71	79	121
Max	555	1656	936	600	708	540	859	1074	1755	1155	840	780	840
Dev	128	351	223	160	206	165	255	250	402	294	222	201	210
Median	130	338	218	203	138	153	175	263	444	241	224	240	294

Table 7 Descriptive statistics of the collected data

	S01	S02	S03	S04	S05	S06	S08	S09	S10	S11	S12	S13
Avg	189	281	205	192	172	201	241	333	302	211	228	261
Min	91	86	60	60	60	93	107	124	60	80	60	90
Max	360	559	431	405	365	382	482	719	544	480	518	488
Dev	63.8	112	83.1	79.7	69.5	60.3	104	130	112	86	99.2	91.3
Median	190	273	179	173	169	208	223	303	294	203	211	260

eliminate all of them, and kept the remaining 64 subjects. Table 7 shows the descriptive statistics of the final set of data used in the replication study.

4.2 Correlation and linear regression study

The first investigated technique was a Spearman's Rho correlation between the individual metrics and the average time for each of the schemas. This first analysis determines whether any binary relationship exists between the independent and the dependent

Table 8 Spearman's rho correlation of the first experiment data

Spearman's Rho	Correlation	Significance
NFT	.68	.01
NDT	.79	.00
NFK	.89	.00
NMFT	.58	.04

Table 9 Spearman's Rho correlation of the replication

Spearman's Rho	Correlation	Significance
NFT	.81	.00
NDT	.83	.00
NFK	.83	.00
NMFT	.75	.01

variables. As usual, a level of significance $\alpha = 0.05$ which means a 95% level of confidence is used to accept the results. The motivation behind the use non-parametric Spearman correlation is that, considering the size of the sample, it is difficult to guarantee the normal distributions of the metrics.

Tables 8 and 9 show that all the metrics are correlated with time with significance values lower than α. This means that the correlation coefficients are all high enough for the degree of freedom of our sample.

Once the presence of a relationship between the time and the metrics was established, the logical following step is to determine whether the relationship is linear. Multivariate regression analysis is used for this purpose. As it is shown in Table 10, the ANOVA analysis allows us conclude that the linear relationship is a good approximation of the relationship between the structural metrics and the understanding time, i.e., a significant F statistic.

Even though the model fit looks positive, all the coefficients in both experiments are non-significant (see Table 11). This indicates the potential for misclassification is high.

Another conclusion that appears from Table 11 is that the two studies produce two very different models for the same schemas. Indeed, except for NFT that have a positive impact (coefficient) for both studies, all the metrics have opposite impacts, negative for one model and positive for another.

4.3 Classification analyses

As discussed in Sect. 3, the size of the sample used makes it difficult to cross-validate and generalize the results obtained using regression techniques. This problem can be

Table 10 ANOVA analysis

Model (Initial)	F	Sig.	Model (Replication)	F	Sig.
(a) Initial study			*(b) Replication study*		
Regression	13.344	.001	Regression	25.707	.000

Table 11 Coefficients of the regression model

	Coefficients	*t*	Sig.
(a) Initial study			
(Constant)	153.355	1.516	.168
NFT	11.652	.114	.912
NDT	−16.293	−.462	.657
NMFT	7.747	.561	.590
NFK	35.996	1.311	.226
(b) Republication study			
(Constant)	75.093	2.096	.074
NFT	83.847	2.330	.053
NDT	20.338	1.583	.157
NMFT	−5.172	−1.063	.323
NFK	−4.677	−.479	.647

circumvented by performing analyses that use classification techniques (presented in Sects. 3.3.2–3.3.4). Indeed, classification analysis reduces the range of possible values for the dependent variable, which usually results in a more accurate prediction with less precision.

4.3.1 Data discretization

Classification techniques require a dependent variable that takes a limited set of values (classes); this is not the case for time. Consequently, it is necessary to transform time into a categorical variable with a finite set of values, two in our case: T1 (short) and T2 (long). The discretization technique takes sample values and domain knowledge into account. Table 12 shows the distribution of the schemas between these two categories. The discretization technique gave the same distribution for the initial experiment and for the replication, i.e., the classification of the schemas was the same for both experiments. This is added proof of the independence of the results coming from the subjects used.

In addition to the dependent variable transformation, FCA manipulates only Boolean attributes. Thus, a classification is also required for the structural metrics by means of the discretization algorithm we used for time values. The different categories obtained are shown in Table 13. For a metric M, each M_i representing a group is defined by an interval. For example, values of NFK are split into three groups (NFK1, NFK2 and NFK3) defined respectively by the three intervals [2,3], [4,7] and [8,12]. This classification led to the binary relation represented in Table 14. In this table, a value equal to 1 means that a schema (row) pertains to a metric category (column), and a value of 0 means the absence of such a relationship. For example, schema S09 pertains to groups NFT2, NDT3, NFK3, NMFT2 and T2 (T2 refers to the group with high average understandability time).

