Requirements Eng (2004) 9: 248–260
DOI 10.1007/s00766-004-0204-6

ORIGINAL ARTICLE

Andrew Gemino · Yair Wand

A framework for empirical evaluation of conceptual modeling techniques

Received: 10 January 2004 / Accepted: 25 May 2004 / Published online: 15 October 2004

Abstract The paper presents a framework for the empirical evaluation of conceptual modeling techniques used in requirements engineering. The framework is based on the notion that modeling techniques should be compared via their underlying grammars. The framework identifies two types of dimensions in empirical comparisons—affecting and affected dimensions. The affecting dimensions provide guidance for task definition, independent variables and controls, while the affected dimensions define the possible mediating variables and dependent variables. In particular, the framework addresses the dependence between the modeling task—model creation and model interpretation—and the performance measures of the modeling grammar. The utility of the framework is demonstrated by using it to categorize existing work on evaluating modeling techniques. The paper also discusses theoretical foundations that can guide hypothesis generation and measurement of variables. Finally, the paper addresses possible levels for categorical variables and ways to measure interval variables, especially the grammar performance measures.

Keywords System analysis and design · Information systems development · Conceptual modeling · Empirical comparison · Requirements engineering

A. Gemino (✉)
Faculty of Business Administration,
Simon Fraser University, 8888 University Drive,
Burnaby, BC, Canada, V5A 1S6
E-mail: gemino@sfu.ca
Tel.: +1-604-2914991
Fax: +1-604-2914920

Y. Wand
Sauder School of Business,
University of British Columbia,
Vancouver, BC, Canada, V6T 1Z2
E-mail: yair.wand@ubc.ca
Tel.: +1-604-8228395
Fax: +1-604-8220045

1 Introduction

Conveying and promoting understanding of the application domain is one of the primary objectives in requirements gathering during information systems analysis. Typically, this information is formalized in models of the application domain, created using conceptual modeling techniques (CMTs). As Mylopoulos [27, p. 2] suggests, "*Conceptual modeling is the activity of formally describing some aspects of the physical and social world around us for the purposes of understanding and communication*."

A conceptual model serves four roles in developing domain understanding [21]: (1) aiding a person's own reasoning about a domain, (2) communicating domain details between stakeholders, (3) communicating domain details to systems designers, and (4) documenting the domain for future reference. Viewed from this perspective, conceptual modeling can be seen as a process whereby individuals reason and communicate about a domain in order to improve their common understanding of it.

Industry studies have revealed high failure rates in information systems projects, and have suggested leading reasons for project failures related to requirements specification that include: (1) lack of user input, (2) incomplete or unclear requirements, and (3) changing requirements [10, 18]. Despite the best efforts of researchers and practitioners, it remains less than obvious how to perform requirements analysis well. As Brooks [6, p. 182] notes " *I believe the hardest part of building software to be the specification, design, and testing of this conceptual construct, not the labour of representing it and testing the fidelity of the representation. We still make syntax errors, to be sure; but they are fuzz compared to the conceptual errors in most systems*." In this respect, as CMTs can help formalize domain understanding, they have the potential to improve the requirements analysis process and thus contribute to project success.

The recognition of the role of conceptual modeling has led to the creation of a large number of CMTs [1, 8, 30]. However, this brings about the question of how to evaluate the performance of CMTs [13, 30, 36]. In this paper, we propose a framework for the empirical evaluation of CMTs. The framework can be used to suggest a common terminology, categorize and compare existing work on empirical evaluation, and identify areas where more work needs to be done.

Section 2 presents some general concepts of CMTs and their underlying grammars. Section 3 proposes a general framework for the empirical evaluation of CMTs. This framework is then used in Sect. 4 to classify current literature on the empirical evaluation of CMTs. Section 5 discusses theoretical guidance for the evaluation of CMTs. Section 6 surveys specific ways to measure the various constructs. Finally, Section 7 suggests possible uses of the framework.

2 Understanding conceptual modeling grammars

Before moving more deeply into the foundations for comparison, it is important to understand what is being compared when alternative modeling methods are considered. In our view, the comparison of methods based simply on specific representations (models or diagrams) is shortsighted. Experience demonstrates that a system model drawn by different individuals using the same modeling technique might yield strikingly different representations. Thus, rather than focusing on the representations, emphasis should be placed on the underlying modeling language.

A conceptual modeling language comprises a set of constructs (often represented by graphic symbols) and rules for combining the construct to create representations. This set of constructs and rules are referred to as a conceptual modeling *grammar* [34–36]. The data flow diagramming grammar, for example, comprises these constructs: process, data flow, data store, and external entity. The grammar includes rules such as ''Every dataflow must begin or end in a process.'' The choice of constructs and rules in conceptual modeling techniques reflects the nature of domains to be represented, as well as trade-offs between expressive power and completeness on the one hand and simplicity on the other.

Taken together, the constructs and rules can be used to form meaningful *scripts* (specific models), which are collections of ''statements'' created by using the grammar. For example, scripts created by the data flow diagramming grammar are data flow diagrams (DFD). A script is the product of the conceptual modeling process [36] while a grammar is used for creating a description of a domain. The importance of distinguishing between grammars and scripts is to further clarify what is being compared across alternative CMTs. If we compare scripts, then we are providing comparative information about a single-modeled domain. If we compare grammars, then we are seeking comparative information about the ability of the grammar to model any domain (of the type for which the grammar is designed). Thus, to compare CMTs we claim researchers should focus on the grammars rather than the scripts made about a particular domain. Accordingly, in this paper we address grammar evaluation.

2.1 Grammars and ontologies

Conceptual modeling techniques focus on modeling real-world (application) domains. In that respect, they are related to ontologies. An ontology is a set of concepts and premises about what can exist and happen in (a certain domain of) the world. More generally, the field of ontology is concerned with the way humans describe the world around them. A major difference between CMTs and ontologies is that the models resulting from CMTs are not intended to be complete descriptions of the domain, but rather abstractions representing useful aspects of a domain for the purpose of information systems development. However, to represent domains, the constructs and rules of the CMT grammar should reflect some ontological commitment to the nature of the domains being modeled.

