Defining and validating metrics for assessing the understandability of entity–relationship diagrams

Marcela Genero [a,*], Geert Poels [b], Mario Piattini [a]

[a] *ALARCOS Research Group, Department of Information Systems and Technologies, University of Castilla-La Mancha, Paseo de la Universidad, 4 – 13071 Ciudad Real, Spain*
[b] *Management Informatics Research Unit, Faculty of Economics and Business Administration, Ghent University – UGent, Tweekerkenstraat 2, 9000 Ghent, Belgium*

Abstract

Database and data model evolution cause significant problems in the highly dynamic business environment that we experience these days. To support the rapidly changing data requirements of agile companies, conceptual data models, which constitute the foundation of database design, should be sufficiently flexible to be able to incorporate changes easily and smoothly. In order to understand what factors drive the maintainability of conceptual data models and to improve conceptual modelling processes, we need to be able to assess conceptual data model properties and qualities in an objective and cost-efficient manner. The scarcity of early available and thoroughly validated maintainability measurement instruments motivated us to define a set of metrics for Entity–Relationship (ER) diagrams. In this paper we show that these easily calculated and objective metrics, measuring structural properties of ER diagrams, can be used as indicators of the understandability of the diagrams. Understandability is a key factor in determining maintainability as model modifications must be preceded by a thorough understanding of the model. The validation of the metrics as early understandability indicators opens up the way for an in-depth study of how structural properties determine conceptual data model understandability. It also allows building maintenance-related prediction models that can be used in conceptual data modelling practice.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Conceptual data modelling; ER diagram; Understandability; Structural properties; Metrics; Measurement theory; Experimental validation

1. Introduction

In the highly dynamic business environment we are in these days, existing business models become obsolete at an ever-increasing pace and must therefore be designed with flexibility in mind to satisfy the needs of agile companies. The constant reshaping of business models increases the volatility of the data requirements that must be supported by companies' data resource management policies and technologies. The ability to rapidly

change databases and their underlying data models to support the needs of changing business models is a main concern of today's information managers. Database evolution produces significant challenges, high on the research agenda of information system researchers.

It has been observed recently that conceptual data model quality, which is postulated as a major determinant of the efficiency and effectiveness of data model evolution, is a main topic in current conceptual modelling research [52]. Empirical studies point out that the quality of conceptual models affects the quality of the system that is finally implemented [25]. Researchers who realize the importance of good conceptual models have proposed quality frameworks to define the field and advance the discipline [51]. However, despite the proliferation of conceptual modeling quality frameworks, generally agreed quality measures still have to be developed [49].

Model maintainability, i.e. the ability to easily change a model [38], seems to be a key factor in conceptual data model evolution. In order to evaluate and, if necessary, improve the maintainability of conceptual data models, data analysts need instruments to assess the maintainability characteristics of the models they are producing. The earlier this assessment can be done, the better, as it has proven much more economical to evaluate and improve quality aspects during the development process than afterwards [8].

Maintainability is, however, an external quality property meaning that it can only be assessed with respect to some operating environment or context of use [28]. The maintainability of a conceptual data model depends on its understandability (also called comprehensibility [31]) because the model must be understood first before any desired changes to it can be identified, designed and implemented.[1] Model understanding depends on model properties such as structure, clarity and self-expressiveness that determine the ease of understanding, but also on the data analyst's familiarity with the model. In large organizations, the data analysts or data model administrators that are responsible for maintaining the models may be different from the ones having originally developed them. It is therefore in the interest of the organization to closely monitor the understandability of its evolving set of conceptual data models.

External qualities such as understandability and maintainability are hard to measure objectively early on in the modelling process. They generally need to be assessed in a subjective way, for instance using expert opinions expressed through a formal or informal scoring system. For a more objective (and more cost-efficient) assessment of external quality attributes, an indirect measurement based on internal model properties is required [28]. If a significant relationship with conceptual data model understandability can be demonstrated, then metrics of internal model properties can be used as indicators of sufficient or insufficient understandability. Once constructed and validated, a measurement-fed prediction model can be implemented and employed at a relatively low cost compared to an a-posteriori understandability assessment.

From a practical perspective, the availability of early applicable understandability metrics would allow data analysts to perform

- A quantitative comparison of design alternatives, and therefore an objective selection between several conceptual data model alternatives.
- An early assessment of conceptual data model understandability, even during the modelling activity, and therefore a better resource allocation based on this assessment (e.g. redesigning high-risk models with respect to understandability).

In this paper we define a set of 12 metrics for measuring structural properties of ER diagrams. Our focus on ER modelling is justified by the observation that in today's database design world, it is still the dominant method of conceptual data modelling [15,50]. Our focus on structural properties is motivated by similar research in the field of empirical software engineering, where properties that determine how software is structured like coupling, cohesion, and inheritance structures, have been shown to be major determinants of external software quality properties, including understandability and maintainability [2,9,29,36,42,51]. From this body of literature, a research model was derived that postulates a relationship between the structural properties of an ER diagram and the understandability of the ER diagram. In this model, the impact of the structural

---

[1] Even though understandability has not been considered as a maintainability sub-characteristic by the ISO 9126 standard [38], software quality research considers understandability to be a main factor influencing maintainability [9,28,36].

properties on understandability is mediated through the cognitive complexity experienced by the person that needs to understand the diagram.

The focus on structural properties is further motivated by our search for metrics that do not require human judgement for calculating them. As syntactic model properties, the measurement of ER diagram structural properties can be automated, which assures the cost-efficiency, the consistency and repeatability of the (indirect) understandability measurements.

Apart from defining the metrics, the paper pays extensive attention to their validation. For the validation of the metrics as ER diagram understandability indicators an experiment was conducted in which the relationship between the metrics values and direct understandability assessments was statistically tested. Out of the 12 proposed metrics, six metrics were investigated in this experiment. For three metrics a statistical significant correlation with understandability measurements was found. The results of the experiment indicate that the understandability of an ER diagram is affected by the structural complexity that is contributed by the diagram's attributes and relationships, in particular the 1:1 and 1:N relationships. The more attributes and relationships, especially 1:1 and 1:N relationships, a diagram contains, the less understandable it is.

This paper is structured as follows. In section 2 we discuss related work in the field of metrics-based conceptual data model quality assurance. After proposing our research model and metrics suite in section 3, we proceed with the empirical validation of the metrics as ER diagram understandability indicators in section 4. Finally, concluding remarks and an outline of our future work is presented in section 5.

2. Related work

A number of quality assurance frameworks for conceptual data models have been proposed [49], amongst them the frameworks of Lindland et al. [43], Krogstie et al. [41], Moody et al. [48], Schütte and Rotthowe [59], and Cherfi et al. [62]. Most of these frameworks provide quality definitions and criteria for conceptual models; few of them, however, include quantitative measures (i.e. metrics) to evaluate the quality of conceptual data models in an objective and cost-efficient way [49]. In this section we review the literature searching for metrics that have been proposed for evaluating the quality of conceptual data models.

Eick [20] proposed a single quality metric for use with S-diagrams[2] [18,19]. Three quality aspects, expressiveness, complexity and normality, are meant to be captured by this metric. To our knowledge there is no published work confirming the validity of Eick's metric.

Gray et al. [34] proposed a suite of metrics for evaluating quality characteristics of ER diagrams (complexity and deviation from third normal form). These authors commented that empirical validation of these metrics has been performed, but they do not provide the results. Independently, Ince and Shepperd [37] used the algebraic specification language OBJ to demonstrate the correctness of the underlying syntax of the metrics. While useful for verifying the precise definition of the metrics, the study does not validate the metrics as being related to the quality of ER diagrams.

Kesh [39] proposed a single metric for the quality of ER diagrams, combining different metrics of the ontological and behavioural quality of ER diagrams. In most cases these metrics are Likert scales that need to be rated in a subjective way. Kesh's proposal also requires that, to obtain an overall quality assessment, each measurement for each ontological quality factor to be weighted, but he did not suggest how to determine the weights. No empirical evidence on the validity of the Kesh's metric has been collected.

Moody et al. [48] proposed a data model quality management framework for evaluating the quality of ER diagrams. This framework includes a set of metrics [47] which capture different quality characteristics of ER diagrams. Some of the metrics are objectively calculated while others are based on subjective expert ratings. In [47] an action research study is reported that used the framework to improve the quality of the data models (product quality) and the process of developing data models (process quality) in a large Australian bank. Based on this study, it was concluded that only four out of the 25 proposed metrics have benefits that outweigh their cost of collection. Among these four metrics, there is only one product metric, which is the number of entities and relationships in an ER diagram. The other three metrics are estimates (development cost esti-

---

[2] The S-diagram is a data model which was influenced by the work on the binary relation model [1] and the Semantic Database Model (SDM) [35].

Table 1
Summary of quality metrics for conceptual data models

Authors	Quality focus	Scope	Objective/ subjective	Validation?	Tool?
Eick [20]	Expressiveness, complexity, normalizedness	S-diagrams	Objective	No	No
Gray [34]	Complexity, deviation from third normal form	ER diagram	Objective	Unknown	Yes
Kesh [39]	Ontological quality, behavioural quality	ER diagram	Objective and subjective	No	Yes
Moody [46]	Completeness, integrity, flexibility, understandability, correctness, simplicity, integration, implementability	ER diagram	Objective and subjective	Number of entities and relationships metric validated in an action research study	Yes
Cherfi et al. [62]	*Specification dimension:* legibility, expressiveness, simplicity, correctness *Usage dimension:* completeness, understandability *Implementation dimension:* implementability, maintainability	Extended ER diagram	Objective and subjective	No	Yes
Maes and Poels [44]	Perceived semantic quality, ease of use, usefulness, user satisfaction	Graphical conceptual models	Subjective	Yes (via two controlled experiments)	No

mate) and process metrics useful for monitoring data model quality over time (the percentage of reuse and the number of defects classified by quality factor).

Cherfi et al. [62] defined a framework considering three dimensions of quality: usage, specification and implementation. For each dimension, quality criteria and corresponding metrics (not all of them are objective) were defined. An example was shown to illustrate the application of the framework, but no empirical study was carried out for demonstrating its validity.

Maes and Poels [44] proposed an instrument to measure the user perception of a conceptual model's semantic quality, ease of use and usefulness, and to measure how satisfied users are with the model. Two experiments were conducted to validate the measures and to develop an underlying model relating the different measured variables. Because they are perception-based, the measures are not objective and cannot be automatically calculated.