Table 12 Schemas of each time interval

Time	First exp	Replication	Schemas
T1	≤ 321 s	<235 s	1, 3, 4, 5, 6, 11, 12
T2	≥ 321 s	>235 s	2, 8, 9, 10, 13

Table 13 Data intervals

NFT	NFT1	NFT2	
	[1]	[2]	
NDT	NDT1	NDT2	NDT3
	[2, 3]	[4, 5]	[6, 8]
NFK	NFK1	NFK2	NFK3
	[2, 3]	[4, 7]	[8, 12]
NMFT	NMFT1	NMFT2	
	[2, 3]	[4,12]	
Time	T1	T2	

Table 14 Binary table of data intervals

	NFT1	NFT2	NDT1	NDT2	NDT3	NFK1	NFK2	NFK3	NMFT1	NMFT2	T1	T2
S01	1	0	1	0	0	1	0	0	1	0	1	0
S02	0	1	0	1	0	0	0	1	1	0	0	1
S03	1	0	1	0	0	1	0	0	1	0	1	0
S04	1	0	0	1	0	0	1	0	1	0	1	0
S05	1	0	1	0	0	1	0	0	1	0	1	0
S06	1	0	1	0	0	1	0	0	1	0	1	0
S07	1	0	0	0	1	0	1	0	0	1	1	0
S08	1	0	0	0	1	0	1	0	0	1	0	1
S09	0	1	0	0	1	0	0	1	0	1	0	1
S10	0	1	0	1	0	0	1	0	0	1	0	1
S11	1	0	0	1	0	0	1	0	1	0	1	0
S12	0	1	1	0	0	1	0	0	0	1	1	0
S13	0	1	0	1	0	0	0	1	1	0	0	1

The observation of Table 14 revealed surprisingly that schemas 7 and 8 have the same interval values for the metrics (independent variables) but different intervals for time (dependent variable). In this case, it is usual to eliminate one of the two data points. Schema 8 was a good candidate for elimination as it involves a very complex terminology that could have biased the time.

4.3.2 Similarity study (CBR)

As described in Sect. 3.3.2, the goal is to find the nearest neighbour of each schema and see if it has a similar time interval. Using the distance measure defined in Sect. 3.3.2, we obtained the nearest neighbour for each schema, as shown in Table 15. These values can be obtained from the data shown in Table 5. As the distance between two schemas depends on their structural characteristics and not on time, the distances are the same for the experiment and for the replication. Moreover, as the discretization algorithm results in the same time intervals, the analysis is the same for both studies.

By observing Tables 14 and 15, it is easy to notice that, except for schemas 8 (with 11) and 12 (with 10), the nearest neighbours also belong to the same time interval. However, these two exceptions can be explained by the relative large distances between the schema

Table 15 CBR results

Schema	Nearest	Distance
S01	S05	0.000
S02	S13	0.017
S03	S06	0.006
S04	S11	0.000
S05	S01	0.000
S06	S03	0.006
S08	S11	0.069
S09	S13	0.063
S10	S13	0.019
S11	S04	0.000
S12	S10	0.102
S13	S02	0.017

structures (0.069 and 0.102, the two largest distances in Table 15). In general, in a similarity-based reasoning, a threshold is defined, beyond which no decision can be produced. This usually occurs for small case sets.

In conclusion, considering the small number of cases (schemas), the results are very interesting in the sense that, if two schemas are structurally similar enough (in our experience, distance less than 0.069), they share the same level of understandability. It is clear, however, that more cases are required in order to draw a final conclusion.

4.3.3 Metric impact study (FCA)

Formal Concept Analysis provides all the possible groups of schemas using their metrics. These groups are organized in a concept lattice. Each node contains two sets I and E, and can be read as the group of schemas E share the same set of metric values I. As with the previous technique, the lattice, shown in Fig. 7, is built using data from Table 14. Consequently, the results of the FCA analysis are the same for the initial and for the replicated experiments.

Table 16 shows the rules derived from the lattice of Fig. 7. These rules can be useful for identifying the cognitive complexity of the schemas. Each node in the lattice, where T1 or T2 is a classification factor, is considered as a rule. A rule, for a node, means that the schemas in E with time T (T1 or T2) share the same metric values in I-{T}. For example, in node 12, schemas that have low NDT (NDT1) and NFK (NFK1) have low average time (T1). These rules can be used for classifying a new schema. Rule 25, for example, specifies that a new star schema with two fact tables (NFT2) and nine foreign keys (NFK3) may require a long time to understand (T2).