2.2 Previous empirical comparisons of grammars

Before outlining our proposed framework, we discuss previous empirical studies comparing CMTs. These studies are presented in Table 1 (expanding on a table from [36]). As can be seen in the table, although the number of studies has been relatively small [33] they employed quite a variety of instruments and measurement constructs. The first empirical comparisons [7, 29, 31] focused on the comprehension of elements within a diagram. This approach assumed that grammars with the ability to represent more information would eventually transfer more information to the user. Consequently, attention was mostly placed on the diagram and the symbols provided by the grammar.

Yadav et al. [38] were the first to separate the product of modeling (usually a script) from the process of understanding the model. In their empirical comparisons, both Jarvenpaa and Machesky [17] and Batra et al. [4] considered, in addition to the product of modeling, the effort required to construct the model. The primary focus remained, however, with the product of the conceptual modeling process, which was a diagram.

In the works of Vessey and Conger [33], Siau [32] and Kim et al. [20] the focus shifted from observing the product to observing the process of understanding. To capture process information, verbal protocols of participants engaged in a model-related task were observed. The ''process'' could be viewed as the cognitive activity involved in creating or understanding a diagram. The cognitive activity involves searching, integrating, and recording, either in a script or storing in memory, the

Table 1 Summary of previous research

Authors	Comparison	Task	Constructs
Brosey and Schniederman [7]	Relational and hierarchical models	Interpret diagrams	Comprehension, query accuracy
Ramsey et al. [31]	Flowcharts vs. program design language (PDL)	Create diagrams	Design style, level of detail, constructs, model correctness
Nosek and Ahrens [29]	DFD, task-oriented menus	Interpret diagrams	Accuracy of understanding
Yadav et al. [38]	DFD and IDEF0	Create diagrams	Semantic and syntactic correctness, completeness, ease of use, learning
Jarvenpaa and Machesky [17]	Logical data structure (LDS) and relational data structures (RDS)	Create diagrams	Model correctness, interpretation accuracy, ease of learning, approach
Batra et al. [4]	Relational data model (RDM) and extended entity relationship (EER) model	Create diagrams	Model correctness, ease of use
Vessey and Conger [33]	Object (OOA), process (DFD), and data (ERD) methods	Create diagrams	Breakdowns, procedural or declarative knowledge
Kim and March [19]	EER diagram and Nijssen information analysis method (NIAM)	Interpret or create diagrams	Users: discrepancies and comprehensionAnalysts: syntactic and semantic errors
Agarwal et al. [2]	DFD and object-oriented (structure) diagrams	Create diagrams	Structure, behavior, overall quality, Jaccards similarity
Wang [37]	DFD vs. object-oriented diagrams	Create diagrams	Syntactic and semantic correctness
Siau [32]	Cardinality constraints and noun vs. verb description	Interpret diagrams	Number of "chunks" correctly interpreted
Agarwal et al. [3]	DFD and object-oriented (structure) diagrams	Interpret diagrams	Comprehension, qualitative pattern analysis
Bodart et al. [5], Gemino [14]	Optional and mandatory properties in ERD	Interpret diagrams	Items remembered, comprehensions, number of solutions provided
Kim et al. [20]	Object modeling technique (OMT), unified modeling language (UML)	Interpret multiple diagrams	Number of correct problems identified, number of diagram transitions, returning episodes
Burton-Jones and Meso [9]	UML decomposition	Interpret diagrams	Comprehensions, number of problem-solving solutions provided, semantic recall
Gemino [15]	OMT using narration and animation	Interpret diagrams	Comprehension, number of problem-solving solutions provided, recall

information about the domain. In the case of a model understanding task, the outcome is cognitive and therefore not directly observable. Hence, verbal protocols were used. The focus in these studies shifted from the diagram itself to cognition about the diagram as the ultimate product in the process.

More recently, researchers have investigated measuring understanding using problem-solving tasks that require participants to reason about the domain being represented [5, 9, 14–16]. The distinction between earlier studies and recent ones using problem-solving (also known as transfer) tasks is the shift from understanding the diagram itself to the mental model of the domain developed by using the diagram. The problem-solving studies recognize that the ultimate objective of conceptual modeling is to communicate information about a domain to a relevant stakeholder. The measurement, therefore, focuses on the "deep" understanding developed about the domain when a person views a diagram, and not the elements of the diagram itself. The measurement of this understanding is accomplished by asking participants to reason about the domain being described rather than to recognize the elements in a model.

3 A general framework

The discussion of the relatively small number of previous empirical comparisons indicates two things. First, conclusive empirical comparisons of CMTs are difficult to create. Second, the diversity of constructs and procedures suggests that no widely accepted instruments, procedures, or constructs for comparative performance exist. This diversity impairs the ability to compare and contrast empirical results and to summarize literature about the comparison of CMTs. To address the above concerns, we propose a framework for categorizing empirical work in the area of CMTs. The framework intends to classify empirical work with respect to the dimensions researchers can choose in designing experiments and measuring relevant outcomes. The framework is based on two main types of dimensions in the comparison of CMTs: (1) the dimension of affecting factors, and (2) the dimension of affected variables (outcomes) for CMTs. A discussion of these dimensions is provided below.

3.1 Affecting factors

The affecting dimension includes factors that can influence the outcome of the conceptual modeling process. The states of the corresponding variables reflect the decisions a researcher makes about the phenomena to be studied. In experimental design, some of these variables will be manipulated while others might serve as controls. Hence, they will be manifested in an experimental design via the task, definition of independent variables, and definition of controls.

The knowledge construction model proposed by Gemino and Wand [16] provides a guide for considering what to include in this dimension. This model suggests three antecedents to the process of communicating conceptual modeling information: (1) the content to be delivered, (2) how the content is presented, and (3) the characteristics of the person participating in the communication. In addition, the task must also be considered. Each of these areas represents a set of variables that can be controlled by the empirical researcher as discussed below.

The *contents* variable refers to the type of information contained within the cases. For instance, some cases may be focused on process elements while others are focused on static data structure. In choosing the initial contents for the study, it is important to identify the appropriate information and to use cases that provide information relevant to the comparison.

