GUIDELINES FOR EMPIRICAL EVALUATIONS OF CONCEPTUAL MODELING GRAMMARS

Andrew Burton-Jones
Sauder School of Business
The University of British Columbia
Vancouver BC
Canada V6T 1Z2
e-mail: Andrew.Burton-Jones@sauder.ubc.ca

Yair Wand
Sauder School of Business
The University of British Columbia
Vancouver BC
Canada V6T 1Z2
e-mail: Yair.Wand@ubc.ca

Ron Weber
Faculty of Information Technology
Monash University
PO Box 197
Caulfield East, Victoria
Australia 3145
e-mail: Ron.Weber@infotech.monash.edu.au

Working paper version of accepted paper.
Original Submission July 2007
Revision and Resubmission August 2008
Final Revision and Resubmission, April 2009

GUIDELINES FOR EMPIRICAL EVALUATIONS OF CONCEPTUAL MODELING GRAMMARS

ABSTRACT

Conceptual modeling grammars are used to create scripts that represent someone's perception, or some group's negotiated perception, of domain semantics. For many years, researchers have evaluated conceptual modeling grammars to determine ways that they can be improved. One way to evaluate them is to empirically evaluate the strengths and weaknesses of the grammars in terms of their effectiveness and efficiency in generating scripts. A number of researchers have proposed guidelines for the design of empirical research to conduct such evaluations. Although these guidelines have proved useful, further clarification is needed in relation to (1) criteria for evaluating grammar performance, (2) characteristics of grammars that can influence grammar performance, and (3) factors that must be considered when testing the effect of grammar characteristics on grammar performance. We review past conceptual modeling research and provide guidelines for addressing these three issues. We also illustrate how the guidelines would apply to studies that evaluate conceptual modeling grammars from an ontological perspective. Finally, we discuss how the guidelines extend those offered in past research and the implications of our work for future research.

Keywords: Conceptual modeling grammars; Grammar quality; Grammar performance; Script creation; Script interpretation; Ontology; Experimental design

INTRODUCTION

Information systems provide representations of the semantics of a domain (Kent 2000). These representations are the result of a design process that often begins with conceptual models, also known as conceptual modeling scripts, which represent the semantics of the domain as perceived by stakeholders of the information system. As the name suggests, conceptual modeling focuses on the conceptual aspects of a domain. Unlike software and database modeling, conceptual modeling eschews design and implementation considerations. This is because it is critical to have a good understanding of the domain to be supported by the information system before launching into design and programming work (Yourdon 1989).

Because of the importance of conceptual modeling during the development of information systems, the evaluation of conceptual modeling-related phenomena is an active research area (Khatri et al., 2006; Corral et al., 2006; Maes and Poels, 2007). In particular, much work has focused on evaluating conceptual modeling grammars, such as the entity-relationship modeling grammar or the business process modeling notation (Siau and Rossi, in press). In this vein of work, researchers are interested in improving the extent to which grammars enable their users to produce high-quality conceptual modeling scripts.

Conceptual modeling grammars might be evaluated in a number of ways. Analytical evaluations, for example, might focus on measuring characteristics of a grammar such as the degree to which its constructs have mnemonic value or the degree to which it offers a complete set of constructs for modeling a domain. Empirical evaluations, on the other hand, might focus on associating characteristics of a grammar with empirical outcomes. Many outcomes might be examined such as the usefulness of the grammar and individuals' adoption of it in practice (Recker 2008). Because the purpose of conceptual modeling grammars is to create scripts, however, empirical evaluations have traditionally focused on the strengths and weaknesses of

alternative grammars in terms of their effectiveness and efficiency in generating scripts (e.g., Parsons and Cole 2005). We adopt this focus because it has been the dominant approach in prior literature.

Several studies have offered guidelines or frameworks to help researchers who wish to evaluate conceptual modeling grammars via scripts (Wand and Weber, 2002; Gemino and Wand, 2004; Siau, 2004; Parsons and Cole, 2005; Aguirre-Urreta and Marakas, 2008). Nonetheless, several matters still need to be clarified: (1) performance criteria that can be used to evaluate conceptual modeling grammars, (2) characteristics of grammars that influence their performance, and (3) factors that researchers should consider when testing the effect of grammar characteristics on grammar performance. The aim of our paper, therefore, is to provide guidelines to address these issues. By so doing, we hope to contribute in two ways. First, we wish to provide guidelines that are broad enough that they can be used by all researchers who wish to evaluate conceptual modeling grammars empirically, irrespective of the theory or research method they employ. Second, we wish to clarify some issues discussed in prior studies that are easily misunderstood. In particular, we seek to clarify the types of research questions for which guidelines offered in prior studies will or will not apply, thereby extending the contribution of these prior studies. More generally, we hope this paper will help clarify the ways in which conceptual modeling grammars can be, and have been, evaluated in the information systems literature, so as to highlight opportunities for future research.

The remainder of the paper is structured as follows. We begin by reviewing some basic concepts that underpin our analyses. We then outline our proposed guidelines. Next, we illustrate how these guidelines could work in practice by describing how they could be used by researchers who employ ontological theories to evaluate conceptual modeling grammars. We then discuss the implications of our guidelines and the extent to which they extend guidelines offered in past research. Finally, we present some brief conclusions.

BASIC CONCEPTS

A *conceptual modeling grammar* provides a set of constructs and a set of production rules that enable a user of the grammar to represent someone's perception, or some group's negotiated perception, of the semantics of a domain. For example, in the entity-relationship conceptual modeling grammar, the constructs are an entity, a relationship, and an attribute. A production rule in the grammar is that an entity can have an attribute.

A *conceptual modeling script* (sentence/string) is a representation of the semantics of a domain, often diagrammatic, generated using a conceptual modeling grammar. For example, using the entity-relationship conceptual modeling grammar, a script might be a "man" entity joined to a "woman" entity via a "married to" relationship.

A *conceptual modeling language* is the set of all scripts that can be generated via a conceptual modeling grammar. In other words, it comprises all scripts that can be produced using a conceptual modeling grammar to represent all domains in which the grammar might be applied.