Furthermore, the rules can be processed to generate a smaller rule set. For example, as rule 19 is already subsumed by rule 12, it can be removed. It is also possible to combine rules. For example, nodes 12 and 9 can prodice the following rule: if a schema has a low value for NDT and for NFK (NDT1 & NFK1) or if it has a low value for NFT and for NMFT (NFT1 & NMFT1), then maintainers are likely to need a short time for performing the tasks (T1).

Finally, when looking at Table 16, in overall terms, two general trends appear: lower values for the metrics lead to low times in accomplishing the tasks. Conversely, the higher

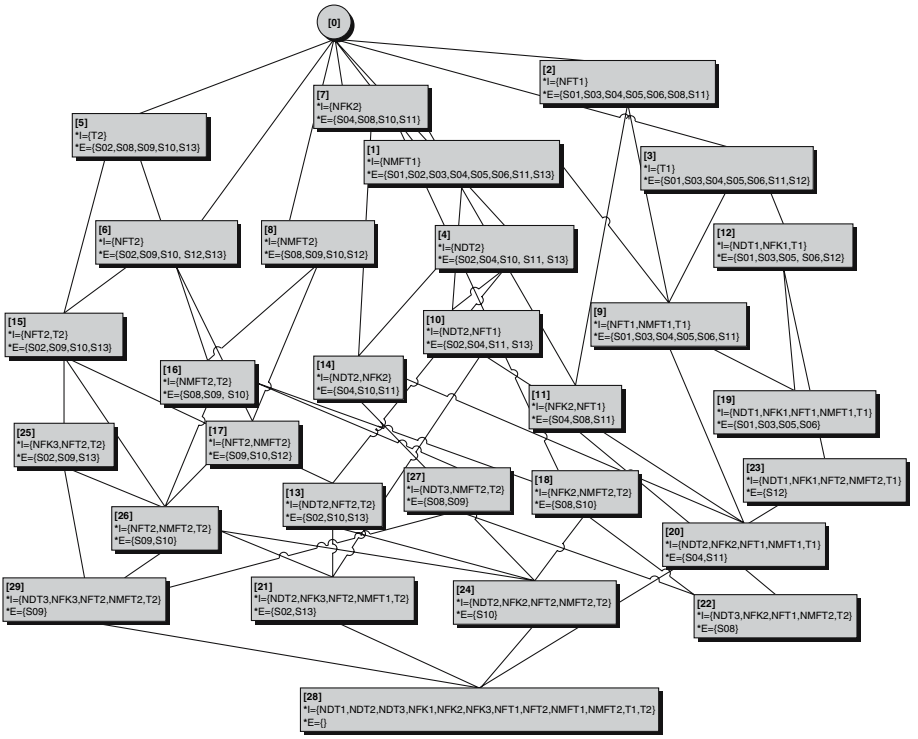

Fig. 7 Lattice for the experiment data

Table 16 Classification rules obtained from the lattice

Node	Metrics	Schemas	Time
12	NDT1 & NFK1	1, 3, 5, 6, 12	T1
9	NFT1 & NMFT1	1, 3, 4, 5, 6, 11	T1
19	NDT1 & NFK1 & NFT1 & NMFT1	1, 3, 5, 6	T1
23	NDT1 & NFK1 & NFT2 & NMFT2	12	T1
20	NDT2 & NFK2 & NFT1 & NMFT1	4,11	T1
15	NFT2	2, 9, 10, 13	T2
25	NFK3 & NFT2	2, 9, 13	T2
16	NMFT2	8, 9, 10	T2
26	NFT2 & NMFT2	9, 10	T2
13	NDT2 & NFT2	2, 10, 13	T2
27	NDT3 & NMFT2	8, 9	T2
18	NFK2 & NMFT2	8, 10	T2
29	NDT3 & NFK3 & NFT2 & NMFT2	9	T2
21	NDT2 & NFK3 & NFT2 & NMFT1	2, 13	T2
24	NDT2 & NFK2 & NFT2 & NMFT2	10	T2
22	NDT3 & NFK2 & NFT1 & NMFT2	8	T2

the metrics values, the higher the time. These conclusions allow us to claim, reasonably, that the four metrics are good indicators of the understandability of the schemas. More precisely, in many case, only one or two aspects (metrics) are enough to cover a large set of schemas (see for example nodes 9 and 15 for NFT).