The *method of presenting* the material relates to the nature of the grammar and to the way scripts are presented. The nature of the grammar provides the primary focus for empirical comparisons of CMTs. Wand and Weber [36] suggest several dimensions to consider. These include (1) the choice of grammar constructs to consider, (2) the nature of comparison (within or between grammars), (3) rules regarding the use of the grammar and how it is applied, and (4) the way the script is presented (text, graphics, narrated, animated, etc.). The choice of grammar often represents the main difference being compared. When different grammars are compared this represents an intergrammar comparison. In some studies, however, the same grammar is used, and different rules within the grammar are applied. This would represent an intragrammar comparison. Together, these dimensions, along with the media used to present the information, describe the way in which the participant will be presented with information about a domain or creating a model of the contents provided.

In addition to the content and presentation method, it is important to consider the characteristics of participants using the CMT. These characteristics might include modeling or domain experience, cognitive style, and other relevant characteristics.

Finally, we must also consider the task that the grammar will be used for; either a model interpretation (''reading'') task or model creation (''writing'') task. Table 2 summarizes the variables affecting outcomes in the conceptual modeling process. These variables comprise the dimensions to consider in the design of empirical comparisons of CMTs.

3.2 Affected variables (outcomes) dimension

The affected variables dimension comprises observable outcomes of conceptual modeling tasks. These variables are the source of dependent measures in empirical comparisons. Affected variables will be the phenomena that reflect possible quality measures of the grammar. For example, the correctness of a model is an affected variable. Two categories of affected variables are proposed: (1) the focus of observation and (2) the criterion for comparison.

The focus of observation indicates whether the emphasis in measurement has been placed on observing: (1) the process of using the CMT, or (2) the product resulting from the use. Of course, it is possible to consider both in an empirical study. The focus of observation will depend on the objectives and empirical design outlined by the researcher. For example, when comparing grammars on their ability to create models, the relevant product of this process is the final script developed. By contrast, in interpretation of an existing script, the final product is the understanding created in the viewer's cognition. These two different products would require different measures but both would place the focus on the product of the CMT process.

The criteria for comparison indicates whether the researcher is comparing (1) the effectiveness of the CMT or (2) the efficiency with which the CMT can be used. For example, a focus could be placed on how effective the CMT is in generating quality scripts, regardless of how long that process may take. This would represent a focus on effectiveness as opposed to efficiency. A study can refer to both criteria. The criteria for comparison are shown in Table 2.

It is important to note that some of the phenomena discussed above might in turn affect other outcome variables. For example, the time to complete a modeling task might depend on the grammar, but might also affect the quality of the outcome. Such variables will be considered as moderators. A moderator can be used to either provide an additional test for the effect of the independent variables, or be tested for its impact as a moderator.

The specific effectiveness or efficiency measures used in a study will depend on the nature of the task. For example, the outcome of script creation is a specific script that can be evaluated for correctness. In contrast, the outcome of script interpretation is tacit—understanding created in the model viewer's cognition. These dependencies are shown in Table 3.

4 How the framework works: classifying the literature

To demonstrate the use of our proposed framework, we now reclassify the empirical research presented earlier (Table 1). To illustrate the classification, we discuss briefly a few of the mentioned works. Table 4 summarizes the classification of the works listed in Table 1 and shows how the various experiments described in the literature map into the various dimensions noted above.

Consider first the article by Vessey and Conger [33]. The study focused on an intergrammar comparison between DFD (process), entity relationship diagrams (structure), and object-oriented analysis (process and structure). Accordingly, the content included information about both process and data structure. Rules within grammars were not varied. The media was paper-based using text and graphics and was not varied between

Table 2 Description of affecting variables and dimensions

Affecting variables and dimensions	Description	Examples and notes
Variables		
Content	The case material used in the study	Process models Data structure descriptions Object-oriented descriptions
Grammar(s) constructs	The specific constructs available in the grammar and their possible relationships. Identify single grammar vs. multigrammar	For example, a grammar containing explicit relationships versus one that does not Single: ERD, multiple: UML
Nature of comparison (inter- vs. intragrammar)	Comparing different ways of using a single grammar (intragrammar) or comparing two or more grammars (intergrammar)	Inter: ERD vs. class diagram Intra: ERD with and without optional relationships
Use of grammar(s)	Rules for using the grammar constructs	For example, use of mandatory properties only versus using optional properties in ERM
Medium of content delivery	The media used to present content developed through the grammar	Text, graphics, annotation, narration, animation, and other combinations
User characteristics	Individual traits that can affect outcomes such as modeling experience (novice or expert)	Level of modeling expertise Level of domain expertise
Task	Interpretation/"reading" Creation/"writing"	Either reading from an existing script or creating a new script
Dimensions		
Focus of observation	Product of task Process of performing task	The product of a model creation is tangible—a model The product of model interpretation is tacit—the domain understanding created in the viewer's mind
Criterion for comparison	Effectiveness measures: how well CMT assists in accomplishing its objectives Efficiency measures: resources needed to use the CMT	A variety of measures can be used Choice of measures will depend on the focus of observation

Table 3 Relationships between focus of observation and criterion for comparison

Criterion for comparison	Focus of observation			
	Script creation		Script interpretation	
	Product	Process	Product	Process
Effectiveness	Physical model (script)	Creating a model	Cognitive model in viewer	Understanding the model
Efficiency	Effort required to create a script		Effort required to interpret a script and develop domain understanding	

treatment groups. The task was diagram creation and participants were novice modelers. The focus was placed on the effectiveness of the process of creating diagrams. Effectiveness was measured by considering the number of breakdowns, or errors that occurred in the process of creating the script. These errors were identified in verbal protocols collected as participants constructed the scripts.