In the field of linguistics, languages are often studied from the perspectives of syntax, semantics, and pragmatics (Parker and Riley, 2005). The study of *syntax* is concerned with how words are combined to form phrases and sentences. One focus is the nature of the grammar's rules, which prescribe the valid ways that phrases and sentences can be constructed. Another is how users of the grammar form phrases and sentences in practice. Accordingly, with a conceptual modeling grammar, the study of syntax might involve examining valid ways in which scripts can be created using a grammar or examining alternative ways that individuals form scripts using the grammar (e.g., by examining the effects of arranging grammatical constructs on a diagram in different ways or the effects of using "nouns" to label "entities" and "verbs" to label "relationships" when creating an entity-relationship diagram).

The study of *semantics* focuses on the meaning of words, phrases, or sentences in a language. Because humans construct meaning from language in complex ways, the study of semantics has been a focus in many disciplines (e.g., linguistics, psychology, sociology, philosophy, and computer science). With a conceptual modeling grammar, the study of semantics might involve examining the meaning of the constructs in the grammar, the meaning of production rules in the grammar, and the meaning of scripts generated via the grammar.

The study of *pragmatics* focuses on how languages are used in practice. The meaning that exists, *prima facie*, in the words, phrases, and sentences of a language might differ from the meaning that individual users of the language ascribe to them. In particular, pragmatics might reflect the context in which language phrases are formed and used. With a conceptual modeling grammar, the study of pragmatics might focus on the meaning that different users assign to the constructs and production rules in the grammar and the scripts generated via the grammar. An example would be how users ascribe meaning to entity types when they are used to represent both things and events in a particular domain.

The study of pragmatics in language is motivated in part by the distinction between the *denotational* meaning and the *connotational* meaning of a word, or phrase, or sentence. The study of *denotational semantics* focuses on the *prima facie* (sometimes called “objective”) relationship between words, phrases, sentences, and their referents. The study of pragmatics has shown, however, that humans do not always interpret words, phrases, or sentences in the same way. They consider their meaning in the context of the meaning of other words, phrases, and sentences. Moreover, they interpret words, phrases, and sentences based on their prior knowledge and the circumstances in which they undertake the interpretation task or the purpose for which they construct phrases. The study of *connotational semantics*, therefore, focuses on how humans create meaning and interpret it in practice. It recognizes that the ways individuals interpret the meaning of words, phrases, or sentences often differ from their *prima facie*

meaning. Likewise, it recognizes that individuals often account for the context in which their words, phrases, and sentences will be interpreted when they determine how to communicate.

GUIDELINES

We propose three guidelines for research that empirical evaluates conceptual modeling grammars in terms of their effectiveness and efficiency in generating scripts (Figure 1). The following subsections discuss each guideline in turn.

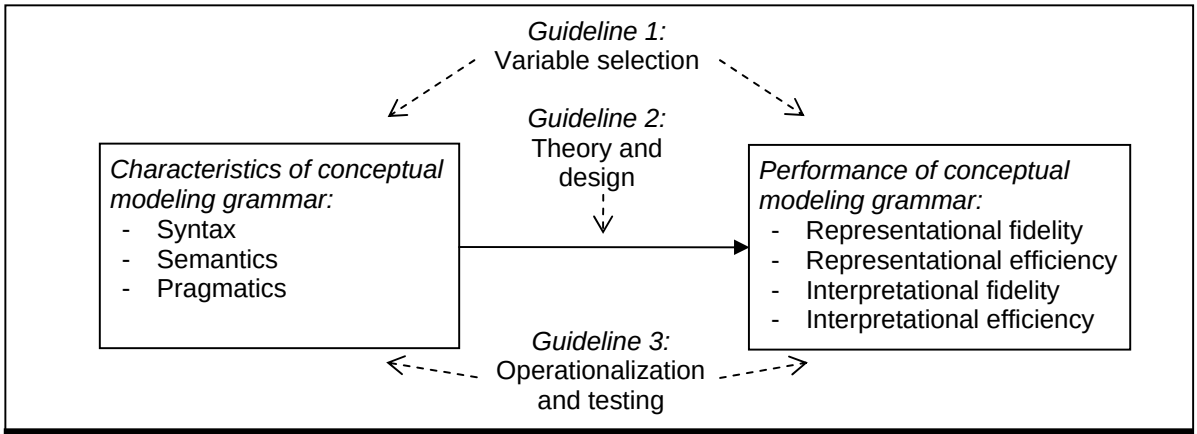


Figure 1: Guidelines for Empirical Evaluations of Conceptual Modeling Grammars

Guideline 1: Variable selection

When evaluating a conceptual modeling grammar, we believe researchers would benefit from understanding the range of potential predictor variables and outcome variables that they might use. We will first suggest a set of outcome variables and then a set of predictor variables that researchers might employ.

Outcome variables. Like any tool, we cannot evaluate the 'truth' of a conceptual modeling grammar, only its performance (i.e., its effectiveness and efficiency) (Moody 2003, p. 210). To evaluate a grammar's performance, we must know how it is used. Past research highlights two important ways in which grammars are used: (1) to create scripts (when individuals use their

knowledge of the grammar to create a script); and (2) to interpret scripts (when individuals use their knowledge of the grammar to interpret a script) (Gemino and Wand, 2004). Thus, we propose that an important way in which the performance of a conceptual modeling grammar can be evaluated empirically is to assess its effectiveness and efficiency in supporting script creation and script interpretation.

Because conceptual models are created to reflect domain semantics, we assess their *effectiveness* in terms of *fidelity* (Parsons and Cole, 2005). We assess their *efficiency* in terms of the amount of resources needed to prepare or interpret them. Accordingly, as Figure 1 shows, we propose four outcome variables that researchers might use to evaluate the scripts produced using a conceptual modeling grammar:

- *Representational fidelity*: how faithfully does the script represent someone's perception, or some group's negotiated perception, of the semantics of the domain?
- *Representational efficiency*: what resources are used to create the script?
- *Interpretational fidelity*: how faithfully does the interpretation of the script represent the semantics in the script?
- *Interpretational efficiency*: what resources are used to interpret the script?