4.3.4 Metric impact degree study (BC)

A BC defines the probability that a case (schema) belongs to a category (time) knowing the values of its attributes (structural metrics). Using Bayes' theorem, the probabilities are computed, starting from individual conditional probabilities of metrics (see Sect. 3.3.4). Metrics with high probabilities have an important impact on the decision.

Tables 17 and 18 show the individual conditional probabilities for the initial and replicated studies according to the two categories of time. In Table 17, the probability that, for example, NFT is equal to 1, knowing that the time is low (T1), is 0.679. This probability is used to calculate the probability of having a low time (T1) knowing that NFT is equal to 1, according to the formula provided in Sect. 3.3.4.

The first conclusion is that the two studies produced very close results. To be specific, NFT has the greatest impact when deciding for high understandability, i.e., low time T1 (P(NFT = 1|T1) = 0.7, compared to all the probabilities $P(M_i = v_j|T1)$), and for low understandability, i.e., high time T2 (P(NFT = 2|T2) = 0.68, compared to all the probabilities $P(M_i = v_j|T2)$). NMFT has an interesting impact on the decision for high understandability (P(NMFT = 2|T1) = 0.42). NDT and NFK have a reasonable impact for both decisions T1 and T2. Indeed, the corresponding conditional probabilities decrease with the values when deciding for T1, and increase when deciding for T2.

4.3.5 Conclusions of the whole study

Both initial and replication experiments confirmed our hypotheses. The correlation study revealed that the four structural metrics were correlated with understandability time. Using multivariate regression study, we showed that the relation between schema structure and understandability can be approximated by linear function. With CBR analysis, it appears that, even if a prediction function cannot be established, a similarity-based prediction can

Table 17 ROC Results for the First Experiment.

NFT	1	2					
T1	0.679	0.321					
T2	0.375	0.625					
NDT	2	3	4	5	7	8	
T1	0.226	0.298	0.226	0.083	0.083	0.083	
T2	0.097	0.097	0.181	0.264	0.181	0.181	
NFK	2	3	4	7	8	9	12
T1	0.286	0.214	0.214	0.071	0.071	0.071	0.071
T2	0.083	0.083	0.083	0.250	0.167	0.167	0.167
NMFT	2	3	4	5	12		
T1	0.457	0.171	0.100	0.100	0.171		
T2	0.117	0.283	0.200	0.283	0.117		

Table 18 ROC results for the replication

NFT	1	2					
T1	0.700	0.300					
T2	0.318	0.682					
NDT	2	3	4	5	7	8	
T1	0.211	0.278	0.211	0.078	0.144	0.078	
T2	0.106	0.106	0.197	0.288	0.106	0.197	
NFK	2	3	4	7	8	9	12
T1	0.267	0.200	0.200	0.133	0.067	0.067	0.067
T2	0.091	0.091	0.091	0.182	0.182	0.182	0.182
NMFT	2	3	4	5	12		
T1	0.427	0.160	0.093	0.160	0.160		
T2	0.127	0.309	0.218	0.218	0.127		

be conducted. Regarding the influence of each metric in the prediction, FCA-based analysis determined which levels of metrics values are better predictors of each level of understandability. More generally, we found that low values for schema metrics are indicators of low values for understanding time and, conversely, high values are synonyms for low understandability. Finally, with a BC, we established that some metrics have more impact on prediction. More concretely, the number of fact tables is by far the most important factor for prediction. In the next section we discuss threats to the validity of the experiment and the lessons learned from our study.

5 Discussion

5.1 Validity of results

As usual, different threats can affect the validity of the results of an experiment. In this section, following the framework proposed in (Wohlin et al. 2000), we discuss some threats that affect construct, internal, external and conclusion validity.

Construct validity: Construct validity is concerned with the relationship between theory and observation. We proposed, as a reasonable measure of understandability, the time for executing a given task. It is important to notice that part of the time recorded was used to answer the test and the other part was dedicated to the analysis and understanding of the data warehouse schema. As we designed equally complex tasks, we assume that the time spent answering the question was similar in all the tasks, and that the variation was due to the time spent in analysing and understanding the schema.

To better ensure construct validity, more experiments would need to be performed, varying the tasks to be carried out. Another possibility is to consider only the time of analysis without including the answering time. This can be done by the subjects themselves or using an eye tracking system.