Table 1 compares the current proposals of conceptual data model quality metrics. The first column of the table contains a reference to the study where the proposal has been published. In the second column, the quality focus of the metrics (i.e. the quality properties that are measured) is presented. The third column refers to the scope of the metrics, meaning the kind of data model that is measured. The fourth column shows whether the metrics are objective (e.g. measures) or subjective (e.g. quality scores or rating assigned by 'expert judges'). The fifth column questions whether the metrics have been validated. The last column reflects whether an automated tool exists for the metric calculation.

Summarising the related work, we can conclude that apart from the number of entities and relationships metric of Moody [46], the metrics for ER diagrams found in the literature have not been validated. In other words, their relationship with the external quality of ER diagrams has not been demonstrated.

3. Proposal of metrics for ER diagrams

We first present the research framework that justifies our choice of metrics and validation approach. Next the metrics suite is defined, both informally and formally (using measurement theory). In a last sub-section the calculation of the metrics is illustrated using an example ER diagram.

3.1. Research framework

The framework for our research is based on similar research in empirical software engineering, where metrics-based prediction models have been proposed for external software qualities such as (lack of)

fault-proneness and maintainability [10]. Although this stream of research is highly explorative in nature, consensus is emerging regarding the role that structural software properties such as coupling, cohesion, size, and inheritance structures play in determining software quality. Moreover, these properties already emerge at the software design stage, allowing software quality to be controlled in an early phase of the software life-cycle.

Reasons why structural properties have an impact on quality have been suggested by Briand et al. [11,12] (see Fig. 1). According to Briand et al., software that is big (i.e. having many composing elements) and has a complex structure (i.e. showing many relationships between the software's composing elements) results in a high cognitive complexity, which is defined as the mental burden of the people that perform tasks on the software. It is the high cognitive complexity that causes the software to display undesirable properties such as being fault-prone or requiring much effort to maintain, simply because it is more difficult to understand, develop, modify or test such software.

Briand et al.'s model, which they refer to as a 'causal chain', has been the basis for much of the recent research on empirically-derived metrics-based software quality prediction [23,33,45,57]. In addition, this model was also used by Erickson and Siau [26] to investigate the complexity of UML [53]. Although there have been attempts to provide a theoretical basis to the model, using cognitive models that consider the information processing limitations of human memory [21,22], the model has not been directly tested due to the difficulty involved in measuring cognitive complexity. Nevertheless, the (indirect) relationship between software's structural properties and external quality properties has been repeatedly demonstrated. According to Briand et al. [12], it is difficult to imagine what could be alternative explanations for these results besides cognitive complexity mediating the effect of structural properties on software quality.

We will use Briand et al.'s model as a working hypothesis and framework for our research, and thus consider an ER diagram as a software artefact. Given that people have to deal with conceptual data models (e.g. data analysts developing them, future system users validating them, database developers using them to design databases, etc.), cognitive complexity is at stake. Hence, if it is assumed that the structural properties of an ER diagram affect its understandability (via cognitive complexity), then metrics that measure these structural properties can be used as early and low-cost understandability indicators. So what is needed is a metrics suite that captures to the best possible extent the ER diagram's structural properties, and secondly a validation of these metrics as understandability indicators.

3.2. Metric definition

According to our research framework, an ER diagram's structural properties affect the cognitive complexity that is associated with the diagram. Cognitive complexity is, however, difficult to measure. To distinguish with *cognitive complexity*, we will refer to the collection of structural properties as *structural complexity*, which is a measurable kind of complexity. According to Systems Theory, the complexity of a system is based on the number of (different types of) elements and on the number of (different types of) (dynamically changing) relationships between them [54]. Hence, the structural complexity of an ER diagram is determined by the different elements that compose it.

Fenton has formally proven that a single metric of complexity cannot capture all possible aspects or viewpoints on complexity [27]. Therefore, it is not advisable to define a single metric for ER diagram structural complexity (like Moody's number of entities and relationships metric [46]). The approach that we take is to propose several metrics, each one focussing on the use of a different ER modelling construct in the ER diagram. This approach also allows investigating which of the ER diagram's structural properties have an impact on understandability, and which do not.
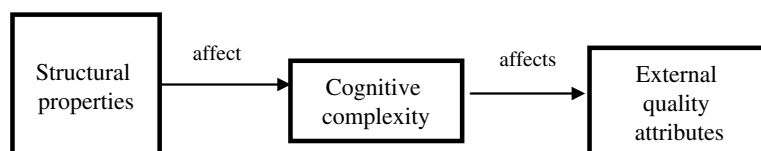

Fig. 1. Relationship between structural properties, cognitive complexity, and external quality attributes [11,12].

The ER diagrams for which the structural complexity metrics are proposed, build upon the constructs offered in the original ER model [14], plus a few additional constructs such as IS-A relationships. Instead of presenting a well-defined meta-model, we only list the modelling constructs (see Table 2). Precise definitions of their semantics can be found in classic textbooks on ER modelling [24]. Although the constructs listed are the ones that are frequently encountered in ER diagrams, other constructs can be added to the list. Likewise, the metrics suite presented here can be extended to cover the use of such other constructs.

Note also that, although we do not use the word 'type' (mainly in order to shorten the names of the metrics proposed in the next sub-section), all constructs are considered at the type level. The information contained within an ER diagram does not allow the measurement of structural complexity at the instance level.

The metrics included in the suite are shown in Table 3. As can be seen these 'count' metrics capture the usage of the constructs listed in Table 2.

In Table 3 the metrics are informally defined, in natural language. For each metric we have also formulated a formal definition based on measurement theory [40,56], allowing a precise and unambiguous interpretation of what is measured and how this is done. Measurement theory is a normative theory prescribing the conditions that must be satisfied in order to use mathematical functions as 'metrics'. Measurement theoretic approaches to metrics definition propose methods to verify whether these conditions hold.

The approach used to formally define the ER diagram metrics of Table 3 is the DISTANCE approach [56], which uses the mathematical concept of 'distance' and its measurement theoretic interpretation as main cornerstones. The basic idea of DISTANCE is to define properties of objects in terms of distances between the objects and other objects that serve as reference points (or norms) for measurement. The larger these distances, the greater the extent to which the objects are characterised by the properties. This particular definition of object properties allows them to be measured by functions that are called 'metrics' in mathematics. Metrics are functions that satisfy the metric axioms, i.e. the set of axioms that are necessary and sufficient to define distance metrics (in the sense of measurement theory) [63].

The application of DISTANCE to the ER diagram metrics of Table 3 is straightforward. Each metric quantifies the extent to which a particular ER modelling construct is used in an ER diagram. The NA metric, for instance, captures the usage of the attribute construct in an ER diagram. For the purpose of measuring the NA metric of an ER diagram, the diagram can be abstracted into its set of attributes. The larger this set (i.e. the bigger the distance between this set and the empty set), the greater the structural complexity of the ER diagram due to the usage of attributes. In [55], the symmetric difference model, i.e. a particular instance of Tversky's set theoretic contrast model [63], was used to show that when properties are represented as sets, the cardinality of the set qualifies as a metric (in both the mathematical and measurement theoretic sense). Hence, the NA metric, which returns the cardinality of the set of attributes defined in an ER diagram, measures the distance between an ER diagram's set of attributes and the empty set (which is used as the reference point for measurement). In other words, the NA metric measures the structural complexity of an ER diagram that is contributed by the usage of the attribute construct.

Table 2
ER modelling constructs measured by the metrics suite

ER modelling construct			Remarks
Entities	Strong	Weak	
Attributes	Simple	Composite	
	Non-derived	Derived	
	Single-valued	Multi-valued	
Relationships	'common'	IS-A	'common' relationships, sometimes also called 'associations', are the ones considered in Chen's original ER model; IS-A relationships are generalization/specialization relationships as considered in most extensions of the ER model.
	Reflexive	Binary	n-ary	Kinds of 'common' relationships according to the number of entity types involved.
	1:1	1:N	M:N	Kinds of 'common' reflexive or binary relationships according to the connectivities (maximum multiplicities) involved.

Table 3
Metrics for ER diagrams

Metric name	Metric description
NE	The Number of Entities metric is defined as the number of entities within an ER diagram, considering both weak and strong entities.
NA	The Number of Attributes metric is defined as the total number of attributes defined within an ER diagram, taking into account not only entity attributes but also relationship attributes. In this number all attributes are included (but not the composing parts of composite attributes).
	By convention, the NA metric does not count the attributes that a sub-type inherits from its super-type (i.e. these attributes are counted only once, as attributes of the super-type).
NDA	The Number of Derived Attributes metric is defined as the number of derived attributes within an ER diagram. The value of NDA is always strictly less than the value of NA.
NCA	The Number of Composite Attributes metric is defined as the number of composite attributes within an ER diagram. This value is less than or equal to the NA value.
NMVA	The Number of Multi-valued Attributes metric is defined as the number of multi-valued attributes within an ER diagram. Again, this value is less than or equal to the NA value.
NR	The Number of Relationships metric is defined as the total number of relationships within an ER diagram, excluding IS-A relationships.
NM:NR	The Number of M:N Relationships metric is defined as the number of M:N relationships within an ER diagram. The value of NM:NR is less than or equal to the NR value.
N1:NR.	The Number of 1:N Relationships metric is defined as the total number of 1:N and 1:1 relationships within an ER diagram[a]. Also this value is less than or equal to the NR value.
NN_AryR	The Number of N-Ary Relationships metric is defined as the number of N-Ary relationships within an ER diagram. Its value is less than or equal to the NR value.
NBinaryR	The Number of Binary Relationships metric is defined as the number of binary relationships within an ER diagram. Again, the value is less than or equal to the NR value.
NRefR	The Number of Reflexive Relationships metric is defined as the number of reflexive relationships within an ER diagram. Its value is less than or equal to the value of the NR metric.
NIS_AR	The Number of IS_A Relationships metric is defined as the number of IS_A relationships within an ER diagram. In this case, we consider one relationship for each super-type/sub-type pair.

[a] The number of 1:1 relationships was not used as a separate metric because these relationships are considered a subset of the 1:N relationships.

As an example, the complete formal definition of the NA metric is presented in Appendix A. Analogous definitions for the other ER diagram metrics can be found in [32].