Next, we consider the work of Bodart et al. [5]. The study focused on an intragrammar comparison between two options for drawing entity relationship diagrams. One version used optional properties while the other allowed only mandatory properties with subtypes. The contents, therefore, focused primarily on data structure. Since this was an intragrammar comparison, the differences were not in constructs, but in the rules used for creating diagrams. The media was paper-based using text and graphics and was not varied between treatment groups. The task was model interpretation and the subjects were novice modelers. The focus was placed on evaluating the cognitive model created by viewing the diagrams. Product (namely domain knowledge) effectiveness was measured in several ways using recall, comprehension, and problem-solving instruments. In addition, the efficiency of the script interpretation process was assessed using instruments for ease of use.

The paper by Kim et al. [20] provides a final example. The study focused on a multi-grammar comparison using the UML and OMT. To reflect this, the contents focused on both the data structure and the process. Since this was an intergrammar comparison, the rules within grammars were not varied. The media was paper-based using text and graphics and was not varied between treatment groups. The task was model interpretation, and subjects were novice modelers. The focus was placed on evaluating the cognitive model created by viewing the diagrams. Process effectiveness was measured via verbal and action protocols by considering the number of correct problems identified. Process efficiency was assessed using the number of transitions and returning episodes made between diagrams.

5 Theoretical guidance for framework

The framework above is intended to provide guidance in designing and evaluating empirical comparisons of CMTs. This section describes some potential theoretical guidance in developing these comparisons. The discussion of the theoretical guidance focuses on the four elements of the knowledge construction model: content, presentation method, individual characteristics, and task.

5.1 Theory to guide contents: informational and computational equivalency

When comparing scripts created with different grammars, the outcome might be confounded because one model might contain more information than the other, or one may be easier to understand than another. This issue can be addressed by considering the concepts of *informational* and *computational equivalency*. Larkin and Simon [22, p. 67] define informational equivalence thus: " *two representations are informationally equivalent if all information in one is also inferable from the other and vice versa*." The need to control for informational equivalence has been raised as a concern by Agarwal et al. [3]. A way to test if representations created with different conceptual modeling grammars are informationally equivalent is to ask participants questions related to the contents of the presented script. Referring to our affecting variables classification, such a task entails using some aspects of content as experimental controls.

Informational equivalence of scripts created with different grammars does not mean that the alternative grammars are computationally equivalent. Larkin and Simon [22, p. 67] define computational equivalency thus " *two representations are computationally equivalent if they are informationally equivalent and, in addition, any inference that can be drawn easily and quickly from the information given explicitly in the one can also be drawn easily and quickly from the information given explicitly in the other, and vice versa*." This suggests that while two grammars may be informationally equivalent, one of the grammars may be easier to use, or more efficient at getting the point across, and, therefore, the grammars would not be computationally equivalent. Carefully considering computational and informational equivalence can help researchers focus on where the differences between techniques are expected to arise.

5.2 Theory for defining presentation methods: ontology and metamodels as benchmarks

When comparing grammars, one needs to know if any significant differences in the grammar are to be expected. Addressing differences in CMTs via their grammars can

Table 4 Classifying the literature using the framework

Authors	Content	Nature	Rules	Media	Subject	Task	Observation focus	Comparison criteria
Brosey and Schniederman [7]	Structure	Intergrammar	Held constant	Held constant	Novices	Interpret diagrams	Product	Effectiveness
Ramsey et al. [31]	Process	Intergrammar	Held constant	Held constant	Experts	Create diagrams	Product	Effectiveness
Nosek and Ahrens [29]	Process	Intergrammar	Held constant	Held constant	Novices	Interpret diagrams	Product	Effectiveness
Yadav et al. [38]	Process	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Product and process	Effectiveness and efficiency
Jarvenpaa and Machesky [17]	Structure	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Product and process	Effectiveness and efficiency
Batra et al. [4]	Structure	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Product and process	Effectiveness and efficiency
Vessey and Conger [33]	Process and structure	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Process	Effectiveness
Kim and March [19]	Process and structure	Intergrammar	Held constant	Held constant	Novices and experts	Interpret or create diagrams	Product	Effectiveness
Agarwal et al. [2]	Process and structure	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Product	Effectiveness
Wang [37]	Process	Intergrammar	Held constant	Held constant	Novices	Create diagrams	Product	Effectiveness
Siau [32]	Structure	Intragrammar	Varied	Held constant	Novices	Interpret diagrams	Product and process	Effectiveness and efficiency
Agarwal et al. [3]	Process and structure	Intergrammar	Held constant	Held constant	Novices	Interpret diagrams	Product	Effectiveness
Bodart et al. [5], Gemino [14]	Process and structure	Intragrammar	Varied	Held constant	Novices	Interpret diagrams	Product and process	Effectiveness and efficiency
Kim et al. [20]	Process and structure	Intergrammar	Held constant	Held constant	Novices	Interpret multiple diagrams	Process	Effectiveness and efficiency
Burton-Jones and Meso [9]	Process and structure	Intragrammar	Varied	Held constant	Novices	Interpret diagrams	Product and process	Effectiveness and efficiency
Gemino [15]	Process	No difference	Held constant	Varied-narration, animation	Novices	Interpret diagrams	Product and process	Effectiveness and efficiency

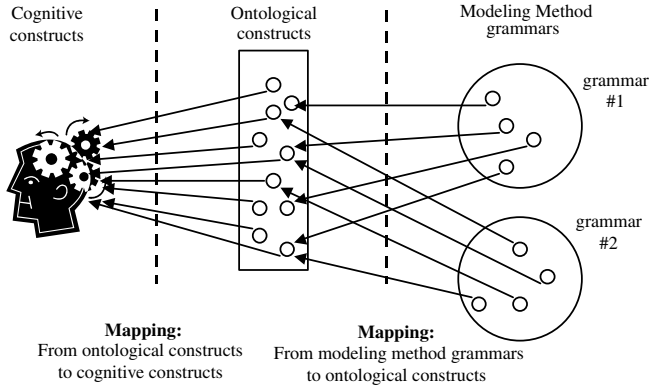

Fig. 1 Mapping modeling grammars, ontological constructs, and cognitive constructs

enable a theoretical approach to empirical comparison. Grammars can be evaluated by comparing them to a grammatical benchmark that contains a set of basic constructs expressing what should be modeled. If the benchmark is a set of generic constructs, it is referred to as a metamodel [30]. If the benchmark is based on a set of beliefs on what might exist and happen in the modeled domain, it is called an ontology. An ontology can be specific to a certain domain (e.g., to health care) or general (as Bunge's ontological model for systems [34]).