Internal validity: Internal validity is the degree to which conclusions can be drawn about the causal effect of independent variables on dependent variables. A lack of internal validity could lead to results that are not derived from causal relationships. Regarding internal validity, we considered the following issues carefully:

- *Differences between subjects:* The used within-subject experiments reduce the variability among subjects.
- *Differences between schemas:* The domains of the schemas were different. On one hand, this reduces the dependence of time on domain knowledge. On the other hand, we are aware that there is a risk that some understandability can be impacted by domain knowledge. This was the case for schema 7, which was removed (see Sect. 4.3.1).
- *Differences between tasks:* All tasks were similar for all schemas. This way of processing helps minimizing time differences due to task complexity.
- *Precision in time values:* Subjects were responsible for recording the start and finish times of each test. We believe this method is more effective than having a supervisor who records the time for each subject. Nevertheless, there is a possibility that a subject could introduce some imprecision. For this reason, we eliminated the time outliers using the box plot technique.
- *Problems with the language:* The first experiment involved subjects from Spain and Canada who did not speak English as a native language. This limitation was the source of some problems with language understanding. In the replication, all the schemas and tasks were in Spanish, as all the subjects were Spaniards. However, we found no significant differences between both studies.
- *Learning effects:* Tasks were assigned in a different order to the subjects; the objective being to prevent learning effects.
- *Fatigue effects:* The average time for completing the experiment was around one hour. Considering the nature of the tasks, it is reasonable to consider that the fatigue effect is minimal. Even if it exists, the variation of the order of the tasks helps circumvent this effect.
- *Persistence effects:* In our case, persistence effects were not present because the subjects had never participated in a similar experiment. The subjects of the replication studies were different from those of the first study.
- *Subject motivation:* Subjects were volunteers and were convinced that their contribution was very important for research in the field of data warehouse metrics development. The experiment was conducted on a volunteer basis and was not part of the students' formal assignments. We can reasonably claim that the subjects were motivated enough. Moreover, subjects with many outliers were eliminated as an a posteriori action to prevent the absence of motivation.
- *Other factors:* Plagiarism and subjects' influence on each other were controlled. They were informed that they should not talk to or share answers with other subjects.

External validity: External validity is the degree to which the results of the research can be generalized to the population under study and to other research settings. The greater the external validity, the more the results of an empirical study can be generalized to actual software engineering practice. If external validity is not assured, the empirical results cannot be generalized to the population. Regarding external validity, the following issues were considered:

- *Materials and tasks used:* We tried to use schemas and operations in the experiment with enough variation to cover a spectrum representative of real cases, although more experiments with larger and more complex schemas are necessary.
- *Subjects:* Due to the difficulty of getting those who are already working in the profession to participate in the experiments, these experiments were conducted using

students and teachers. Nevertheless, many authors agree that, for many phenomena, using students has little impact on the validity of a study (Basili et al. 1999; Hörst et al. 2000; Carver et al. 2003). In our particular case, the tasks performed do not require a high level of industrial experience and can easily be carried out by students. However, to ensure the external validity of our study, more experiments with industrial subjects are necessary.

Conclusion validity: Conclusion validity defines the extent to which conclusions are statistically valid. The only issue that could affect the statistical validity of this study is the size of the sample data (13 objects). However, as explained above, we designed the experiment in such a way as to get around this limitation. We are currently working on the collection of a larger data set, to conduct a replication study. This data collection is a long and complex task, owing first of all to the nature of the experiment objects (data warehouses) and, secondly, to the used sampling method (purposing sampling).

5.2 Lessons learned

During the development, execution and analysis phases of this study, we faced several problems that could have had an impact on the validity of our experiment. Solving/circumventing these problems was an opportunity for us to learn more about empirical investigation and thus improve subsequent empirical studies.

The first challenge was to build large representative data samples. Published material usually contains only one small example, and companies are generally not willing to share their designs with others (as data warehouse schemas are usually considered strategic information).

Another problem was the impact of the language on the behaviour of the subjects (time in answering the questions). Indeed, the experimental package was presented in English. The subjects were from Ciudad Real (Spain) and Montreal (Canada) with their native/working languages being Spanish and French, respectively. As a consequence, we found some errors or abnormal response times that were explained by misunderstandings regarding certain words, such as "Rent". Moreover, some of the responses were provided in Spanish or French. For the replication of the experiment, this problem was eliminated, as we translated the experimental material into Spanish and all the subjects were Spaniards; accordingly, language problems did not arise during the replication.

The third problem was the level of detail in the answers. As the only constraint was deciding which information had to be recovered from which table in order to obtain a specific result, we found significant differences in the level of detail presented by subjects. For the replication, the subjects took a mini-tutorial on how to perform the tasks with an example showing the appropriate level of detail for the answers. For future replications, a more objective approach to answering the questions can be considered in order to avoid this possible bias.

After exploring the possible threats, we can reasonably claim that this study confirmed our hypotheses. Overall, the analyses have determined the nature of the relationship between the proposed metrics and the understandability (and to some extent, the cognitive complexity) of data warehouse schemas, and have provided useful information regarding this relationship. It is important to note, however, that more data is required in order to determine a precise definition of the prediction function.