The measurement theoretic theorems associated with distance measurement are incorporated in the DISTANCE approach, meaning that the conditions specified by these theorems are met when defining metrics with DISTANCE. This ensures that the validity of the metrics as measures of the ER diagram structural properties is formally proven within the framework of measurement theory. An important pragmatic consequence of the explicit link with measurement theory is that the resulting metrics define ratio scales.

3.3. Illustration

As an example, we will apply the outlined metrics to the ER diagram shown in Fig. 2, taken from Elmasri and Navathe [24].

Table 4 shows the value of each metric calculated for the example presented above.

4. Empirical validation[3]

To validate the structural complexity metrics as understandability indicators, we conducted an empirical study using the laboratory experiment format. The structural complexity of the ER diagrams used in this experiment was measured using the metrics suite defined in the previous section. We also independently assessed the understandability of the diagrams, either by measuring the study participants' performance on

---

[3] All the experimental material can be provided by the corresponding author upon request.
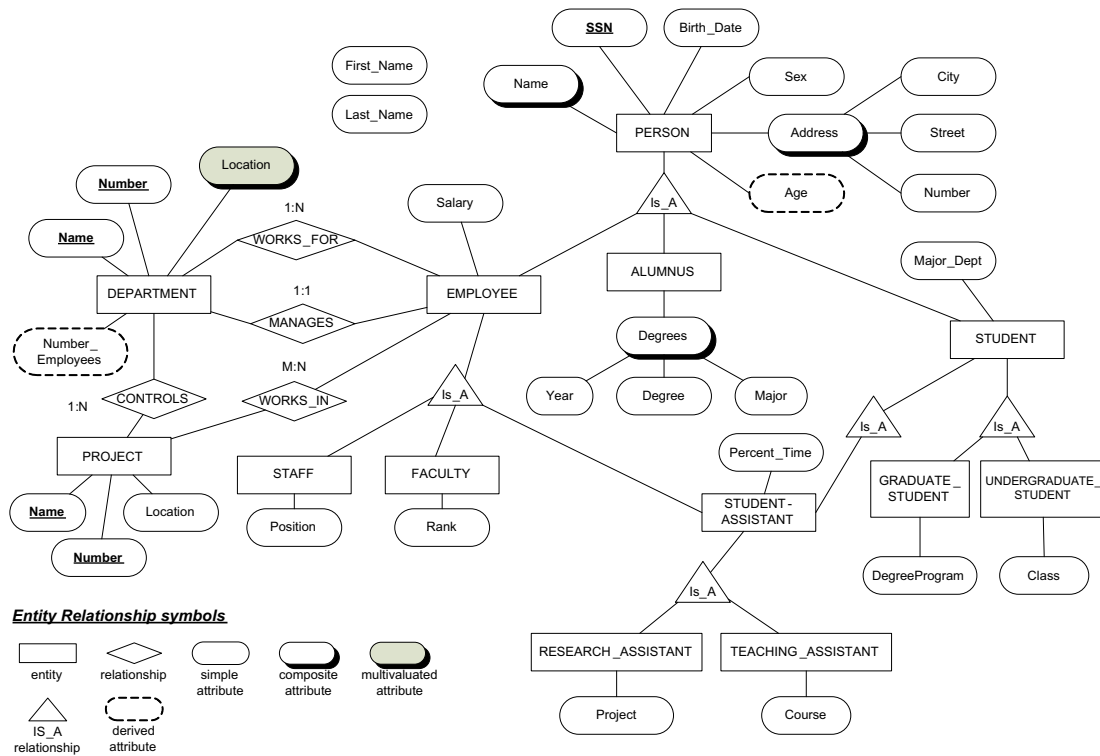
Fig. 2. Example of an ER diagram [24].

Table 4
Metric values for the ER diagram of Fig. 2

Metric	Value	Explanation
NE	13	Entities = PROJECT, DEPARTMENT, EMPLOYEE, STAFF, FACULTY, STUDENT-ASSISTANT, ALUMNUS, PERSON, STUDENT, GRADUATE_STUDENT, UNDERGRADUATE_STUDENT, RESEARCH_ASSISTANT, TEACHING_ASSISTANT
NA	23	Non-derived, single-value, simple attributes (not part of composite attributes) = PROJECT(Name, Number, Location), DEPARTMENT(Name, Number), EMPLOYEE(Salary), STAFF(position), FACULTY(Rank), STUDENT_ASSISTANT(Percent_Time), PERSON(SSN, Sex, Birth_Date) STUDENT(Major_Dept), GRADUATE_STUDENT(Degree_Program), UNDERGRADUATE_STUDENT(Class), RESEARCH_ASSISTANT(Project), TEACHER_ASISSTANT(Course) Derived attributes = DEPARTMENT(Number_Employees), PERSON(Age) Composite attributes = PERSON(Name, Address) and ALUMNUS(Degrees) Multivalued attributes = DEPARTMENT (Location)
NDA	2	Derived attributes = DEPARTMENT(Number_Employees), PERSON(Age)
NCA	3	Composite attributes = PERSON(Name, Address) and ALUMNUS(Degrees)
NMVA	1	Multivalued attributes = DEPARTMENT (Location)
NR	4	Relationships = CONTROLS, WORKS_FOR, MANAGES, WORKS_IN
NM:NR	1	M:N relationships = WORKS_IN
N1:NR	3	1:N relationships = WORKS_FOR, CONTROLS, MANAGES
NN_AryR	0	N-ary relationships = {}
NBinaryR	4	Binary relationships = CONTROLS, WORKS_FOR, MANAGES, WORKS_IN
NIS_AR	11	IS_A relationships = (PERSON, EMPLOYEE), (PERSON, ALUMNUS), (PERSON. STUDENT), (EMPLOYEE, STAFF), (EMPLOYEE, FACULTY), (EMPLOYEE, STUDENT_ASSISTANT), (STUDENT, STUDENT_ASSISTANT), (STUDENT, GRADUATE_STUDENT), (STUDENT, UNDERGRADUATE_STUDENT), (STUDENT_ASSISTANT, RESEARCH_ASSISTANT), STUDENT_ASSISTANT, TEACHING_ASSISTANT)
NRefR	0	Reflexive relationships = {}

maintenance-related tasks that require an understanding of the diagram (i.e. diagram comprehension tasks) or by measuring the participants' perceptions of diagram understandability. The validity of the metrics was examined by investigating the existence and statistical significance of the correlation between the measurement values for structural complexity and understandability.

Three pilot studies were conducted before the experiment proper (see Section 4.1). These pilot studies were mainly aimed at testing the experimental materials and measures and informed us on potential threats to study validity. The lessons drawn from these preliminary studies were used to alleviate the risks of these threats occurring in the 'real' experiment (which is presented in Section 4.2). The experimental process followed was based on a framework for experimental software engineering research proposed by Wholin et al. [65].

4.1. Pilot studies

The participants in the *first pilot study* were nine staff members of the Department of Information Systems and Technologies at the University of Castilla-La Mancha and seven students enrolled in the final year (i.e. fifth year of studies) of Computer Science at the same university. We selected 24 ER diagrams from a larger repository of diagrams that was collected specifically for the purpose of experimentation and metrics validation from case-studies and educational textbooks on database design [16,17]. The diagrams in this repository modelled different Universes of Discourse (UoD) covering a wide range of application domains such as the medical field, science, education, business (both commerce and industry), culture (e.g. arts, leisure, sports), and the broad societal/political field (including administration and law). They represent typical examples of conceptual data models that are used as main input for the logical database design process. The ER diagram of Fig. 2 is an example from this repository.

When selecting diagrams for the sample used in the pilot study, we tried to obtain a wide range of metric values (see Table 5). However, as our repository did not show sufficient variability in the values of four of the proposed 12 metrics (i.e. NDA, NCA, NMVA, NRefR), most of the diagrams in the sample had a value of zero for these metrics. These metrics were therefore not considered in the subsequent data analysis.

The dependent variable, understandability, was measured by assessing participants' perceptions, using a scale consisting of seven linguistic labels (see Table 6). A within-subject design was used, i.e. each participant had to rate the understandability of each of the 24 diagrams (in a different order for each participant).

The analysis of the obtained data indicated a significant positive correlation between the understandability rating of a diagram and seven (out of eight tested) metrics (NE, NA, NR, N1:NR, NM:NR, NBinaryR and NIS_AR). The correlation was not significant for the NN_AryR metric. A plausible reason is that the NN_AryR metric is zero for most of the 24 diagrams (with hindsight we could have omitted also this metric from the data analysis).

In the *second pilot study*, we tested an alternative procedure for measuring the understandability of ER diagrams. In the first pilot study the participants' perception of a diagram's understandability was not based on any experience with performing tasks using the diagram considered. The judgement of the participants might therefore be based on previous experiences with understanding ER diagrams, and is in this sense subjective. So, in the second pilot study a more objective measurement of understandability was obtained based on the performance of participants on a task that requires understanding of the diagram.

The participants were 31 students enrolled in the third year of Computer Science in the Department of Information Systems and Technologies at the University of Castilla-La Mancha in Spain. Four diagrams (different from the diagrams in the first pilot study) were selected from the repository. As the study was within-subject and hence for each diagram a task had to be performed, we limited the number of diagrams to four. We made sure that the selected diagrams showed a good spread in metric values for the metrics that had significant correlations with maintainability ratings in the first pilot study, the only exception being NIS_AR (see Table 7). Hence, the metrics considered were NE, NA, NR, NM:NR, N1:NR and NBinaryR. Because in the four diagrams all relationships were binary, the NR and NBinaryR values were the same for each diagram.

The task to be performed consisted of answering questions about the diagram, which allowed us to evaluate if the participant really understood the information conveyed by the diagram. For each diagram there were the same number of questions (five) and these questions were conceptually similar and posed in an identical order.