These outcome variables are similar to those identified in some prior studies (Wand and Weber, 2002; Gemino and Wand, 2004; Aguirre-Urreta and Marakas, 2008). For instance, as in these studies, our outcome variables for fidelity can be viewed in terms of Norman's (1986) theory of action. As Figure 2 shows, limitations in representation fidelity create what Norman calls the gulf of execution (a difference between the semantics understood by the stakeholders and the semantics represented in the script). Limitations in interpretation fidelity create what Norman calls the gulf of interpretation (a difference between the semantics reflected in the script and the

semantics interpreted by the reader).

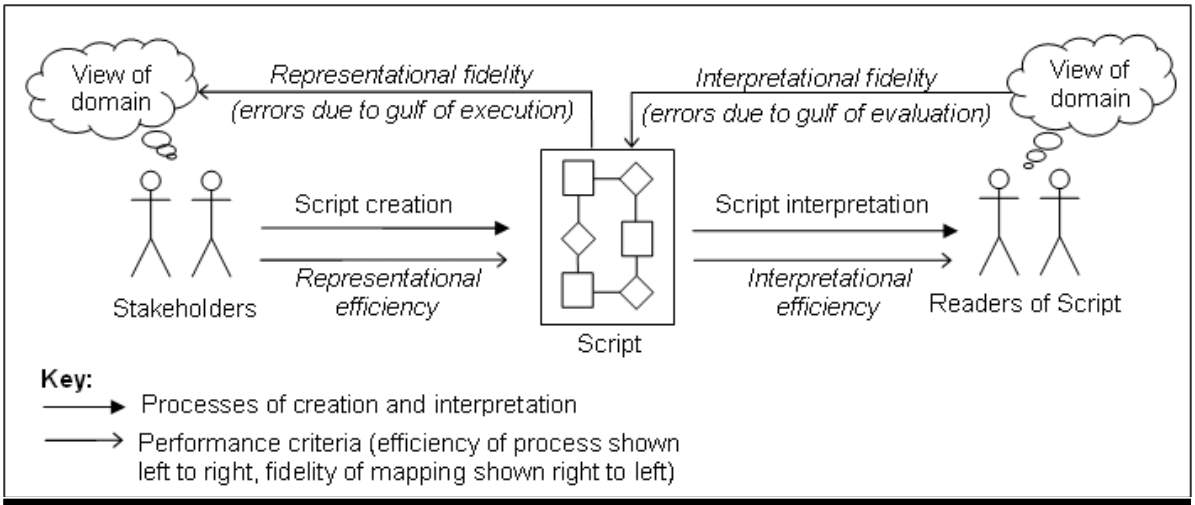


Figure 2: Grammar Evaluation Criteria

Another way in which our outcome variables are similar to those in prior studies is that they apply broadly. They are not tailored to a specific context. That is, our four outcome variables should be relevant to any study in which researchers examine one or more grammars and, based on this evaluation, make predictions about script creation or script interpretation phenomena. For example, an individual or a group in collaboration with end-users or in isolation from them may create a script. In all these contexts, researchers can evaluate the extent to which a grammar enables the creator(s) of the script to construct an effective (high-fidelity) script in an efficient manner.

Likewise, a script might be interpreted soon after or long after its creation by the individual(s) who created it or by other stakeholders. Moreover, it might be interpreted to support many tasks (e.g., systems analysis, communication, design, project management, end-user querying, organizational change management) (Kung and Solvberg, 1986, Hirschheim et al., 1995). Rather than examine context-specific measures of effectiveness, we focus on the more-immediate issue of interpretability because it is relevant in all contexts of use (Aguirre-Urreta

and Marakas, 2008, p. 12). That is, many people may interpret a script, at many different times, and for many different tasks. Nonetheless, in all contexts, it is useful to know if a grammar enables a reader to obtain an effective (high-fidelity) interpretation in an efficient manner.

Predictor variables. Many factors could affect the outcome variables mentioned above. Our focus, however, is the extent to which the characteristics of a grammar affect them. Similar to Lindland et al. (1994), we suggest that three characteristics of grammars are especially relevant:

- *Syntax:* the constructs in the grammar and their rules for arrangement.
- *Semantics*: the meaning of the constructs in the grammar.
- *Pragmatics*: the context in which a grammar is used.

Two points should be noted about these characteristics. First, when we assess the performance (efficiency or fidelity) of script creation, the relevant predictor variables are the syntax and semantics of the grammar and the pragmatic context in which the script is created (such as the skills of the modeler who created the script). When we assess the performance (efficiency or fidelity) of the interpretation process, however, the relevant predictor variables are the syntax and semantics of the grammar instantiated in the script and the pragmatic context in which the script is interpreted (such as the skills of the reader who interpreted the script).

Second, as noted earlier, two types of semantics exist: denotational and connotational. The distinction is important because a potential criticism of the outcome variables we have proposed is that representational fidelity cannot be assessed without making an interpretation and, therefore, the distinction between representational fidelity and interpretational fidelity is moot. We accept this criticism, but we believe the distinction between representational fidelity and interpretational fidelity is still useful analytically. Representational fidelity is a function of the denotational semantics manifested in the script, whereas interpretational fidelity is a function of

both the denotational semantics of the script and the connotational semantics that arise when someone interprets the script. These two outcomes variables, therefore, are not the same.

Summary. Based on the aforementioned outcome variables and predictor variables, Table 1 shows the range of studies that researchers can perform to evaluate a conceptual modeling grammar empirically via scripts. Table 1 also lists examples of some of these types of studies. To populate Table 1, we reviewed all articles published from 1998-2008 in the six journals listed by the Association for Information Systems as "top journals" in the IS field (*European Journal of Information Systems, Information Systems Journal, Information Systems Research, Journal of the Association for Information Systems, Journal of Management Information Systems, and MIS Quarterly*). 1602 articles were published in this sample of journals in this timeframe. Of these articles, we identified 13 candidate articles that focused on modeling in analysis or design. Of these 13 articles, we classified seven as having empirically evaluated a conceptual modeling grammar. Although some other journals publish more conceptual modeling research, Table 1 provides a useful snapshot of the research that has been published in this area recently. In Appendix 1, we describe how we determined which articles were included in the seven relevant to our purpose and how we classified these articles according to the cells of Table 1. We also describe heuristics that we found useful for classifying conceptual modeling work.