6 Conclusions

One of the primary obligations of IT professionals must be to assure the quality of information, as this is one of the primary assets of an organization. Quality considerations have accompanied multidimensional data model research from the beginning (Jarke et al. 2000). Although some interesting guidelines have been proposed for designing "good" multidimensional models, more objective indicators are still needed. Our work aims specifically to produce a set of metrics that are valid for measuring the quality of data warehouse schemas. These can help designers in their daily work, e.g., when choosing among different design alternatives that are semantically equivalent.

In this article, we have proposed and validated a set of metrics, through an empirical study. Although this study was subject to many threats to validity (see Sect. 5), we found that the four metrics presented (NFT, NDT, NFK and NMFT) are good indicators of data warehouse quality (cognitive complexity). However, we are aware that more experiments are needed to draw a final conclusion.

Acknowledgements This research is part of the CALIPO project, supported by Dirección General de Investigación of the Ministerio de Ciencia y Tecnologia (TIC2003-07804-C05-03). This research is also part of the ENIGMAS project, supported by Junta de Comunidades de Castilla – La Mancha – Consejería de Ciencia y Tecnología (PBI-05-058). This work was performed during the stay of Houari Sahraoui at the University of Castilla-La Mancha under the "Programa Nacional De Ayudas Para La Movilidad de Profesores en Régimen de año sabático", from Spanish Ministerio de Educación y Ciencia, REF: 2004-0161. We would like to thank all of the volunteer subjects who participated in these experiments whose inestimable assistance helped us reach the conclusions in this paper. We also want to thank the reviewers for their valuable comments.

Appendix: Collected time

Appendix: Collected time

Subj.	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13
Collected data from the initial experiment													
1	485	1259	1050	993	933	608	978	811	1992	889	515	895	1224
2	555	1656	550	280	708	297	859	–	1046	536	500	–	–
3	240	540	420	300	300	180	60	360	420	660	360	240	240
4	216	384	218	316	328	193	380	521	586	608	415	393	306
5	862	1779	468	479	496	485	498	489	903	679	834	690	712
6	622	1706	476	494	585	531	509	465	935	520	935	662	625
7	735	2130	936	538	365	427	817	1227	1755	1155	570	910	782
8	75	270	95	110	55	110	175	145	420	145	95	120	145
9	116	254	203	203	65	92	205	194	410	271	186	194	389
10	165	199	280	87	74	153	97	93	200	118	109	249	268
11	100	251	255	180	75	126	147	168	495	206	164	228	121
12	240	600	480	300	180	300	480	600	900	300	300	240	600
13	226	590	167	124	166	153	141	319	173	206	205	307	392

continued

Subj.	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13
14	170	433	220	198	571	535	213	1074	920	460	462	190	270
15	75	340	72	210	83	52	157	265	450	199	148	100	185
16	97	129	62	137	56	88	61	66	215	166	71	126	194
17	120	384	134	188	74	131	147	203	226	145	272	79	355
18	115	335	149	163	84	117	77	52	258	160	108	135	252
19	109	221	142	110	65	90	340	148	195	159	153	184	241
20	140	291	136	206	138	290	359	352	444	241	243	324	287
21	75	160	100	75	110	80	120	260	250	240	160	100	180
22	115	255	38	50	50	70	130	110	196	120	145	270	308
23	360	540	600	470	405	179	107	625	540	660	300	430	300
24	420	1020	480	600	420	540	840	1200	1140	1020	840	780	840
Subj.	S01	S02	S03	S04	S05	S06	S08	S09	S10	S11	S12	S13	
Collected data from the replication													
1	360	559	196	293	199	230		353	499	244	236		
2	161	526	166	137	88	166	326	431	363	250	426	237	
3	233	199	140	317	274	244	180	277	305	142	109	284	
4	109	360	332	195	127	228	139	304	445	159	233	238	
5	238	251	117	123	169	295	214	253	294	344	163	285	
6	235	243	176	106	181	208	270	331	236	201	122	260	
10	229		150	189	169	184	210			278	187	260	
11		359	204	239		169	242		287	243	268	179	
12	105	178	158	162	126	175	182	376	312	119	195	258	
13	177	311			270	382			540		344	278	
15	198	86	90	117	363	99	107	301	163	80	130	280	
16	251		285	145	220			300			201	322	
17	125	187	161	105	188	208	120	218	190	130	159	118	
18	97	110	92	149	90	131	117	189	169	152	61	90	
19	191	94	357	122	173	249	124	270		93	107	117	
21	221	132	196	140	179	174	125	544	221	131	146	251	
24	149	366	243	328	142	206		465	227			369	
26	171	208	165	303	143	240	251	159	339		243	130	
27	245	460	388	266	190	245	324	220					
30	96	262	120	60	63	93	223	219	217	120	230	320	
33	130	190	386	100	120	130	448	398	248	110			
34	228	120	175	155	240	150	110	150	170	115	177	193	
35	91	259	135		122	250	280	366	176	158	135	286	
37	149		160	186	125		225	400	325	132	162	205	
38	150	201	153	119	127	197	115		158	128	150	192	
39	296	270	241	337	294	216		381	332				
40	117		169	132	108	129		628			201	462	
44		285	196	196	70			514		292	340	352	
45	190		278	145		109	259		456	353	152		
46			242		202			273	405	213	518	302	