Wand and Weber [34] proposed evaluating modeling methods by analyzing the mappings between the constructs of a modeling grammar and a set of fundamental ontological constructs. In this way, an ontology is used as a benchmark to evaluate constructs within a modeling grammar. Further, since ontology is intended to formalize the way the real world is modeled by humans, the ontological constructs might also be viewed as related to constructs in human cognition. If the ontological constructs map well to cognitive constructs, then modeling grammars that map well to the ontological constructs, by extension, also map well to human cognition. This mapping is depicted in Fig. 1.

The expressiveness of a grammar is its ability to generate scripts that capture information about a modeled domain. A grammar's expressiveness can be evaluated by examining the mapping that exists between the benchmark concepts and the grammar's constructs. Deficiencies in the grammar might then, in principle, be revealed in an examination of this mapping. A benchmark concept for which there is no matching grammar construct, for example, would indicate that the grammar is incomplete.

Ontological analysis can be accomplished independent of any consideration of participants in the CMT process. However, even if we know a certain grammar is more ontologically expressive than an alternative, we may have no theoretical line of reasoning to establish it. For example, while one modeling grammar may be very expressive and hence superior in a grammar-based comparison, a representation created with that grammar might be overly complicated, resulting either in frequent mistakes in applying the grammar or little impact on the understanding developed by the person viewing a script. If a theoretical link cannot be established between the features of a grammar and increased human performance, the differences outlined in an ontological analysis may not be important in regard to overall CMT performance.

5.3 Theory on individual differences: using cognitive theories for generating hypotheses

Gemino and Wand [16] claim that observations cannot provide explanations of why observed differences exist. To establish more general principles for effective design of modeling techniques, we need theories to hypothesize why differences between grammars might matter. Theoretical considerations can be used both to guide empirical work and to suggest how to create more effective grammars.

The focus on cognitive theory is natural when considering CMTs. For example, when "reading" a model, the outcome is implicit—in the reader's mind. Thus, the evaluation of a grammar on the basis of its ability to convey information should take into account cognitive considerations. Specifically, the evaluation should recognize the difference between a representation and the resulting cognitive model developed by the viewer. It follows that grammar evaluations should take into account the viewer's information processing activity. Moreover, since cognitive processes cannot be evaluated except by observing tasks performed by humans, evaluating the performance of grammars in interpretation tasks is necessarily empirical.

A review of the literature shows that the focus in early comparisons of CMTs was on empirical observations [17, 33]. Batra et al. [4] utilized a framework from cognitive theory and recent work has incorporated cognitive theory in the comparison of grammars. For example, Bodart et al. [5] and Burton-Jones and Meso [9] have used the theories of semantic memory from Collins and Quillan [11], while Gemino [15] relies on the cognitive theory of multimedia learning [24] to generate hypotheses.

5.4 Theory guiding choice of task: Norman's Theory of Action

There are two possible functions in any modeling exercise: script interpretation ("reading"), and script creation ("writing") [36]. Norman's theory of action [28] is useful in considering both sides of these functions. The theory is depicted in Fig. 2 and is best understood by considering two persons who are involved in modeling. The model creator has a goal to communicate his/her understanding by representing it in a model. A model viewer has the goal of understanding the domain being represented by interpreting the model. Norman's [28]

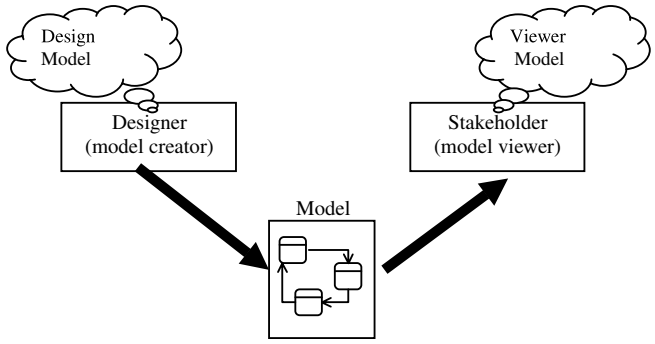

Fig. 2 Norman's [28] theory of action as applied to IS modeling.

theory clearly distinguishes between a person's understanding, held in cognition, and the model that is expressed physically. Discrepancies between a person's understanding of the system, and the model used to represent the system leads to issues in both creating and interpreting the diagram. Norman [28] labels these discrepancies the "gulf of execution" and the "gulf of evaluation."

A gulf of execution forms when discrepancies exist between the conception of the domain from a model creator's viewpoint and the model of the domain. In this case, the model creator is frustrated because the model does not represent his/her view of the domain accurately. These discrepancies may occur due to: (1) constraints on the expressiveness of the modeling technique, (2) lack of skill of model creator, or (3) confusion in the model creator's conception of the domain.

A gulf of evaluation occurs when a discrepancy exists between a model viewer's cognitive model of the domain and the script representing the domain. In this case, the model viewer is not getting the correct idea, because there is a difference between what the script shows and what the viewer understands. The discrepancy may occur when: (1) the user misinterprets the diagram due to lack of experience with the method, (2) the user develops a different conception of the domain being represented from the one conveyed by the diagram, or (3) there exists ambiguity within the script itself.

Viewed through Norman's [28] theory, the modeling process can be seen as an effort to bridge the gulfs of execution and evaluation in order to bring the model creator's and the model viewer's conceptions of the system (the design model and the viewer model) closer together. Norman's [28] theory is useful because it provides a clear separation between the domain represented in a model and the understanding of that domain as represented in the cognitive model of either the designer or the viewer. These insights suggest that researchers need to clarify whether their focus is placed on the gulf of execution (where the models are created) or the gulf of evaluation (where the models are interpreted). More significantly, Norman's [28] model suggests that a complete evaluation must refer to how the CMT bridges both these gulfs.