Overall, Table 1 shows 28 types of studies. All reflect feasible research studies. Nonetheless, the citations in Table 1 show that only a limited number of the different types of studies have been undertaken. To illustrate the feasibility of each type of study, we provide a description in Appendix 2 of studies that could be undertaken to examine all main effects and two-way interactions in the table.

Table 1: Possible Research Studies for Evaluating Conceptual Modeling Grammars Via Scripts and Examples

		Process and performance criteria *(Outcome variables)*			
		Script Creation		**Script Interpretation**	
		Representational fidelity as outcome	**Representational efficiency as outcome**	**Interpretational fidelity as outcome**	**Interpretational efficiency as outcome**
Effect of grammar characteristics *(Predictor variables)*	**Main effect of syntax**	1. Kim et al. 2000	2.	3. Kim et al. 2000	4. Kim et al. 2000
	Main effect of semantics	5. Kim et al. 2000 Bodart et al. 2001 Hadar and Soffer 2006 Soffer and Hadar 2007 Parsons & Wand 2008 Shanks et al. 2008	6.	7. Kim et al. 2000 Bodart et al. 2001 Shanks et al. 2008	8. Kim et al. 2000 Bodart et al. 2001[1] Shanks et al. 2008[1]
	Main effect of pragmatics	9. Soffer & Hadar 2007[1]	10.	11. Bodart et al. 2001[1] Khatri et al. 2006	12.
	Interaction effect of syntax and semantics	13.	14.	15.	16.
	Interaction effect of syntax and pragmatics	17.	18.	19.	20.
	Interaction effect of semantics and pragmatics	21.	22.	23.	24.
	Interaction effect of syntax, semantics, and pragmatics	25.	26.	27.	28.

* Citations in more than one cell reflect that more than one issue was examined in the same study.

[1] Indicates that this cell was a minor focus of the paper.

Guideline 2: Theory and design

Although all cells in Table 1 reflect feasible research topics, we are not suggesting that researchers must study every single cell. Rather, in any given study, it is important that researchers justify why the variables in the cell (or cells) examined in that study are interesting and important and, to the extent possible, present a theory to explain the relationships among

the variables. We give an example of how a researcher could do so later in the paper, when we discuss how researchers could use the theory of ontological expressiveness (Wand and Weber 1993) to evaluate a conceptual modeling grammar.

When researchers conduct a study in any of the cells in Table 1, they also need to design the study, or analyze the study's data, in such a way that they can (a) identify an effect of the predictor variable (if one exists), and (b) control for the effects of other variables that are not the study's focus. This requirement is necessary to ensure the study faithfully tests the theory and, as a result, has high internal validity.

As an example of the first practice, consider studies that examine the impact of the semantics in a grammar on the interpretational fidelity of scripts produced using that grammar. In such a study, the researcher must identify how a variation in the semantics of the grammar affects readers of scripts produced using the grammar. A common way to design such a study is to give alternative scripts with different semantics to a random sample of readers and ask the readers to answer questions based on the script (e.g., Shanks et al. 2008). As Parsons and Cole (2005) note, a problem that can occur in such studies is that the readers might not answer the questions based only on the scripts they received. Rather, they may use their background knowledge of the domain shown in the script to answer the questions. If this outcome occurred, researchers might find no significant difference between groups in the answers the groups provide. Importantly, the outcome would not reflect that the semantics in the script were unimportant. Rather, it would reflect that experimental participants did not refer to the semantics in the script (i.e., the task was not salient to experimental participants).

Researchers can address the issue of salience in three ways. First, prior to the conduct of their research, they can ask individuals who are representative of their participant cohort to assess the extent to which they believe the scripts are salient to the tasks that have to be performed.

Low-salience tasks then should not be used in the research. Second, after participants have completed their tasks, they can be asked to provide feedback on the salience of each task in light of the scripts they received. Low-salience tasks can be excluded from data analysis. Third, to the extent tasks fail to manifest differences between different treatment groups, their salience must be questioned. Alternatively, other explanations must be found for the absence of differences between treatment groups–for instance, a poor theory or poor research method.

Researchers must also ensure that they control for the effects of other variables that are not the study's focus. For example, in studies that focus on the creation or interpretation of scripts, Aguirre-Urreta and Marakas (2008) highlight the importance of controlling for "pragmatic" factors such as the level of mastery that an individual (modeler or reader) has in the modeling grammar and other individual difference factors (e.g., cognitive abilities). Thus, if researchers wish to examine cells in Table 1 that are associated with syntax and/or semantics (but not pragmatics), they must control for "pragmatic" factors (whether in the design of the study, in the analysis of data, or both) to ensure that the study's results are not confounded. Such pragmatic factors also affect the external validity of a study because the only way to evaluate the performance of a grammar is to evaluate its ability to support script creation and script interpretation processes. These processes, in turn, always occur in some pragmatic context. As a result, the specific *properties* of this pragmatic context (such as the level of experience of the modeler or reader, the time allowed for tasks, the incentives to perform, and so on) will affect the extent to which the results of the study can be generalized to other settings. Empirical researchers must remain mindful of the pragmatic contexts in which their studies are undertaken and understand how these contexts affect the generalizability of their findings (Lee and Baskerville 2003).

Guideline 3: Operationalization and testing

Once researchers have selected variables and theorized relationships among them, they need

to operationalize their constructs and test the relationships posited among them. Typically, empirical evaluations of grammars are relative rather than absolute. That is, researchers wish to say that two or more grammars with different syntactic, semantic, or pragmatic characteristics perform differently (some better, some worse) rather than to quantify a grammar's absolute performance. To measure relative differences, researchers must have a way to detect whether their predictor variables, control variables, and outcome variables display variance.