Table 5
Metric values for each ER diagram in the first pilot study

ER diagram	Metric values								Median of understandability ratings
	NE	NA	NR	NM:NR	N1:NR	NN_AryR	NBinaryR	NIS_AR	
DP1	5	19	5	2	3	0	5	0	2
DP2	12	44	11	3	8	2	9	4	3
DP3	6	23	7	0	7	0	7	2	4
DP4	9	33	11	5	6	0	11	2	5
DP5	13	15	11	8	3	0	11	5	2
DP6	6	16	4	4	0	1	3	3	3
DP7	6	16	40	4	0	1	3	3	2
DP8	11	33	8	1	7	0	8	4	4
DP9	6	45	7	0	7	0	7	0	3
DP10	4	2	3	0	3	0	3	0	1
DP11	9	31	6	1	5	0	6	2	3
DP12	6	17	3	0	3	0	3	2	2
DP13	12	11	5	4	1	0	5	9	3
DP14	10	35	11	4	11	1	10	0	3
DP15	8	29	6	0	6	0	6	2	3
DP16	6	29	4	4	0	2	2	0	3
DP17	6	34	5	2	3	0	5	0	3
DP18	6	15	5	0	5	0	5	0	2
DP19	6	16	4	2	2	0	4	0	2
DP20	8	19	5	5	0	0	5	3	3
DP21	5	10	3	2	1	0	3	1	3
DP22	9	26	5	2	3	0	5	4	2
DP23	8	12	5	3	2	0	5	4	3
DP24	11	32	7	1	6	0	7	4	4

Table 6
Understandability linguistic labels[a]

Extremely difficult to understand	Very difficult to understand	A bit difficult to understand	Neither difficult nor easy to understand	Quite easy to understand	Very easy to understand	Extremely easy to understand

 [a] For carrying out the data analysis we associated numbers to each linguistic label from 1 to 7. In this way 1 relates to extremely easy to understand and 7 relates to extremely difficult to understand.

Table 7
ER diagrams used in the first experiment

ER diagrams	Metric values				
	NE	NBinaryR	NA	N1:NR	NM:NR
DP25	3	2	11	0	2
DP26	5	6	15	1	5
DP27	8	10	21	4	6
DP28	5	6	13	1	5

Each participant had to write down the time spent answering the questionnaire, by recording the initial time and final time.

The dependent variable, understandability was measured using the following performance-based measures:

– Understandability Time (UT): The time needed to understand an ER diagram (expressed in minutes).
– Understandability Effectiveness (UEffec): The number of correct answers reflects how well the participants performed the required understandability tasks.

– Understandability Efficiency (UEffic): The number of correct answers divided by UT relates the understanding performance of the participants to their effort (in terms of time spent).

Using Spearman's correlation coefficient, each of the collected metrics was correlated separately to the measures for the dependent variable. All metrics showed a significant positive correlation with UT and a significant negative correlation with UEffic. The correlation with UEffec was not significant. All metrics (except NA) had exactly the same correlation coefficient for an understandability measure. This is a consequence of the use of the non-parametric Spearman's correlation coefficient, which is based on the relative ranking of the diagrams with respect to structural complexity. Each of the four diagrams takes the same position in this ranking according to each of the metrics (only NA gives a different ranking), which explains the observed pattern in correlation coefficients.

To be able to identify possible differences in the effects of the structural complexity metrics, diagrams should be used that can be ranked in different ways according to the metric values. This was tested in a *third pilot study* were nine diagrams (different from those in the previous pilot studies) were carefully selected such that differences in relative rankings were obtained depending on the metric considered. The metrics considered were those of the second pilot study, as well as NN_AryR (as in the first pilot study) and NRefR which for this sample showed reasonable variability (see Table 8). Participants were 27 students enrolled in the third year of Computer Science in the Department of Information Systems and Technologies at the University of Castilla-La Mancha. For measuring understandability the procedures of the first and second pilot studies were combined, so both a perception-based measure and the UT, UEffec and UEffic performance-based measures were used. The questions for the understanding task were similar as in the second pilot study (i.e. five questions per diagram). The understandability rating was made after performing the task.

In the *third pilot study*, the Spearman's correlation coefficients revealed that all ER diagram structural complexity metrics considered showed a significant positive correlation with the subjective ratings of understandability (i.e. diagrams with higher values for these metrics were perceived as more difficult to understand), a significant positive correlation with UT (i.e. task completion took longer for diagrams with higher metric values) and a significant negative correlation with UEffic (i.e. the efficiency of understanding decreases with increasing metric values). Consistent with the second pilot study, there was no significant correlation with UEffec.

Although the values of the correlation coefficients were now different for the different metrics, they were either all significant or all non-significant. Hence, again no differences in the effects of different structural complexity metrics could be distinguished. We realized that for the diagrams used in the pilot studies the structural complexity metrics were heavily intercorrelated, and thus for example if NE is correlated with NA, whenever NE is correlated with the measures of understandability, NA is also likely to be correlated. A particular concern for the design of the experiment is therefore to reduce as much as possible the correlations between the structural complexity metrics, without developing artificial diagrams (e.g. a diagram with 20 entities and only 2 relationships).

Table 8
ER diagrams used in the third pilot study

ER diagrams	Metric values							
	NE	NA	NR	N1:NR	NM:NR	NBinaryR	NN_AryR	NRefR
DP29	2	2	6	2	0	2	0	0
DP30	5	15	5	5	0	5	0	0
DP31	8	27	9	9	0	9	0	0
DP32	11	45	15	12	3	13	2	3
DP33	12	38	7	5	2	5	2	0
DP34	13	54	17	14	3	15	2	3
DP35	7	30	5	5	0	4	1	0
DP36	13	55	17	14	3	15	2	3
DP37	15	41	9	6	3	7	2	0

4.2. The experiment

Using the GQM template for goal definition ([3,4]), the goal of the experiment is defined as follows:

Analyse	*ER diagram structural complexity metrics*
For the purpose of	*Evaluating*
With respect to	*the capability to be used as indicators of ER diagram understandability*
From the point of view of	*the researchers*
In the context of	Professors and Ph.D. students *at the Department of Information Systems and Technologies at the University of Castilla-La Mancha*

4.2.1. Design

The participants were 17 professors and 11 Ph.D. students at the Department of Information Systems and Technologies at the University of Castilla-La Mancha. A within-subjects design was used to cancel out differences between the participants. The participants were more experienced in ER modelling than the participants in the pilot studies. They were chosen for convenience (being colleagues of the corresponding author), but were not informed about the goal of the study (in order to avoid experimenter bias). Being colleagues, they were motivated to perform well in the study.

The experimental material consisted of a guide explaining the ER notation and eight ER diagrams related to different UoDs, including culture/leisure (library operations, planning of tourist trips), science (biological processes), commerce (purchasing), industry (production process), and conducting business in general (debt financing, order taking and invoicing). The familiarity of the participants with the domains modelled was assessed as moderate. None of the participants was an expert in either of the domains modelled, but some personal (rather than professional) experience with one or more subject areas was plausible.

Like the diagrams used in the pilot studies, six diagrams were taken from the repository we built for experimentation and they originated from educational textbooks and real cases. Two diagrams were taken from [7,31], where they were used in other empirical studies on conceptual data modelling. Given their origin, the diagrams in our sample were mainly proposed for educational and research purposes. However, they are representative examples of conceptual data models that are used as input for the logical database design process. Although not extremely large, some diagrams were realistically large with NE values going up to 39. An example (diagram DE4) is shown in Appendix B.

We considered only six metrics (NE, NA, NR, NM:NR, N1:NR, NIS_AR) out of our initial proposal of 12 metrics (see Table 3). Our experience with the pilot studies motivated us to focus on this core set of metrics, trying to provide as much variance as possible and at the same time controlling (to the best possible extent) the correlation between these metrics. Once selected, most of the diagrams were repeatedly manipulated to obtain more favourable sample characteristics with respect to the range of metric values and the correlation between the metrics. To do this exercise simultaneously for 6 metrics was already a big challenge and extending this core set of metrics was not feasible. One simplification we made was to include only binary relationships and IS_A relationships in the diagrams. As a consequence, the NBinaryR metric takes the same values as the NR metric and the NRefR and NN_AryR metrics take on zero values for all diagrams. We further avoided the use of composite, multi-valued and derived attributes.

The structural complexity values of the diagrams are presented in Table 9, while Table 10 provides a correlational analysis of the metrics. As can be seen, some metric pairs are still significantly correlated (NE and NR, NR and N1:NR, NE and N1:NR) but further reducing the correlation between the number of entities and the number of relationships (most of which are binary 1:N) would have resulted in artificial diagrams. In a 'normal' ER diagram, the number of entities and relationships would tend to be correlated.

Each diagram had a test enclosed which included a questionnaire to assess whether the participants really understood the information conveyed by the diagram. Each questionnaire contained exactly the same number of questions (10) and the questions for each of the eight diagrams were conceptually similar (for an example

Table 9
ER diagrams used in the experiment

ER diagram	Metric values					
	NE	NA	NR	NM:NR	N1:NR	NIS_AR
DE1	7	35	5	1	4	2
DE2	10	33	5	1	4	4
DE3	8	34	9	6	3	2
DE4	15	30	12	3	9	4
DE5	11	50	17	10	7	2
DE6	37	79	24	8	16	4
DE7	39	52	17	4	13	11
DE8	20	28	16	1	15	2
Mean	18.38	42.63	13.13	4.25	8.88	3.88
Max	39.00	79.00	24.00	10.00	16.00	11.00
Min	7.00	28.00	5.00	1.00	3.00	2.00
St. Dev.	12.81	17.17	6.62	3.45	5.22	3.04

Table 10
Spearman's correlation coefficients between the metrics in the experiment (significant correlations shown in bold)

	NE	NA	NR	NM:NR	N1:NR	NIS_AR
NE	1	0.2857	**0.9341**	0.2683	**0.8862**	0.639
		$p = 0.4927$	**p = 0.0006**	$p = 0.5204$	**p = 0.0033**	$p = 0.0880$
NA	0.2857	1	0.3592	0.6587	0.21557	0.33130
	$p = 0.4927$		$p = 0.3820$	$p = 0.0756$	$p = 0.6081$	$p = 0.4503$
NR	**0.9341**	0.3592	1	0.4172	**0.9036**	0.4526
	p = 0.0006	$p = 0.3820$		$p = 0.3037$	**p = 0.0020**	$p = 0.2601$
NM:NR	0.2683	0.6587	0.4172	1	0.1472	0.0267
	$p = 0.5204$	$p = 0.0756$	$p = 0.3037$		$p = 0.7278$	$p = 0.9499$
N1:NR	**0.8862**	0.2155	**0.9036**	0.1472	1	0.4001
	p = 0.0033	$p = 0.6081$	**p = 0.0020**	$p = 0.7278$		$p = 0.3259$
NIS_AR	0.639	0.3129	0.4526	0.0267	0.4001	1
	$p = 0.0880$	$p = 0.4503$	$p = 0.2601$	$p = 0.9499$	$p = 0.3259$	

Table 11
Understandability linguistic labels for the experiment[a]

Very difficult to understand	A bit difficult to understand	Neither difficult nor easy to understand	Quite easy to understand	Very easy to understand

[a] For carrying out the data analysis, we assigned numbers to each linguistic level: 'Very easy to understand' corresponded to 1 and 'Very difficult to understand' corresponded to 5.

see Appendix B). To cancel out potential confounding effects of domain familiarity, participants were instructed that the answers to the questions had to be found in the diagrams.