6 Guidance for measurement

The discussion above illustrates how the literature for empirical research can be categorized in regard to the state of affecting variables and affected (outcome) variables. This categorization can be useful in determining research questions and hypotheses. However, to design empirical comparisons, one needs to determine the levels or values that affecting variables can have, and how to measure the affected variables. This section describes the various measures available for researchers in developing empirical designs for comparing the CMTs and the instruments for measuring relevant outcomes. The discussion has been organized into two sections: one focusing on the affecting dimensions, the other focusing on the affected dimensions.

6.1 Choosing levels for affecting variables

The choices available to researchers in designing empirical comparisons are provided in Table 5. Table 5 builds on the constructs identified in Table 2 and provides additional information regarding the possible states of the constructs and some suggestions on their use.

We begin with the content of the cases used. The contents should provide the appropriate type of information according to the main types of modeling constructs available in tested grammars. These types can be broadly categorized as process-oriented, structure-oriented, or object-oriented (where process and structure elements are "encapsulated" in one construct). For example, testing process models with the data of structure-oriented cases would not generate interesting comparisons. The importance of the cases used in the procedures cannot be overstated. A common reaction to CMT comparisons is to suggest that observed results are simply sensitive to the case used and that a different case would result in a different outcome. These arguments are difficult to rebut. Moreover, the subjects' domain experience can play a part in the conceptual modeling process, further increasing the dependence on the case [26]. Several researchers have addressed this issue by performing tests with more than a single case [5, 14]. While using more than one case increases the costs associated with the experiment, consistent results across two or more cases lead to stronger validity. Furthermore, consistent results across research teams using identical cases can provide an even higher level of validity and should be encouraged. Thus, the use of "standard" cases should be encouraged.

We have discussed earlier the need to clarify differences between chosen grammars with an appropriate grammar benchmark (ontology or metamodel). We have also discussed the issue of intra- and intergrammar comparisons. Note that when an intragrammar comparison is developed, differences in the rules of how to

Table 5 Summarizing levels for affecting variables

Affecting variable	Description	Levels (states)	Suggestions
Content	The case material used in the study	Process	Use standard cases
		Structure	Use multiple cases
		Encapsulated (OO)	
Grammar(s) constructs	The specific constructs available in the grammar and their possible relationships	With construct	Use theory to clearly define differences
		Without construct	
		Multiple grammar	
Nature of comparison inter- vs. intragrammar)	Inter: ERD vs. class diagram	Intragrammar	Test for informational equivalence
	Intra: ERD with and without relationship symbol	Intergrammar	
Use of grammar(s)	Rules for using the grammar constructs. e.g., ERD with and without optional relationships	With rule	Normally a variable in intragrammar comparisons
		Without rule	
Medium of content delivery	Media used to present content developed through grammar	Text, graphics, narration, animation, other	Test for informational equivalence
User characteristics	Individual traits that can affect outcomes: eg. modeling experience (novice or expert)	Novice	Measure
		Expert	Model expertise
		Cognitive style	Domain expertise
		Others	Randomize subjects
Task	Interpretation/"reading" or creation/"writing"	Interpretation	Note impact on relevant outcome measures
		Creation	

apply the grammar often become the variable of interest in the study. It is important to note in creating either inter- or intragrammar comparisons, that the notion of informational equivalency will be central to the usefulness of the results. If the two treatments provide significantly different levels of information, the results for the empirical test may be of little interest because the content differences, and not grammatical differences, might drive the results. Information equivalency is particularly hard to attain in intergrammar comparisons. The problem of information equivalency is also of significance when comparing the different media used to present models. In general, careful controls need to be put into place to make sure significant information advantages are not provided to one of the treatment groups.

Perhaps the most important consideration in developing studies is the relevant participant pool. Many observers have suggested that using novices negates external validity of the results. Prior empirical work has recognized that both the knowledge of the modeling method and the knowledge of the application domain are important considerations in choosing potential participants. These two dimensions reflect several types of participants in the conceptual modeling phase of scope definition. These types can be described using the grid provided in Fig. 3.

The classification in Fig. 3 can be used to identify the project roles in conceptual modeling processes. For example, individuals in Quadrant I (high domain knowledge, but low modeling knowledge) might represent stakeholders such as system users or managers, whereas individuals in Quadrant II (high domain and high modeling knowledge) represent experienced analysts who have worked in the particular application domain.

While this classification is useful in describing roles in the system analysis process, it represents only stereotypical roles for individuals in projects. In reality, there is a wide degree of variation in the level of domain and modeling knowledge even within these general roles. For example, a "user" may have intimate knowledge of his or her area of functional expertise, but little or no knowledge of activities in other parts of a business. An "analyst" may be experienced with a given modeling technique, but possess little or no knowledge of other

Fig. 3 Types of study participants

modeling techniques or of the application domain. Project groups could be formed using stakeholders from different divisions, who know their own area, but have little or no knowledge of the overall application domain. In addition, external consultants may vary in their knowledge of the domain or modeling methods. In summary, neither stakeholders nor analysts can always be expected to be domain experts in the all areas of the business pictured in a conceptual modeling diagram.

In designing an empirical comparison of modeling techniques, it is also important to consider the challenges faced when using participants with high levels of domain and/or modeling method knowledge. For example, analysts typically possess a high level of modeling knowledge of the particular technique they already use [26]. Thus, any comparison of techniques using analysts as subjects must first overcome the bias towards familiar techniques. Batra et al. [4] suggested that novices are representative of typical information systems users who are involved in the analysis and design process in end-user computing. In summary, it is important to recognize that the use of either ''experienced'' analysts or ''real'' stakeholders who are very familiar with the application domain, while seemingly providing more realistic conditions, might create substantial difficulties in an experimental study.

6.2 Measuring affecting variables

The summary of empirical studies provided earlier in Table 1 indicated a wide variety of constructs used in measuring outcomes from the comparison of CMTs. Our discussion of these constructs highlighted the distinction between the product of the CMT process and the process itself. In presenting the framework, we argued that these outcome measures could be summarized into two criteria for comparison, effectiveness, and efficiency. The argument further introduced the different products resulting from either the creation or interpretation of scripts. In this section, we suggest the various outcome measures that can be used for outcomes noted above. This categorization and the related measures are summarized in Table 6.