Researchers can test for the *presence* of variance in several ways. Perhaps the most-common approach is to use analysis-of-variance procedures to test whether a significant difference exists between the mean responses given by two experimental groups (e.g., in their responses to a manipulation check for a predictor variable or in their responses expressed via some measure of an outcome variable, such as time or accuracy).

Researchers can also use various techniques to test for the *absence* of variance. For example, they might obtain ratings of two grammars from expert users of the grammar and use inter-rater agreement statistics to test whether the ratings are similar or even equivalent. Alternatively, researchers might obtain ratings from a sample of end-users and use analysis-of-variance procedures, together with a power calculation, to test for the absence of a significant difference between the mean ratings of the two groups of end-users.

Although researchers have many ways to test for the presence or absence of variance, challenges arise with both types of test. The challenge with testing for the *presence* of variance is that some researchers may criticize the study by saying that the results are obvious. The challenge with testing for the *absence* of variance is that it is difficult, if not impossible, to conclusively verify the absence of variance. Because we believe these challenges are not well understood, we highlight them in the sections below and suggest how they might be addressed.

Challenges when testing for the presence of variance. Whenever researchers investigate whether differences in one variable lead to differences in another, they might be charged with testing the obvious. To illustrate, consider studies in cell 5, Table 1. In such studies, researchers wish to examine whether differences in the semantics of two or more grammars result in scripts that differ in the accuracy with which they represent a domain. Figure 3 provides scripts that might be used in such a study. Figure 3a was produced using a grammar that allows mandatory properties only, while Figure 3b was produced using a grammar that allows mandatory and optional properties (and where the modeler chose to use both types of constructs). Assume in this case that a researcher wished to compare these grammars by randomly assigning these scripts to experimental participants and asking the participants: "Will PhD students who have an advisor have to pay tuition?" Assume also that the correct answer in the domain is "no." Presumably, participants receiving Figure 3a will answer "no," while participants receiving Figure 3b will not. Is such a test worth conducting? Parsons and Cole (2005, p. 330) write: "…if one form provides enough information to answer selected questions correctly, while a second form does not, it would not be surprising to find that participants receiving the first form outperform those receiving the second form on those questions."

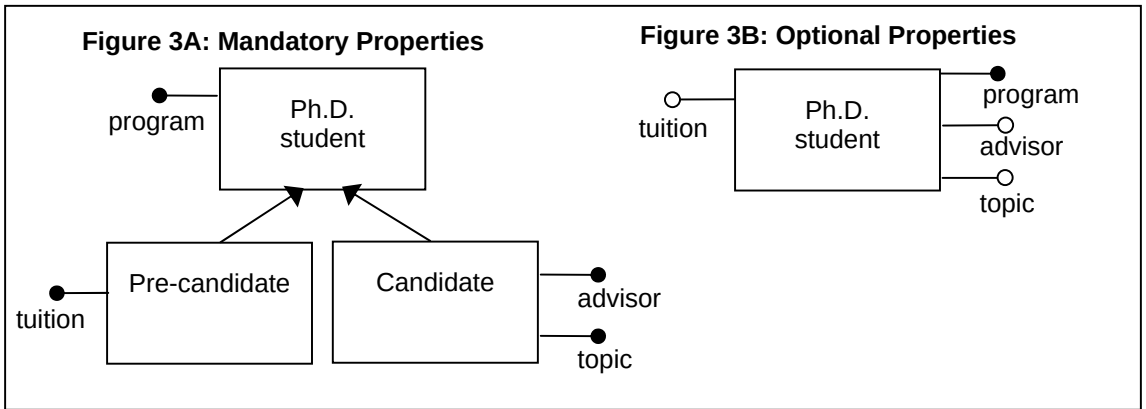


Figure 3: Two Scripts of a Domain (Adapted from Gemino & Wand 2005, p. 303)

Note: → is a subclass of —○ optional property —● mandatory property

Parsons and Cole (2005) appear to be concerned about testing for outcomes that are obvious. It is important to note that this type of criticism can be levied at any study in which a researcher wishes to test whether differences in a predictor variable lead to differences in an outcome variable. We accept that occasions will arise where differences in the predictors will appear to be so substantial that testing for differences in outcomes seems pointless. Nonetheless, when these occasions occur, great care must be taken when decisions are made about whether to proceed with the tests. The history of science is littered with examples of obvious outcomes that have been contradicted by empirical evidence. Moreover, the information systems field has its own examples (e.g., Allen and Parsons, in press). In this regard, we propose that three matters ought to be considered when determining whether an empirical test has merits.

First, if examples of the treatment that are expected to lead to the poorest outcome can be found in extant literature or practice, empirical tests should be undertaken. If such examples exist, clearly the reasons why the different versions of the treatment differ are not obvious to all who have a stake in the conceptual modeling field. One way that researchers who undertake such evaluations might motivate their work, therefore, is to provide examples of the treatments from published literature or practice. For instance, in the context of Figure 3, a researcher could cite textbooks that recommend that modelers create scripts with optional properties.

Second, because many examples of counterintuitive outcomes exist in the history of science, some level of empirical confirmation of "obvious" outcomes is still needed. If the first empirical test confirms the prediction, only a small number of replications might then be needed. From a research viewpoint, however, not to undertake at least one test and not to undertake at least some replications of the test is foolhardy behavior. For example, consider again the two scripts in Figure 3 and the question "Will PhD students who have an advisor have to pay tuition?" It is possible that a researcher could use these materials in an experiment and obtain no significant difference between groups on their answers to the question, a seemingly counterintuitive

outcome. For example, if the participants in the experiment were university students, all of the students might answer “no” to the question, irrespective of the script they are given, purely based on their knowledge of the policies at many universities (i.e., that advisors fund students).