Understandability was measured via the judgement of the participants about how easy or difficult they find it to understand the diagrams (an ordinal scale), according to five linguistic labels (see Table 11). We called this perception-based measure Subjective Understandability (SubUnd). Furthermore, as in the second and third pilot studies, understandability was also measured using performance-based measures:

– The time needed to understand an ER diagram (expressed in minutes, called Understandability Time (UT)).
– The number of correct answers reflects how well the participants performed the required understandability tasks, called Understandability Effectiveness (UEffec).
– The number of correct answers divided by UT relates the understanding performance of the participants to their effort (in terms of time spent). We called this measure Understandability Efficiency (UEffic).

The hypotheses we will to test for achieving our goal are

– $H_{1,0}$: There is no significant correlation between the metrics (NE, NA, NR, N1:NR, NM:NR, NIS_AR) and the Subjective Understandability. $H_{1,1}$:¬$H_{1,0}$
– $H_{2,0}$: There is no significant correlation between the metrics (NE, NA, NR, N1:NR, NM:NR, NIS_AR) and the Understandability Time: $H_{2,1}$:¬$H_{2,0}$
– $H_{3,0}$: There is no significant correlation between the metrics (NE, NA, NR, N1:NR, NM:NR, NIS_AR) and the Understandability Effectiveness. $H_{3,1}$:¬$H_{3,0}$
– $H_{4,0}$: There is no significant correlation between the metrics (NE, NA, NR, N1:NR, NM:NR, NIS_AR) and the Understandability Efficiency. $H_{4,1}$: ¬$H_{4,0}$

4.2.2. Data analysis and interpretation

We obtained 224 data points, from 28 subjects and 8 diagrams per subject. All of the subjects answered all the questions, so the number of answers in all cases is 10.

Before analysing the data for testing our hypotheses and as the time and efficiency measures (UT, UEffic) are meaningless unless a minimum level of quality is delivered, we excluded those data that have UEffec <0.80 (see Table 12), i.e. those data have more than two answers wrong (8 data points). Table 12 reflects that the correctness of the responses is high, because most of the cases (64%) have the UEffec value 1, which means that all the answers were correct.

Analysing the UEffic data, we found five outliers, as the box-plot of Fig. 3 shows. These five points were excluded from the analysis.

From the obtained 224 data points, after cleansing the data, we considered 211 data points for testing our hypotheses.

We started with the Kolmogorov–Smirnov test, which indicated that data distributions were not normal, hence we decided to use a non-parametric test statistic, i.e. Spearman's correlation coefficient (with a significance level set at $\alpha = 0.05$). The obtained Spearman's correlation coefficients are shown in Table 13.

The findings that Table 13 reveals are

– Relating to hypotheses 1 and 3, it seems that only the number of relationships (NR) and the number of 1:N relationships (N1:NR) (which includes by definition also the 1:1 relationships) affect the subjective perception of the understandability of the ER diagrams (SubUnd) and understandability effectiveness as measured by the number of correct answers to diagram understanding questions given (UEffec). The more binary relationships (in general) and 1:1 and 1:N relationships (in particular) an ER diagram has, the lower the number of understanding questions correctly answered and the higher the perceived difficulty in understanding the diagram.

Table 12
Distribution of understandability effectiveness by ER diagram

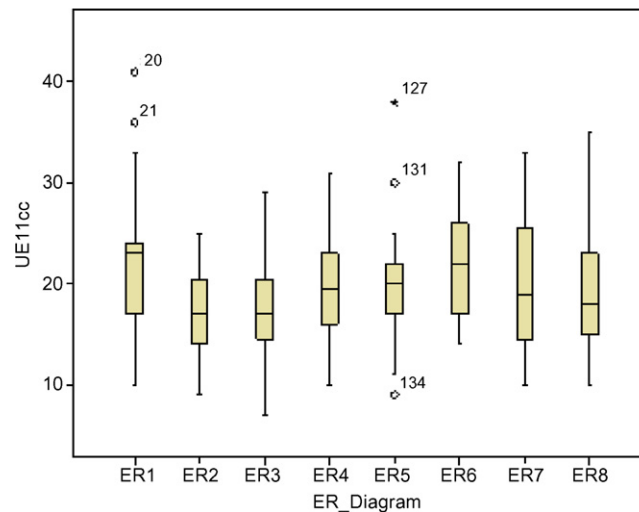| ER diagrams	UEffec					
	0.6	0.7	0.8	0.9	1	
DE1	0	2	1	4	21	28
DE2	0	1	5	13	9	28
DE3	0	1	1	9	17	28
DE4	0	0	3	7	18	28
DE5	0	0	2	8	18	28
DE6	3	0	1	9	15	28
DE7	0	0	1	2	25	28
DE8	0	1	2	4	21	28
	3	*5*	*16*	*56*	*144*	*224*
	1.34%	*2.23%*	*7.14%*	*25.00%*	*64.29%*	*100.00%*

Fig. 3. Box-plot of understandability efficiency by ER diagram.

Table 13
Spearman's correlation coefficients between the structural complexity metrics and the understandability measurements and ratings (significant correlations shown in bold) (experiment)

	NE	NA	NR	NM:NR	N1:NR	NIS_AR
UT	0.1259	**0.1596**	**0.1637**	0.0750	0.1750	0.0600
	$p = 0.071$	**p = 0.023**	**p = 0.017**	$p = 0.275$	p = 0.011	$p = 0.3870$
SubUnd	0.1340	0. 077	**0.1510**	0.0300	**0.2070**	0.0310
	$p = 0.052$	$p = 0.263$	**p = 0.029**	$p = 0.667$	**p = 0.003**	$p = 0.650$
UEffec	−0.1180	−0.0890	−0.1420	−0.0530	−0.1250	0.0350
	$p = 0.088$	$p = 0.200$	**p = 0.039**	$p = 0.447$	**p = 0.004**	$p = 0.613$
UEffic	−0.1230	**−0.1530**	**−0.1630**	−0.0760	−0.1770	−0.0550
	$p = 0.0750$	**p = 0.027**	**p = 0.018**	$p = 0.270$	**p = 0.0100**	$p = 0.423$

- Relating to hypotheses 2 and 4, the diagrams with higher metric values for NA, NR and N1:NR take more time to understand (UT) and have lower understandability efficiency (UEffic) values. Hence, the structural complexity of a diagram that is contributed by attributes and (1:1 and 1:N) relationships has a negative impact on the time required and the efficiency of participants in understanding a diagram.
- A remarkable result (and contrary to the pilot studies) is that the number of entities (NE), which can be considered as the purest measure of diagram size in our metrics suite, was not significantly correlated to any of the understandability measures. Hence, we found no evidence that the size of a diagram, as measured by the number of entities it contains, has an impact on the diagram's understandability.
- Although the presence of many-to-many relationships (NM:NR) and IS-A relationships (NIS_AR) could vary by a factor 1:10 and was quite high for some diagrams (maximum NM:NR was 10 and maximum NIS_AR was 11), these metrics were not significantly correlated to any of the understandability measures. Hence no effect on diagram understandability could be demonstrated.

5. Discussion and conclusions

5.1. Summary, results and research contributions

We proposed a set of metrics for measuring structural properties of ER diagrams. In total 12 metrics were defined for various ER modelling constructs that when used in an ER diagram determine the diagram's

structural complexity. Following empirical software engineering research, an artefact with a high structural complexity is also characterised by a high cognitive complexity which causes problems when understanding and maintaining the artefact. Hence, measurement of internal properties like structural complexity allows the assessment and prediction of the artefact's external quality.

Our metrics measure only syntactic properties of ER diagrams and thus can be calculated automatically, without human intervention or judgment, ensuring the cost-efficiency, consistency and repeatability of the measurements. The metrics are also theoretically valid because they were defined following the DIS-TANCE framework [56], which means that, from a measurement theory point of view, they are proven to be measures of an ER diagram's structural complexity. Moreover, the use of DISTANCE guarantees that the metrics can be used as ratio scale measurement instruments, which greatly facilitates the analysis of measurement data.

We also evaluated the capability of the proposed metrics to be used as indicators of ER diagram understandability. A laboratory experiment was conducted with the aim of testing the hypothesized correlation between the proposed metrics and both performance-based and perception-based measures of ER diagram understandability. The empirical validation was limited to a core set of six metrics capturing the ER diagram structural complexity that is contributed by the diagram's entities, attributes, binary associations between entities (with different types of maximum cardinalities), and IS-A relationships. Six other metrics that were proposed, including fine-grained metrics for different kinds of attributes and metrics for different grades of relationships (i.e. unary, binary, higher-order), were not tested because of difficulties in finding a sufficiently large set of realistic ER diagrams with both a wide spread in values for all metrics considered and little inter-correlation between the metrics. It was therefore decided to focus first on a set of six core metrics, i.e. the NE, NA, NR, NIS_AR, N1:NR, NM:NR metrics.

The result of the experiment was that three metrics, i.e. NA, NR and N1:NR, could be validated as indicators of ER diagram understandability. The correlation between these metrics and some or all of the understandability measures was statistically significant. The correlation between the other three metrics, i.e. NE, NM:NR and NIS_AR, and the understandability measures was not significant.

The experiment showed that the correctness of understanding, as measured through the answers given to diagram comprehension questions (referred to as the UEffec measure in the experiment description), was affected by the number of relationships (NR) and the number of 1:1 and 1:N relationships (N1:NR). The more relationships an ER diagram contains, and in particular the more 1:1 and 1:N relationships it contains, the lower the correctness scores. Also the perceived ease of understanding an ER diagram decreases when the number of (1:1 and 1:N) relationships increases.

The same two metrics were also correlated to measures of the efficiency by which an ER diagram can be understood (i.e. time it takes to understand the diagram and correctness of understanding relative to understanding effort). Also the number of attributes (NA) had an impact here. Hence, the more relationships, and especially the more 1:1 and 1:N relationships, and the more attributes an ER diagram contains, the more time it takes to understand it. Of particular interest here is that the NA metric was not correlated to the NR and N1:NR metrics, meaning that the number of attributes and the number of (1:1 and 1:N) relationships each on their own affect understandability.