continued

Subj.	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13
48	214		126		210	167	167	145		148	183	357	
49	205		295	198	190	242	218	290	258	249		180	
50		278	350	156			434		354	195	255		
52	110	272		201	167	222	318	418		218		202	
53	302	256			365			344	330		378	318	
54	144	476	354	167		139	372			218	202		
55	300	285	133			131		575	319	271		433	
56		409	217	196	172		410	469			440		
57	165	287	119	162	145	272	151	407	262	306	232	176	
58	271		296	271	201	221			232	106	317	329	
59	192	205	172	115	143	208		201	385	191	214	233	
60	202	400	158	133	143	135	204	206	361	204	211	261	
61	159	329	431		142	221	377			166		264	
62	169	276	193	223	291	289	319	442			297		
63		383	246	324	200	224	482			335		240	
64	184	182			296	282	202	243	194	386	301		
65	206	401	305	245	150	290	125	328	314				
66	143	347	181	184	225	146	236	290	277	117	242	175	
67	151	250	193	173	112	185	301	267	191	213	152	238	
68		320		405	225			300		246	204		
69	334	451	227	255	154	282				315			
70	192	131	156	92	138	146	230	420	458	174	96	153	
71			231	297			429	719			431	364	
73	100	205	165	105	106	147	121	124	179	170	167	104	
75	195			198	169	211	223		291	234	245	294	
76	204	252		165	105	153			266	139	230	197	
77		409	158			194	161		544	179	305		
78	283		120					225	194	145	169	174	
79	120	120	60	120	60	120	180	240	60	300	60	420	
80		305	245		80				535	209	336	305	
81	190			204	189	278			389	284	310	345	
82		321	158					326	399	336		488	
83		274	135	360		221	364			480		332	
85	167		209						327	268	257	192	

References

Anahory, S., & Murray, D. (1997). *Data warehousing in the real world*. Harlow, UK: Addison-Wesley.

Basili, V. R., Shull, F., & Lanubille, F. (1999). Building knowledge through families of experiments. *IEEE Transactions on Software Engineering, 25*(4), 456–473.

Bouzeghoub, M., & Kedad, Z. (2002). *Information and database quality, Chapter 8, Quality in data warehousing* (pp. 163–198). Kluwer Academic Publishers.

Briand, L., Morasca, S., & Basili, V. (1996). Property-based software engineering measurement. *IEEE Transactions on Software Engineering, 22*(1), 68–86.

Briand, L., Ikonomovski, S., Lounis, H., & Wüst, J. (1998). *A Comprehensive investigation of quality factors in object-oriented designs: An industrial case study, Technical Report ISERN-98-29*. Germany: Fraunhofer Institute for Experimental Software Engineering.

Calero, C., Piattini, M., Pascual, C., & Serrano, M. (2001). Towards Data warehouse Quality Metrics, International Workshop on Design and Management of Data Warehouses (DMDW'01).

Carver, J., Jaccheri, L., Morasca, S., & Shull, F. (2003). Issues in using students in empirical studies in software engineering education. In *Proceedings of 2003 International Symposium on software metrics (METRICS 2003)*. Sydney, Australia. September 2003, pp. 239–249.

Debevoise, N. T. (1999). *The data warehouse method*. NJ: Prentice Hall Upper Saddle River.

Fenton, N., & Pfleeger, S. (1997). *Software metrics: A rigorous approach* (2nd ed.). London: Chapman & Hall.

Flach, P., & Lachiche, N. (1999). 1BC: A First-Order Bayesian Classifier. In *Proceedings of the Ninth International Workshop on inductive logic programming (ILP'99), volume 1634 of lecture notes in artificial intelligence*, pp. 92–103.

Godin, R., Mineau, G., Missaoui, R., St-Germain, M., & Faraj, N. (1995). Applying concept formation methods to software reuse. *International Journal of Knowledge Engineering and Software Engineering, 5*(1), 119–142.