Third, empirical tests of theoretical predictions provide a means of *calibrating* the consequences of a treatment. For example, even if it is obvious that users of Figure 3a will conclude that PhD students who have an advisor do not have to pay tuition, and that users of Figure 3b will not be able to reach this conclusion, empirical evaluations are still useful methodologically and theoretically. Methodologically, such evaluations can be used for instrument validation. For example, if experimental participants failed to provide the expected pattern of answers, it might indicate that the instruments used to measure the outcomes were not valid (e.g., perhaps participants misunderstood the question or misunderstood the response options available). Theoretically, such evaluations can be used to test the sensitivity of participants to the treatments. For example, even if the overall pattern of results to our question regarding Figures 3a and 3b is obvious, will *all* participants answer in the expected manner? If the difference in outcomes is minor even when the treatment is strong, the experiment is internally valid, and the tests are reliable and valid, then the practical usefulness of the theoretical predictions should be questioned. If the difference in outcomes is substantial, however, greater importance can be ascribed to the theoretical predictions.

In short, whenever researchers examine whether variance in a predictor creates variance in an outcome, they could be criticized for testing the obvious. In all such studies, therefore, we suggest that researchers explain why the difference in the outcomes they are testing are relevant in practice, why they are not obvious, and even if they are somewhat obvious why conducting the test is still important (for theoretical or methodological reasons).

Challenges when testing for the absence of variance. Recently, some researchers have stressed the importance of verifying an *absence* of variance. Specifically, when examining the ability of individuals to interpret scripts, they have sought to explain whether the scripts they are comparing are "informational equivalent" and/or "computational equivalent" (e.g., Agarwal et al., 1999; Siau, 2004; Gemino and Wand, 2004; Parsons and Cole, 2005; Corral et al., 2006; Maes and Poels, 2007, Aguirre-Urreta and Marakas, 2008). Two scripts are *informationally equivalent* when "all information in one is also inferable from the other and vice versa" (Larkin and Simon, 1987, p. 67). Two scripts 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" (ibid). In the absence of information equivalence, a concern has been that "internal validity is threatened, since differences in information content may confound attempts to measure differences in comprehension of alternate semantically equivalent representations" (Parsons and Cole, 2005, p. 330).

As noted above, researchers can use various techniques to assess the absence of variance in measures. At first glance, therefore, verifying the informational equivalence or computational equivalence of scripts may not seem difficult. In our view, however, *conclusively* verifying the equivalence of scripts is not only difficult but impossible.

Informational equivalence cannot be verified conclusively for three related reasons. First, informational equivalence is subjective because users' interpretations of a script are affected by connotational semantics, not just denotational semantics. Because different people have different knowledge, we cannot assume that all people will infer the same connotational semantics from a given representation (Patel et al., 2004). For example, consider once again the two scripts in Figure 3 and the question: Do all Ph.D. students have advisors? A researcher might claim that these two scripts are informationally equivalent with respect to this

question because both scripts indicate that the correct answer is “no.” Nevertheless, not all readers of these two scripts might give this answer. If one reader receives Figure 3A and knows what subclass relationships imply, and another reader receives Figure 3B but does not know what optional properties imply, the two readers will not obtain the same information from the scripts (see Siau, 2004, and Aguirre-Urreta and Marakas, 2008, for a similar argument).

This problem might be alleviated if researchers could identify and control for the background knowledge of each user of a script. Unfortunately, it is not clear whether informational equivalence is defined in terms of *all* users of alternative scripts or a *single* user of the scripts. If it is defined in terms of *all* users, researchers would have to identify the population of possible users, obtain a random sample from this population, and control for the background knowledge of each user, if they wished to verify the equivalence of the scripts. Such sampling strategies are exceedingly difficult, if not impossible, to implement. If researchers did not follow such a strategy, however, they could not verify the equivalence of a script for *all* users because different members of the population may have different knowledge and engage different connotational semantics when they interpret scripts.

Finally, even if informational equivalence pertains to just *one* user, we still do not see how informational equivalence can be established unequivocally through empirical methods. For instance, consider the case of an individual presented with two alternative scripts of a domain. Once the individual has examined one representation, her/his conclusions about the second representation have been confounded. Cognitive processing associated with the first representation could either enhance or undermine cognitive processing associated with the second representation. To establish informational equivalence unequivocally, the individual must be able to examine the second representation from the viewpoint of *tableau rasa*–a requirement that is impossible to fulfill.

The concept of computational equivalence suffers from the same problems we have attributed to the concept of informational equivalence: (a) it is not clear whether computational equivalence is defined in terms all users or a single user of two scripts that provide alternative representations of a domain; (b) different users of scripts may have different knowledge, and this knowledge may influence the cognitive processing required to interpret a script (and thus whether two scripts are computationally equivalent); and (c) because of cognitive confoundings, we cannot see how computational equivalence can ever be shown for a single user.

Moreover, an evaluation of two scripts for computational equivalence can proceed only under the assumption that the user under scrutiny agrees with someone else's assessment that the scripts are informationally equivalent. If on the basis of the scripts the user makes correct inferences, *prima facie* support exists for the validity of this assumption. If the user makes incorrect inferences, however, it is not clear whether (a) she/he considers that the scripts are *not* informationally equivalent, or (b) because of high computational overheads associated with one or both scripts, she/he terminates the task (e.g., through frustration or exhaustion) before the correct inferences can be drawn. On the other hand, if the user is first asked to assess whether the scripts are informationally equivalent and she/he concludes they are, subsequent assessments to determine computational equivalence are then confounded by the cognitive "computation" that has occurred already to determine whether informational equivalence exists.

Overall, because informational equivalence and computational equivalence cannot be verified conclusively, we believe that researchers should be cautious about using these concepts. If researchers wish to use them, they should take two steps. First, they should explain the steps they took to *maximize* the equivalence of the relevant treatments or controls in their study. Kim et al. (2000) give an example. They proposed that two sets of scripts in their study were informationally equivalent. To maximize the degree of equivalence, they transformed their scripts to natural language statements, compared the natural language statements for

equivalence, revised the scripts to improve equivalence, and repeated this process several times to maximize equivalence among the scripts. Second, researchers should obtain *evidence* to indicate whether the operationalizations that they claimed to be equivalent were, in fact, sufficiently similar to be deemed "practically" equivalent. Gemino and Wand (2005) give an example. They proposed that two alternative scripts prepared to represent a domain in their study were informationally equivalent. To check this assumption, they created a set of comprehension questions regarding the semantics in the scripts. After finding that experimental participants receiving one version of the scripts did not perform significantly differently on the comprehension test from participants receiving the alternative script, they concluded that the scripts were practically equivalent. Researchers can also create tests to measure "practical" computational equivalence–e.g., via the time taken to create or interpret a script (Siau, 2004).