An equally interesting result is that the number of entities (NE) had no direct effect on understandability (whatever measure was used). Through its correlation with the number of relationships, the number of entities might of course have an indirect understandability effect (i.e. diagrams with a lot of entities normally also have a lot of relationships). However, the experiment result allows identifying those structural complexity aspects that directly affect diagram understanding (relationships and attributes) as opposed to other aspects, like diagram size (number of entities), for which no evidence of a direct impact on understandability was found. Likewise no understandability effect was found for the number of many-to-many relationships (NM:NR) and the number of IS-A relationships (NIS_AR), both of which were not correlated to any of the other structural complexity metrics for the sample of diagrams used in the experiment.

Summarizing, the research contribution of this paper is the definition and validation of three ER diagram structural complexity metrics (NA, NR and N1:NR) that can be used as early, easily and objectively

calculated indicators of ER diagram understandability. Our research shows that the understandability of an ER diagram is affected by its structural complexity where the contributors to this structural complexity are the diagram's attributes and relationships, in particular the 1:1 and 1:N relationships. The more attributes and (1:1 and 1:N) relationships a diagram contains, the less understandable it is. There is no evidence that the size of a diagram in terms of number of entities affects understandability, unless through its expected correlation with the number of relationships.

5.2. Implications and recommendations

The correlation analysis that we performed suggests that when building maintenance-related prediction models for ER diagrams, it is advised to include at least one metric related to the number of attributes and at least one other metric related to the number of relationships, which should preferably include counts of the binary 1:1 and 1:N relationships. These structural complexity aspects seem to have the strongest impact on subjectively experienced understandability and on the efficiency and effectiveness in understanding ER diagrams. We found no evidence that to predict understandability a model should also include a measure of diagram size (in terms of the number of entities). Of course, indirectly, diagram size has an impact on understandability because diagrams with a large number of entities would normally also have many relationships. This result is interesting given that the only ER diagram structural complexity metric previously validated (the number of entities and relationships metric [48]) does not distinguish between structural complexity that is due to heavy interconnections between entities and structural complexity that stems from diagram size.

Our study helps understanding what factors may inhibit ER diagram comprehension. From this understanding, further research may derive quality-focused ER diagram design knowledge which can then be incorporated into the modelling language, method and process. Our experiment suggests that to control diagram understandability, primary attention should be paid to the number of attributes and the number of relationships, especially binary 1:1 and 1:N relationships, contained in the diagram. An interesting avenue for further research is the development and evaluation of ER diagram refactoring rules that reduce the number of attributes and binary 1:1 and 1:N relationships, replacing them by semantically equivalent modelling solutions. We provide two examples here, while acknowledging and stressing that they are tentative and need further research:

- If a diagram contains entities that have one or more identical attributes, then generalization and inheritance can be applied to reduce the number of attributes. The reduction in NA will improve understandability time and efficiency, while the increase in NE and NIS_AR will not have an effect on understandability. Of course, the formulation of this rule is speculative, given that our experiment does not allow directly testing whether it would indeed improve understandability. We did for instance not separately measured the attributes that were shared by two or more entities. To evaluate this rule, a controlled experiment must be conducted comparing the treatment (application of the rule) to a control group. The diagram(s) in the control group should contain entities with identical attributes.

- Our results do not support the practice of objectifying M:N relationships. Under some conditions an M:N relationship between two entities A and B can be replaced by a connecting entity C which is related to A via an 1:N relationship and to B via another 1:N relationship. The attributes of the M:N relationship would become attributes of the new entity C. Objectification thus reduces the value of NM:NR, keeps NA constant, but increases the values of NE and N1:NR. According to our results, the net effect on understandability would be negative since adding 1:N relationships lowers understandability whereas the other changes have no effect. Our study thus actually supports the opposite of objectification and favours the use of an M:N relationship instead of two 1:N relationships and a connecting entity. Again, we formulate this rule with the necessary precaution because our study does not directly validate it. A problem with our current study is that we did not separately measured entity attributes and relationship attributes (maybe they have different effects on understandability?). Only controlled experimentation comparing both situa-

tions can provide definite answers. One such experiment has been conducted in a related study on UML class diagrams (see [58]) where it was shown that objectification reduces comprehension if users are not modelling experts.

5.3. Limitations and future work

A first limitation of our research is that only half the metrics suite was evaluated in the empirical study. The capability to be used as understandability indicators was not tested for the three detailed metrics that distinguish specific kinds of attributes, i.e. composite attributes (NCA), multi-valued attributes (MVA) and derived attributes (NDA). With hindsight, this is not really a problem given that the majority of tools available for drawing ER diagrams do not incorporate facilities for specifying composite attributes, multivalued attributes and derived attributes, and that therefore most of the designers do not use these constructs in their ER diagrams.

Also the three metrics that distinguish relationships according to their grade, i.e. unary (NRefR), binary (NBinaryR), and higher-order (NN-AryR), were not tested. As commented before, the ER diagrams of the experiment were chosen and further manipulated such that wide metric value ranges were obtained and intercorrelations between the metrics were avoided. This design control turned out to be a difficult exercise to be performed for all metrics simultaneously. Hence, the choice was made to include only the most common form of relationships (i.e. binary relationships) in the diagrams (making the NR values the same as the NBinaryR values). Future research may wish to extend our study to find out if there are understandability differences between relationships of different grades. For instance, a controlled experiment can be conducted to investigate if diagram users understand better binary relationships or ternary relationships.

A second limitation relates to choices made when proposing our suite of 12 metrics. For instance, we have no separate metric for weak entities, although many people have difficulties in understanding them.[4] Another choice made was to make no distinction between entity attributes and relationship attributes (see also our discussion on objectification in the previous sub-section). However, according to the Bunge–Wand–Weber representational model [64], which is often used to evaluate conceptual modelling languages, relationships with attributes should be prohibited because they reduce the ontological clarity of models [61]. Likewise our metrics suite distinguishes types of binary relationships based on maximum cardinalities, but does not include separate metrics for differences in minimum cardinalities, i.e. whether relationships are optional or mandatory (or partial versus full). Here the BWW model prohibits the use of optional entity properties, including optional relationships in which the entities participate [61]. Controlled experiments testing these ontological predictions have been undertaken by fellow researchers (see [7,13,31]). In the absence of definite results more work needs to be done. The results of these studies may inform us on how to further refine our metrics suite.

A third limitation is the weaknesses inherent in our choice of research method, i.e. the experiment, and the design choices we made. Because of the difficulty of controlling the independent variable, the observed correlations do not demonstrate per se a causal relationship between structural complexity and understandability. They only provide empirical evidence of it. Only experiments where all included metrics would be varied in a controlled manner and all other factors would be held constant, could really demonstrate causality. On the other hand, it is difficult to imagine what could be alternative explanations for our results besides a relationship between structural complexity and understandability.

The use of academics and doctoral students as experiment participants might produce a problem to external validity. However, as long as the tasks performed do not require high levels of industrial experience, experiments involving non-professionals (e.g. students) can be justified [5].

Further empirical validation, including internal and external replication of the experiment presented in this paper, is needed. We also need to carry out additional empirical studies, for instance with practitioners, in order to extend our current study and to further build up the cumulative knowledge on ER diagram structural complexity and understandability. Finally, more data related to 'real projects' is needed to

---

[4] We thank an anonymous reviewer for this remark.

strengthen the evidence that these metrics can be used as practical ER diagram understandability indicators.

As several authors remark [22,30] the practical utility of metrics would be enhanced if meaningful thresholds could be identified. However a recent series of studies which looked at thresholds for object-oriented metrics [6,23] demonstrate that there are no threshold values for contemporary object-oriented metrics. To our knowledge, studies trying to find thresholds values for ER diagram metrics do not exist, so we believe that could be a good issue to tackle in the future.

Pending is also to carry out similar experimental work but focusing not only on understandability but also on modifiability, in order to ascertain if there exists a relationship between the structural complexity of ER diagrams and the performance of subjects when doing modification tasks. In addition it could be valuable to investigate if understandability is related with modifiability, i.e. if the performance of the subjects when understanding an ER diagram is related with the performance of the modification tasks [60].

Acknowledgements

Appendix A. Distance-based definition of the NA metric

In Section 3 the NA metric was defined as the number of attributes defined within an ER diagram. The metric intends to measure an aspect of structural complexity which basically states that the more attributes defined within an ER diagram, the higher its structural complexity.

Using the method described in [55,56] the DISTANCE-based definition of a metric consists of five steps:

– *Finding a measurement abstraction.* The set of objects that are characterised by a structural complexity property is the Universe of ER diagrams (notation: UERD), i.e. the set of all conceivable syntactically correct ER diagrams that are relevant to some Universe of Discourse (UoD) (which does not need to be specified any further). The structural complexity property for which the NA metric is proposed (referred to as *pty*) is the number of attributes defined within an element of UERD. Let UA be the Universe of Attributes relevant to the UoD. The set of attributes defined within an ERD $\in$ UERD, denoted by SA(ERD), is a subset of UA. The sets of attributes defined within the ER diagrams of UERD are elements of the power set of UA (notation: $\wp$(UA)). As a consequence we can equate the set of measurement abstractions (referred to as M in the DISTANCE procedure) to $\wp$(UA) and define a mapping function abs_{NA} as:

$$abs_{NA} : UERD \rightarrow \wp(UA) : ERD \rightarrow SA(ERD)$$

This function simply maps an ER diagram onto its set of attributes, which is a sufficient representation of an ER diagram for the aspect of structural complexity considered.
– *Defining distances between measurement abstractions.* A set of elementary transformation functions for $\wp$(UA), denoted by $T_{e-\wp(UA)}$, must be found such that any set of attributes can be transformed into any other set of attributes. It was shown in [55] that to transform sets, only two elementary transformation functions are required: one that adds an element to a set and another that removes an element from a set. So, given two sets of attributes $s_1 \in \wp$(UA) and $s_2 \in \wp$(UA), s_1 can always be transformed into s_2 by removing first all attributes from s_1 that are not in s_2, and then adding all attributes to s_1 that are in s_2, but were not in the original s_1. In the 'worst case scenario', s_1 must be transformed into s_2 via an empty set of attributes. Formally, $T_{e-\wp(UA)} = \{t_{0-\wp(UA)}, \ t_{1-\wp(UA)}\}$, where $t_{0-\wp(UA)}$ and $t_{1-\wp(UA)}$ are defined as

$$t_{0-\wp(\mathrm{UA})} : \wp(\mathrm{UA}) \to \wp(\mathrm{UA}) : s \to s \cup \{a\}, \quad \text{with } a \in \mathrm{UA}$$
$$t_{1-\wp(\mathrm{UA})} : \wp(rmUA) \to \wp(\mathrm{UA}) : s \to s - \{a\}, \quad \text{with } a \in \mathrm{UA}$$

– *Quantifying distances between measurement abstractions.* It was shown in [63] that the symmetric difference model, which is a particular instance of Tversky's contrast model, can always be used to define a metric (in the mathematical sense) for distances between sets. Hence, we define the metric $\delta_{\wp(\mathrm{UA})}$ as

$$\delta_{\wp(\mathrm{UA})} : \wp(\mathrm{UA}) \times \wp(\mathrm{UA}) \to \mathfrak{R} : (s, s') \to |s - s'| + |s' - s|$$

This definition is equivalent to stating that the distance between two sets of attributes, as defined by the shortest sequence of elementary transformations between these sets, is measured by the count of elementary transformations in the sequence. Note that for any element in s but not in s′ and for any element in s′ but not in s, an elementary transformation is required.