In regard to the task of script creation, the product is an actual (physical) script. Measurements of the effectiveness of these scripts can be based on a variety of constructs reflecting expert ratings of the diagrams. A common method is to score models based on different types of errors. To address concerns about subjectivity, Agarwal et al. [2] suggested the use of Jaccards simplicity, which assigns a similarity value between 0 and 1 that expresses the degree of overlap between two scripts, A and B, as the proportion of the overlap from the whole. In measuring the effectiveness of the process, Vessey and Conger [33] introduced the concept of ''breakdowns'' which represent errors that are identified from the verbal protocols of sessions where participants create scripts. The efficiency of the script creation process has been assessed by considering the time taken to create scripts, along with the perceived ease of use and ease of learning reported by participants. Finally, another subjective measure of effectiveness is the confidence subjects convey as to the quality of their models. While this confidence applies to the created model, we view it as reflecting the comfort level generated throughout the model creation process.

The measurements that can be applied to script interpretation tasks are less obvious than those in script creation. The product of script interpretation is the cognitive model developed by the script viewer. The cognitive model cannot be directly observed. Rather, what can be assessed is the performance of tasks based on cognition. Several types of performance measures have been suggested for capturing the understanding a participant develops when interpreting a diagram. Most interpretation studies have employed comprehension tests [3, 17]. A comprehension test is typically comprised of a set of questions about elements in the script being considered. Such a test can be conducted when a model is available, reflecting model understanding, or after it has been removed, reflecting recall.

Table 6 Measures for affected (outcome) variables

Criterion for comparison	Focus of observation			
	Script creation		Script interpretation	
	Product	Process	Product	Process
Effectiveness (measures)	Physical model Expert rating Accuracy Correctness Detail Completeness Quality Discrepancies Jaccards similarity	Creating model Breakdowns Errors Confidence in correctness (subjective)	Cognitive model in viewer Comprehension Items remembered Recall Problem solving Total Acceptable Semantic recall (Cloze test)	Understanding the model Errors in interpreting Confidence in correctness (subjective)
Efficiency (measures)	Ease of use (subjective) Ease of learning (subjective) Measures: elapsed time		Ease of use Ease of learning Elapsed time	

An important distinction has been raised between the comprehension of diagram elements and the understanding of the domain that develops as a result of viewing scripts. If one considers that the objective of the conceptual modeling process is to further domain understanding, then comprehension questions focused on diagram elements do not measure the desired product of the conceptual modeling process. It is for this reason that recent studies [5, 9, 14, 15] have focused on measuring the understanding of the domain using problem-solving questions. While script comprehension is a necessary condition for understanding, significant additional cognitive processing may be required to develop an understanding of the domain. Researchers should therefore consider measuring not only the comprehension of script (typically a diagram) elements but also domain *understanding* created by cognitive processing of script information.

A procedure for assessing understanding is described in Mayer [23, 24]. In his experiments, Mayer [23, 24] created two treatment groups: one was provided with a text description accompanied by a diagram (the "model" group), the other was provided with only the text description (the "control" group). After participants viewed the materials, the materials were removed and the participants were asked to complete a comprehension test and a set of problem-solving questions. The comprehension task included questions regarding the attributes of items in the description or the relationships between them. For example, participants were given information about a car's braking system. The comprehension test included questions such as: "*What are the components of a braking system*," whereas problem solving included questions such as: "*What could be done to reduce the distance needed to stop*." The problem-solving questions required participants to use the mental representation they had developed to suggest answers for which information was not directly available in the diagram. It is premised that the more answers participants provide, the more sophisticated is their (cognitive) model of the domain and the higher is the level of understanding they have developed.

A further possibility to measure the product of script interpretation is provided by semantic recall based on fill-in-the-blanks (Cloze) test. In this test, participants are asked to fill in the blanks in a paragraph after the model has been removed [14, 9]. The higher the semantic recall, the higher the level of understanding of the domain being modeled. In such tests, attention should be paid to identifying synonymous terms participants might use when filling a blank.

As with script creation, the process of script interpretation can be "opened" by using verbal protocols (or other process-tracing methods) to count the number of errors during the interpretation process. This number can be used to measure the effectiveness of script interpretation. In addition, its effectiveness can also be tested by asking subjects about the confidence in their answers.

The efficiency of the interpretation process can be evaluated by measuring the time taken to complete the interpretation. As well, subjective ease of use and ease of learning associated with the grammar can be assessed using instruments such as the ease of use instrument provided by Moore and Benbasat [25] or Davis [12]. It should be noted that, to date, perceived ease of use has not been found to be a significant contributor to effectiveness outcome scores.

7 Summary

Conceptual modeling grammars are used in requirements analysis to promote understanding of the application domain and the communication of this understanding. Given the importance of the requirements analysis phase, the choice of a conceptual modeling grammar can affect the outcome of a systems development project.

Given that CMTs are important, definitive information about CMT performance can only be obtained by empirical methods. However, existing empirical works tend to differ considerably. To be able to compare different works, we have proposed a framework for reasoning about experimental comparisons of conceptual modeling grammars. The framework identifies the main dimensions that have to be considered in setting up the research questions and hypotheses. Specifically, it suggests the possible choices of experimental tasks, independent variables, controls, mediating variables, and dependent variables.

As the framework provides a common set of dimensions to categorize existing work on empirical evaluation, it can be used as a source of common terminology: to compare existing work, to identify where gaps exist, to identify where more work can be done, and to support the formulation of research hypotheses.

As more empirical work on conceptual modeling grammars appears, it is possible that more will be known about the level of categorical variables and the reliability of the measures for interval variables, notably those related to grammar performance. However, it is also possible that more relevant variables will be proposed. Thus, we believe the framework will evolve over time. Yet, we hope the present framework has already contributed towards the creation of a cumulative tradition of empirical work on conceptual modeling grammars.