ILLUSTRATION: THE THEORY OF ONTOLOGICAL EXPRESSIVENESS

We illustrate our guidelines by explaining how they could inform researchers who use the theory of ontological expressiveness. The theory of ontological expressiveness enables researchers to evaluate the ability of a conceptual modeling grammar to reflect domain semantics (Wand and Weber, 1993). The semantics are defined by a mapping between grammatical constructs and ontological constructs. We provide a summary of the theory in Appendix 3. Other theories can also be used to evaluate conceptual modeling grammars, independently or in conjunction with the theory of ontological expressiveness–such as theories of cognitive fit (Vessey, 1991; Khatri et al., 2006), diagrammatic reasoning (Kim et al., 2000), semiotics (Krogstie et al., 2006, Siau and Tian, 2009), and linguistics (Becker et al., 2008). We focus on the theory of ontological expressiveness alone for two reasons. First, it has been used extensively to evaluate conceptual modeling grammars and scripts. Second, much discussion about the need for informational and computational equivalence when evaluating conceptual modeling grammars

and scripts has been motivated by research conducted using ontological theories (Parsons and Cole, 2005).

Guideline 1: Variable selection

The theory of ontological expressiveness describes four defects in a grammar—*redundancy*, *overload*, *excess*, and *deficit*—that could affect the performance of the grammar. All four outcome variables proposed earlier could be used to test the theory. For example:

- *Representational fidelity:* If a grammar contains construct deficit, a researcher might predict that scripts created using the grammar will contain instances of these defects. Such a script will lack representational fidelity because it will fail to contain relevant semantics according to the ontological benchmark.

- *Representational efficiency:* If a grammar contains any of the four defects, a researcher might predict that a modeler using the grammar will take more time trying to decide how to use the constructs in the grammar to model the domain as faithfully as possible.

- *Interpretational fidelity:* If a script contains instances of construct redundancy, construct overload, or construct excess according to the ontological benchmark, a researcher might predict that readers will be confused by the presence of different syntax to represent the same phenomenon (redundancy), the use of one type of syntax to represent different phenomenon (overload), and the presence of seemingly irrelevant information (excess). As a result of this confusion, readers could give an interpretation of the script that ascribes semantics to the domain that are different from the semantics represented in the script.

- *Interpretational efficiency:* If a script contains instances of construct redundancy, construct overload, or construct excess, a researcher might predict that readers will be

confused by these defects in the script. As a result of this confusion, readers will take longer to obtain a faithful interpretation of the semantics in the script.

Likewise, a researcher testing the theory of ontological expressiveness could consider all three predictor variables noted above. For example:

- *Syntax:* Construct redundancy is a syntactic problem because it occurs when a grammar offers multiple types of syntax (symbols) to represent one type of semantics.

- *Semantics:* Construct overload, excess, and deficit are semantic problems because they occur when syntactic elements (symbols) in a grammar fail to distinguish between different types of semantics (overload), when a grammar contains semantics that are meaningless in a domain (excess), or when the grammar does not enable a modeler to show relevant semantics (deficit) according to the ontological benchmark.

- *Pragmatics:* Construct redundancy, overload, excess, and deficit may cause more problems in some contexts than in other contexts. Specifically, the extent of the problems that arise might depend on the expertise of the user (modeler or reader) of the script and the task for which the grammar or script is being used. For example, if readers know the domain being modeled, they might supplement information missing from the model based on their own knowledge of the domain.

Guideline 2: Theory and design

Using the predictor and outcome variables noted above, a researcher could use the theory of ontological expressiveness to evaluate a conceptual modeling grammar in many ways. In Appendix 2, we briefly describe 18 such studies. In Appendix 3, we also provide examples of scripts that could be used in some of these studies. Rather than discuss all such studies here,

we highlight (a) how researchers can use the theory of ontological expressiveness to specify relationships between predictor and outcome variables, and (2) some design issues that researchers should consider when testing such relationships.

Given the predictors and outcomes noted above, researchers could use the theory of ontological expressiveness to propose three types of relationships. First, they might propose a main effect of a predictor on an outcome. For example, they may predict that modelers using a grammar that contains construct redundancy will produce scripts that contain instances of construct redundancy. They might then propose that readers of such scripts will be confused by the use of different grammatical constructs to show one type of ontological construct. As a result, readers of the script might assume wrongly that the different grammatical constructs reflect different types of phenomena, thereby leading to a reduction in interpretational fidelity. We provide an example of this type of prediction in Appendix 3 (Figure A4).

Second, researchers might propose that an outcome depends on an interaction between two predictors. We give examples of such propositions in Appendices 1 and 2. We give another example in Figure 4. As Figure 4a shows, if a modeler uses a grammar that contains construct redundancy (a syntactic factor), a researcher may predict that the outcome will depend on the modeler's expertise (a pragmatic factor). That is, the researcher may predict that novice modelers will produce scripts that contain redundancy but that expert modelers will avoid using the redundant constructs. As a result, if a sample of individuals was randomly assigned scripts created with a grammar that exhibits construct redundancy, with half of the sample assigned scripts created by novice modelers and the other half of the sample assigned scripts created by expert modelers, the researcher might predict that the negative effect of construct redundancy on interpretational fidelity would occur only for the group that received scripts created by the novice modelers.