– *Finding a reference abstraction.* For the structural complexity property that is considered, the obvious reference point for measurement is the empty set of attributes. It can be argued that an ER diagram with no attributes defined has the lowest structural complexity that can be imagined. Therefore we define the mapping function ref_{NA} as

$$ref_{\mathrm{NA}} : \mathrm{UERD} \to \wp(\mathrm{UA}) : \mathrm{ERD} \to \emptyset$$

– *Defining a metric for the property.* The NA metric can be formally defined as a function that returns for any $\mathrm{ERD} \in \mathrm{UERD}$ the value of the metric $\delta_{\wp(\mathrm{UA})}$ for the pair of sets SA(ERD) and ∅:

$$\forall \mathrm{ERD} \in \mathrm{UERD} : \mathrm{NA}(\mathrm{ERD}) = \delta_{\wp(\mathrm{UA})}(\mathrm{SA}(\mathrm{ERD}), \emptyset)$$
$$= |SA(ERD) - \emptyset| + |\emptyset - \mathrm{SA}(\mathrm{ERD})|$$
$$= |SA(ERD)|$$

As a consequence, a metric that returns the number of attributes in an ER diagram qualifies as a metric (in the sense of measurement theory) of the structural complexity property that is determined by the quantity of attributes defined within an ER diagram.

Appendix B. An example of the experimental material

Here, we show, as an example, the ER diagram DE4 and its understandability and rating tasks.

Diagram DE4
 (a) Take the attached diagram.
 (b) Write down the starting time (indicating hh:mm:ss)
 (c) Answer the following questions (YES/NO):
     1. Is an order represented as an entity?
     2. Must all warehouses have warehouses an insurance ID attribute?
     3. Can a payment be related with more than one order header?
     4. Must a supplier supply at least one stock item?
     5. Must each stock item be associated with at least one order line in some order?
     6. Can the stock items be stored in warehouses that are not own warehouses?
     7. Is 'tax' an attribute of order summary?
     8. Must an order have at least one order line?
     9. Must each warehouse contain at least one stock item?
     10. Must each Invoice Header be associated with only one line item?

 (d) Write down the finishing time (indicating hh:mm:ss)
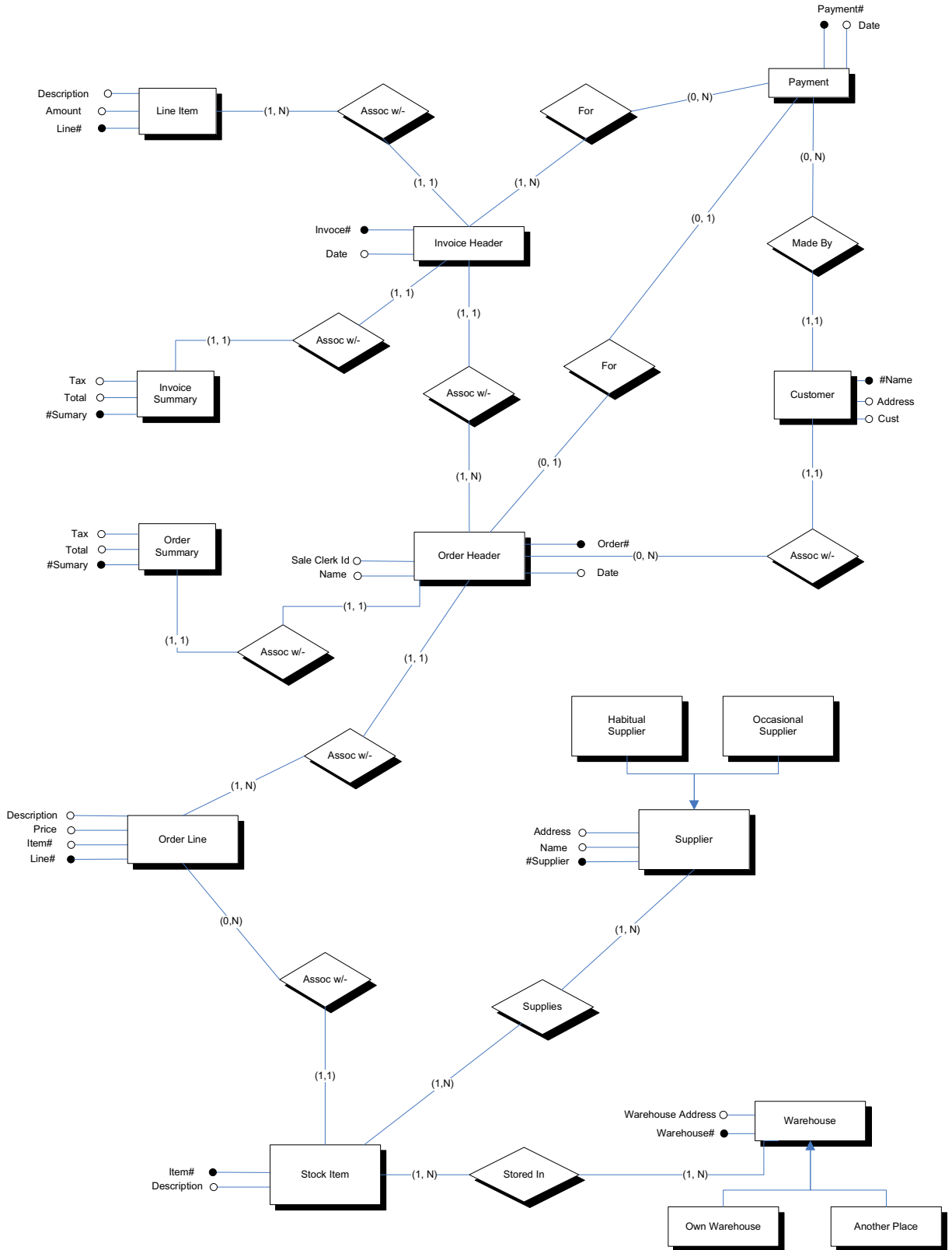
(e) According to your criterion, rate how easy or difficult were for you to understand the diagram (mark with a cross).

Very difficult to understand	A bit difficult to understand	Neither difficult nor easy to understand	Quite easy to understand	Very easy to understand

References

[1] J. Abrial, Data Semantics, in: IFIP TC2 Conference, North Holland, Amsterdam, 1974.

[2] R. Bandi, V. Vaishnavi, D. Turk, Predicting maintenance performance using object-oriented design complexity metrics, IEEE Transactions on Software Engineering 29 (1) (2003) 77–87.

[3] V. Basili, H. Rombach, The TAME project: towards improvement-oriented software environments, IEEE Transactions on Software Engineering 14 (6) (1988) 728–738.

[4] V. Basili, D. Weiss, A methodology for collecting valid software engineering data, IEEE Transactions on Software Engineering 10 (1984) 728–738.

[5] V. Basili, F. Shull, F. Lanubile, Building knowledge through families of experiments, IEEE Transactions on Software Engineering 25 (4) (1999) 435–437.

[6] S. Benlarbi, K. El-Emam, N. Goel, S. Rai, Thresholds for object-oriented measures, NRC/ERB 1073, 2000.

[7] F. Bodart, A. Patel, M. Sim, R. Weber, Should optional properties be used in conceptual modelling? A theory and three empirical tests, Information System Research 12 (4) (2001) 384–405.

[8] B. Boehm, Software Engineering Economics, Prentice-Hall, 1981.

[9] L. Briand, C. Bunse, J. Daly, A controlled experiment for evaluating quality guidelines on the maintainability of object-oriented designs, IEEE Transactions on Software Engineering 27 (6) (2001) 513–530.

[10] L. Briand, J. Wüst, Empirical studies of quality models in object-oriented systems, in: M. Zelkowitz (Ed.), Advances in Computers, vol. 59, Academic Press, 2002, pp. 97–166.

[11] L. Briand, J. Wüst, S. Ikonomovski, H. Lounis, Investigating quality factors in object-oriented designs: An industrial case-study, in: 21st International Conference on Software Engineering, Los Angeles, CA, 1999, pp. 345–354.

[12] L. Briand, J. Wüst, H. Lounis, Replicated case studies for investigating quality factors in object-oriented designs, Empirical Software Engineering 6 (1) (2001) 11–58.

[13] A. Burton-Jones, R. Weber, Understanding relationships with attributes in entity–relationship diagrams, in: 20th Annual International Conference on Information Systems (ICIS'99), Charlotte, NC, USA, 1999, pp. 214–228.

[14] P. Chen, The entity–relationship model: toward a unified view of data, ACM Transactions on Database Systems 1 (1) (1976) 9–37.

[15] I. Davies, P. Green, M. Rosemann, M. Indulska, S. Gallo, How do practitioners use conceptual modeling in practice? Data and Knowledge Engineering 58 (2006) 358–380.

[16] A. De Miguel, M. Piattini, Fundamentos y modelos de bases de datos, second ed., Ra-Ma, Madrid, 1999.

[17] A. De Miguel, M. Piattini, E. Marcos, Diseño de bases de datos relacionales, Ra-Ma, Madrid, 1999.