Acknowledgements This work was supported in part by grants from the Social Sciences and Humanities and Natural Sciences and Engineering Research Councils of Canada.

References

1. Avison DE, Fitzgerald G (1995) Information systems development: methodologies, techniques, and tools, 2nd edn. McGraw-Hill, London

2. Agarwal R, Sinha A, Tanniru M (1996) Cognitive fit in requirements engineering: a study of object and process models. J Manag Inf Syst 13(2):137–162
3. Agarwal R, De P, Sinha AP (1999) Comprehending object and process models: an empirical study. IEEE Trans Softw Eng 25(4):541–556
4. Batra D, J Hoffer, R Bostrom (1990) Comparing representations with relational and EER models. Commun ACM 33(2):126–139
5. Bodart F, Sim M, Patel A, Weber R (2001) Should optional properties be used in conceptual modelling? A theory and three empirical tests. Inf Syst Res 12(4):384–405
6. Brooks FP (1998) The mythical man-month: essays of software engineering, Anniversary edition. Addison-Wesley
7. Brosey M, Schniederman B (1978) Two experimental comparisons of relational and hierarchical database models. Int J Man Machine Stud 10:625–637
8. Bubenko JA Jr (1986) Information system methodologies: a research review. In: Olle TW, Sol HG, Verrjin-Stuart AA (eds) Information system design methodologies: improving the practice. Elsevier, North Holland, pp 289–317
9. Burton-Jones A, Meso P (2002) How good are these UML diagrams? An empirical test of the Wand and Weber good decomposition model. In: Applegate L, Galliers R, DeGross JI (eds) Proceedings of the international conference on information systems 2002, December, 2002
10. Chaos (1995) Standish group report on information system development. http://www.standishgroup.com/chaos.html. Accessed 12 Jul 1996]
11. Collins AM, Quillan MR (1969) Retrieval time from semantic memory. J Verbal Learn Behav 8:240–247
12. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technologies. MIS Q 13(3):319–340
13. Floyd C (1986) A comparative evaluation of systems development methods. In: Olle TW, Sol HG, Verrjin-Stuart AA (eds) Proceedings of the IFIP WG 8.1 working conference on comparative review of information systems design methodologies: improving the practice. North-Holland, Amsterdam, pp 19–54
14. Gemino A (1999) Empirical comparison of system analysis techniques. PhD Thesis, University of British Columbia
15. Gemino A (2004) Empirical comparisons of animation and narration in requirements validation. Req Eng J 9(3):153–168. DOI 10.1007/s00766-003-0182-0
16. Gemino A, Wand Y (2003) Evaluation of modeling techniques based on models of learning. Commun ACM 46(10):79–84
17. Jarvenpaa S, Machesky J (1989) Data analysis and learning: an experimental study of data modeling tools. Int J Man Mach Stud 31:367–391
18. Johnson J, Boucher KD, Connors K, Robinson J (2001) The criteria for success. Softw Mag 21(1):s3–s8
19. Kim YG, March S (1995) Comparing data modeling formalisms. Commun ACM 38(4):103–115
20. Kim J, Hahn J, Hahn H (2000) How do we understand a system with (so) many diagrams? Cognitive integration processes in diagrammatic reasoning. Info Syst Res 11(3):284–303
21. Kung CH, Solvberg A (1986) Activity modelling and behaviour modelling. In: Olle TW, Sol HG, Verrjin-Stuart AA (eds) Proceedings of the IFIP WG 8.1 working conference on comparative review of information systems design methodologies: improving the practice. North-Holland, Amsterdam, pp 145–171
22. Larkin J, Simon HA (1987) When a diagram is (sometimes) worth ten thousand words. Cogn Sci 11(1):65–99
23. Mayer RE (1989) Models for understanding. Rev Educ Res 59(1):43–64
24. Mayer RE (2001) Multimedia learning. Cambridge University Press, New York
25. Moore GC, I (1991) Development of an instrument to measure perceptions of adopting an information technology innovation. Inf Syst Res 2(3):192–222
26. Morris MG, Spier C, Hoffer JA (1999) An examination of procedural and object-oriented system analysis methods: does prior experience help or hinder performance. Decis Sci 30(1):107–136
27. Mylopoulos J (1992) Conceptual modeling and telos. In: Loucopoulos P, Zicari R (eds) Conceptual modeling, databases, and case: an integrated view of information systems development, chap 2. Wiley, New York, pp 49–68
28. Norman D (1986) Cognitive engineering. In: Norman D, Draper S (eds) User centered design: new perspectives on human computer interaction. Lawrence Erlbaum Associates, Hillsdale, pp 31–61
29. Nosek J, Ahrens J (1986) An experiment to test user validation of requirements: data flow diagrams vs. task oriented menus. Int J Man Mach Stud 25:675–684
30. Oei JLH, van Hemmen LJ, Falkenberg ED, Brinkkemper S (1992) The meta model hierarchy: a framework for information systems concepts and techniques. Technical Report No. 92-17, Department of Informatics, Faculty of Mathematics and Informatics, Katholieke Universiteir, Nijmegen, pp 1–30
31. Ramsey R, Atwood M, Van Doren J (1993) Flowcharts versus program design languages: an experimental comparison. Commun ACM 26(6):445–449
32. Siau KL (1996) Empirical studies in information modeling. PhD Thesis, University of British Columbia
33. Vessey I, Conger S (1994) Requirements specification: learning object, process, and data methodologies. Commun ACM 37(5):102–113
34. Wand Y, Weber R (1993) On the ontological expressiveness of information systems analysis and design grammars. J Inf Syst 3:217–237
35. Wand Y, Weber R (1995) On the deep structure of information systems. Inf Syst J 5:203–223
36. Wand Y, Weber R (2002) Information systems and conceptual modeling a research agenda. Inf Syst Res 13(4):203–223
37. Wang S (1996) Two MIS analysis methods: an experimental comparison. J Educ Bus Jan/Feb:136–141
38. Yadav S, Bravoco R, Chatfield A, Rajkumar T (1988) Comparison of analysis techniques for information requirements determination. Commun ACM 31(9):1090–1097