Grosser, D., Sahraoui, H. A., & Valtchev, P. (2003). An analogy-based approach for predicting design stability of Java classes. In *International Symposium on Software Metrics (METRICS'03)*, pp. 252–262.

Hörst, M., Regnell, B., & Wohlin, C. (2000). Using students as subjects – A comparative study of students & professionals in lead-time impact assessment. In *4th Conference on empirical assessment & evaluation in software engineering*, EASE, Keele University, UK.

Huang, K.-T., Lee, Y. W., & Wang, R. Y. (1999). *Quality information and knowledge*. Prentice Hall: Upper Saddle River.

Inmon, W. H. (1997). *Building the data warehouse* (2nd ed.). John Wiley and Sons.

ISO. (2001). *Software product evaluation-quality characteristics and guidelines for their use*. Geneva: ISO/IEC Standard 9126.

Jarke, M., LenzerinI, I. M., Vassilou, Y., & Vassiliadis, P. (2000). *Fundamentals of data warehouses*. Springer.

Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The data warehouse lifecycle toolkit*. John Wiley and Sons.

Kitchenham, B., Pfleegger, S., Pickard, L., Jones, P., Hoaglin, D., El-Emam, K., & Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions of Software Engineering, 28*(8), 721–734.

Poels, G., & Dedene G. (1999). *DISTANCE: A framework for software measure construction*. Belgium: Dept. Applied Economics Katholieke Universiteit Leuven.

Ramoni, M., & Sebastiani, P. (1999). Bayesian methods for intelligent data analysis. In: M. Berthold & D. J. Hand (Eds.), *An introduction to intelligent data analysis*. Springer: New York.

Schneidewind, N. (2002). Body of knowledge for software quality measurement. *IEEE Computer, 35*(2), 77–83.

Serrano, M., Calero, C., & Piattini, M. (2002). Validating metrics for data warehouses. *IEE Proceedings SOFTWARE, 149*(5), 161–166.

Serrano, M., Calero, C., & Piattini, M. (2005). An experimental replication with data warehouse metrics. *International Journal of Data Warehousing & Mining, 1*(4), 1–21.

Wilson, D., & Martinez, T. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*, 1–34.

Wohlin, C., Runeson, P., Höst, M., Ohlson, M., Regnell, B., & Wesslén, A. (2000). *Experimentation in software engineering: An introduction*. Kluwer Academic Publishers.

Zuse, H. (1998). *A framework of software measurement*. Berlin: Walter de Gruyter.

Author Biographies

Manuel Serrano is MSc and PhD in Computer Science by the University of Castilla – La Mancha. Assistant Professor at the Escuela Superior de Informática of the Castilla – La Mancha University in Ciudad Real. He is a member of the Alarcos Research Group, in the same University, specialized in Information Systems, Databases and Software Engineering. His research interests are: DataWarehouses Quality & Metrics, Software Quality. His e-mail is Manuel.Serrano@uclm.es

Coral Calero is MSc and PhD in Computer Science. Associate Professor at the Escuela Superior de Informática of the Castilla – La Mancha University in Ciudad Real. She is a member of the Alarcos Research Group, in the same University, specialized in Information Systems, Databases and Software Engineering. Her research interests are: advanced databases design, database/datawarehouse quality, web/portal quality, software metrics and empirical software engineering. She is author of articles and papers in national and international conferences on these subjects. Her e-mail is: Coral.Calero@uclm.es

Houari Sahraoui received a Ph.D. in Computer Science from Pierre Marie Curie University, Paris in 1995. He is currently an associate professor at the Department of Computer Science and Operational Research, University of Montreal where he is leading the software engineering group (GEODES). His research interests include object-oriented software quality, software visualization and software reverse and re-engineering. He has published more than 80 papers in conferences, workshops and journals and edited two books. He has served as program committee member in several major conferences and as member of the editorial boards of two journals. He was the general chair of IEEE Automated Software Engineering Conference in 2003. His e-mail is sahraouh@iro.umontreal.ca

Mario Piattini is MSc and PhD in Computer Science by the Polytechnic University of Madrid. Certified Information System Auditor by ISACA (Information System Audit and Control Association). Full Professor at the Escuela Superior de Informática of the Castilla – La Mancha University. Author of several books and papers on databases, software engineering and information systems. He leads the ALARCOS research group of the Department of Computer Science at the University of Castilla – La Mancha, in Ciudad Real, Spain. His research interests are: advanced database design, database quality, software metrics, object oriented metrics, software maintenance. His e-mail address is Mario.Piattini@uclm.es