[18] C. Eick, Lockermann, Acquisition of terminological knowledge using database design techniques, in: ACM-SIGMOD Conference on Management of Data, 1985, pp. 84–94.

[19] C. Eick, T. Raupp, Towards a formal semantics and inference rules for conceptual data models, Data and Knowledge Engineering 6 (1991) 297–317.

[20] C. Eick, A methodology for the design and transformation of conceptual schemas, in: 17th International Conference on Very Large Data Bases, 1991, pp. 25–34.

[21] K. El-Emam, The prediction of faulty classes using object-oriented design metrics, NRC/ERB1064, National Research Council Canada, 1999.

[22] K. El-Emam, Object-oriented metrics: a review on theory and practice, NRC/ERB 1085, National Research Council Canada, 2001.

[23] K. El-Emam, S. Benlarbi, N. Goel, S. Rai, The confounding effect of class size on the validity of object-oriented metrics, IEEE Transactions on Software Engineering 27 (7) (2001) 630–650.

[24] R. Elmasri, S. Navathe, Fundamentals of Database Systems, second ed., Addison-Wesley, Massachussets, 1994.

[25] A. Enders, H.D. Rombach, A handbook of Software and Systems Engineering: Empirical Observations, Laws and Theories, Addison-Wesley, 2003.

[26] J. Erickson, K. Siau, Theoretical and practical complexity of UML, in: Tenth Americas Conference on Information Systems, New York, USA, 2004, pp. 1669–1674.

[27] N. Fenton, Software measurement: a necessary scientific basis, IEEE Transactions on Software Engineering 20 (3) (1994) 199–206.

[28] N. Fenton, S. Pfleeger, Software Metrics: A Rigorous Approach, second ed., Chapman & Hall, London, 1997.

[29] F. Fioravanti, P. Nesi, Estimation and prediction metrics for adaptive maintenance effort of object-oriented systems, IEEE Transactions on Software Engineering 27 (12) (2001) 1062–1083.

[30] V. French, Establishing software metric thresholds, International Workshop on Software Measurement (IWSM'99), 1999.

[31] A. Gemino, Y. Wand, Complexity and clarity in conceptual modeling: comparison of mandatory and optional properties, Data and Knowledge Engineering 55 (2005) 301–326.

[32] M. Genero, Defining and validating measures for conceptual models, Ph.D. Thesis, Department of Computer Science, University of Castilla-La Mancha, Spain, 2002.

[33] M. Genero, MaE. Manso, A. Visaggio, G. Canfora, M. Piattini, Building a metric-based prediction model for UML class diagram maintainability, Empirical Software Engineering 12 (5) (2007) 517–549.

[34] R. Gray, B. Carey, N. McGlynn, A. Pengelly, Design metrics for database systems, BT Technology 9 (4) (1991).

[35] M. Hammer, D. McLeod, Database description with SDM: a semantic database model, ACM TODS 6 (3) (1981) 351–386.

[36] R. Harrison, S. Counsell, R. Nithi, Experimental assessment of the effect of inheritance on the maintainability of object-oriented systems, The Journal of Systems and Software 52 (2000) 173–179.

[37] D. Ince, M. Shepperd, Algebraic validation of software metrics, ESEC 1991, 1991.

[38] ISO 9126, Software product evaluation-quality characteristics and guidelines for their use, ISO/IEC Standard 9126, Geneva, 2001.

[39] S. Kesh, Evaluating the quality of entity relationship models, Information and Software Technology 37 (12) (1995) 681–689.

[40] D. Krantz, R. Luce, P. Suppes, A. Tversky, Foundations of Measurement, vol. 1, Academic Press, New York, 1971.

[41] J. Krogstie, O. Lindland, G. Sindre, Towards a deeper understanding of quality in requirements engineering, in: Proceedings of the Seventh International Conference on Advanced Information Systems Engineering (CAISE 1995), Jyvaskyla, Finland, 1995, pp. 82–95.

[42] W. Li, S. Henry, Object-oriented metrics that predict maintainability, Journal of Systems and Software 23 (2) (1993) 111–122.

[43] O. Lindland, A. Sindre, A. Solvberg, Understanding quality in conceptual modeling, IEEE Software 11 (2) (1994) 42–49.

[44] A. Maes, G. Poels, Evaluating quality of conceptual modelling scripts based on user perceptions, Data and knowledge Engineering (2007), doi:10.1016/j.datak.2007.04.008.

[45] MaE. Manso, M. Genero, M. Piattini, No-redundant metrics for UML class diagrams structural complexity, in: J. Eder, M. Missikoff, (Eds.), CAISE 2003, Lecture Notes in Computer Science, vol. 2681, Springer-Verlag, 2003, pp. 127–142.

[46] D. Moody, Metrics for evaluating the quality of entity relationship models, in: Seventeenth International Conference on Conceptual Modelling (E/R'98), Singapore, 1998, pp. 213–225.

[47] D. Moody, G. Shanks, Improving the quality of data models, information systems, Empirical Validation of a Quality Management Framework 28 (6) (2003) 619–650.

[48] D. Moody, G. Shanks, P. Darke, Improving the quality of entity relationship models – experience in research and practice, in: Seventeenth International Conference on Conceptual Modelling (ER'98), Singapore, 1998, pp. 255–276.

[49] D.L. Moody, Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions, Data and Knowledge Engineering 55 (3) (2005) 243–276.

[50] R. Muller, Database Design for Smarties. Using UML for Data Modelling, Morgan Kaufman, San Francisco, 1999.

[51] J. Nelson, G. Poels, M. Genero, M. Piattini, Quality in conceptual modeling – five examples of the state of art, Data and Knowledge Engineering 55 (3) (2005) 237–242.

[52] A. Olivé, Specific relationship types in conceptual modeling: the cases of generic and with common participants, unpublished keynote lecture, in: Fourth International Conference on Enterprise Information Systems (ICEIS'02), Ciudad Real, Spain, 2002, <http://www.iceis.org/iceis2002/keynote.htm>.

[53] OMG (Object Management Group), Unified Modeling Language (UML) Specification, Version 1.4. Object Management Group (OMG), 2001.

[54] N. Pippinger, Complexity theory, Scientific American 238 (6) (1978) 1–15.

[55] G. Poels, On the formal aspects of the measurement of object-oriented software specifications, Ph.D. Thesis, Faculty of Economics and Business Administration. Katholieke Universiteit Leuven, Belgium, 1999.

[56] G. Poels, G. Dedene, Distance-based software measurement: necessary and sufficient properties for software measures, Information and Software Technology 42 (1) (2000) 35–46.

[57] G. Poels, G. Dedene, Evaluating the effect of inheritance on the modifiability of object-oriented business domain models, in: Fifth European Conference on Software Maintenance and Reengineering (CSMR 2001), Lisbon, Portugal, 2001, pp. 20–28.

[58] G. Poels, F. Gailly, A. Maes, R. Paemeleire, Object class or association class? Testing the user effect on cardinality interpretation, Lecture Notes in Computer Science 3770 (2005) 33–42.

[59] R. Schuette, T. Rotthowe, The guidelines of modeling – an approach to enhance the quality in information models, in: Seventh International Conference on Conceptual Modelling (ER'98), 1998, pp. 240–254.

[60] T. Shaft, I. Vessey, The role of cognitive fit in the relationship between software comprehension and modification, MIS Quarterly 30 (1) (2006) 29–55.

[61] G. Shanks, E. Tansley, R. Weber, Using ontology to validate conceptual models, Communications of the ACM 46 (2003) 85–89.

[62] Si-Said Cherfi, J. Akoka, I. Comyn-Wattiau, Conceptual modelling quality – from EER to UML schemas evaluation, in: 21st International Conference on Conceptual Modeling (ER 2002), Tampere, Finland, 2002, pp. 499–512.

[63] P. Suppes, D. Krantz, R. Luce, A. Tversky, Foundations of measurement: geometrical, Threshold and Probabilistic Representations, vol. 2, Academic Press, San Diego, 1989.

[64] Y. Wand, R. Weber, An ontological model of an information system, IEEE Transactions on Software Engineering 16 (11) (1990) 1282–1292.

[65] C. Wohlin, P. Runeson, M. Höst, M. Ohlson, B. Regnell, A. Wesslén, Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, 2000.

Marcela Genero is Associate Professor at the Department of Information Systems and Technologies at the University of Castilla-La Mancha, Ciudad Real, Spain. She received her MSc degree in Computer Science in the Department of Computer Science of the University of South, Argentine in 1989, and her Ph.D. at the University of Castilla-La Mancha, Ciudad Real, Spain in 2002. Her research interests are: empirical software engineering, software metrics, conceptual models quality, database quality, quality in product lines, quality in model-driven development, etc. Marcela Genero has published in prestigious journals (Information and Software Technology, Journal of Software Maintenance and Evolution: Research and Practice, L'Objet, Data and Knowledge Engineering, Empirical Software Engineering, Journal of Object Technology, Journal of Research and Practice in Information Technology, Empirical Software Engineering, etc.), and conferences (CAiSE, ER, MODELS/UML, ISESE, METRICS, ESEM, SEKE, etc. She edited with Mario Piattini and Coral Calero the books titled ''Data and Information Quality'' (Kluwer, 2001), and ''Metrics for Software Conceptual Models'' (Imperial College, 2005). She is member of the International Software Engineering Research Network (ISERN).

Geert Poels is a Professor in the rank of Lecturer at the Department of Management Information, Operations Management and Technology Policy of Ghent University (Belgium). He holds degrees in Business Engineering and Computer Science, and a Ph.D. in Applied Economic Sciences. His research interests include software metrics, conceptual modeling, business ontology, and accounting information systems. Dr. Poels has published in IEEE Transactions on Software Engineering, Data & Knowledge Engineering, Software and Systems Modeling, Information and Software Technology and Lecture Notes in Computer Science, and presented at conferences such as ER and CAiSE. In 2002, 2003, 2006 and 2007 he co-organized the IWCMQ/QoIS workshops on conceptual model and information system quality at the ER conference.

Mario Piattini is MSc and Ph.D. in Computer Science by the Technical University of Madrid. Certified Information System Auditor by ISACA (Information System Audit and Control Association). Full Professor at the Department of Information Systems and Technologies at the University of Castilla-La Mancha, in Ciudad Real, Spain. Author of several books and papers on databases, software engineering and information systems. He leads the ALARCOS research group at the University of Castilla-La Mancha.