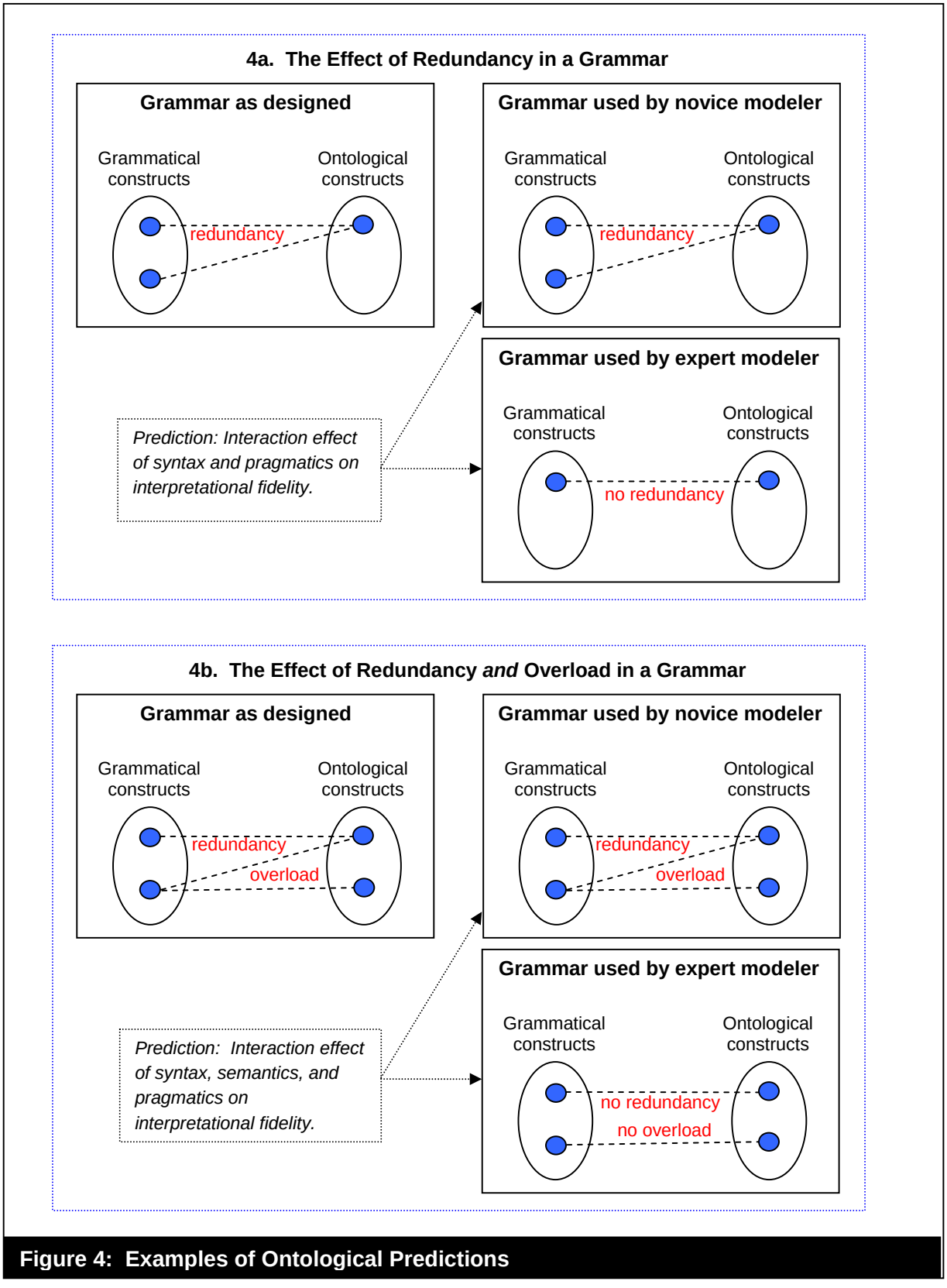


Figure 4: Examples of Ontological Predictions

Third, researchers might propose that an outcome depends on an interaction among all three predictors. Because of the complexity of three-way interactions, we do not provide examples of them in our Appendices. Nonetheless, we give a brief illustration in Figure 4b, in which the level of a pragmatic factor (the modeler's experience) determines whether the presence of both syntactic and semantic defects leads to (a) *lower* performance than a situation in which only one defect is present, or (b) *no change* in performance compared with a situation in which only one defect is present.

Specifically, when *novice modelers* use a grammar that contains both construct redundancy and construct overload, Figure 4b suggests they will produce a script that contains both types of defects. If users read the script containing both types of defects, a researcher might predict that users' interpretations will have lower fidelity than their interpretations of the script that has just one type of defect. When both defects are present, the researcher may predict that readers will assume wrongly that different constructs have different meanings (due to redundancy). As a result, readers may make mistakes about which phenomena a given grammatical construct represents (due to overload). Readers may also spend more time interpreting the model.

In contrast, when *expert modelers* use a grammar that contains both construct redundancy and construct overload, Figure 4b suggests they can take advantage of the redundancy to overcome problems caused by the overload. They can achieve this outcome by ensuring they use a different grammatical construct for each ontological construct. In short, when expert modelers use a grammar that has both syntactic and semantic defects, defective scripts need not result.

Whether researchers propose a main effect or an interaction effect, they must design their study to control for possible confounds. For example, if researchers wish to test the impact of construct redundancy on representational efficiency, they should attempt to hold other syntactic

and semantic factors constant (e.g., by ensuring that the syntax and semantics of the grammars are identical except for the presence of construct redundancy in one grammar). They should also seek to control the effects of pragmatic factors (e.g., in an experimental context, by ensuring that participants are a relatively homogenous set of modelers, randomly assigning a grammar to a modeler, obtaining reliable measures of each modeler's experience with the grammar and the domain to be modeled, and including these measures as covariates in the data analysis to control for their effects).

Guideline 3: Operationalization and testing

As noted earlier, when researchers test for the presence of variance, they might be accused of testing the obvious. Moreover, when researchers test for the absence of variance, they may be unable to do so. We discuss each challenge in turn.

Testing for the presence of variance. Any test of the theory of ontological expressiveness will require researchers to test for the presence of variance. Specifically, to use the theory to evaluate a conceptual modeling grammar, researchers will need to identify (in the case of correlational research) or create (in the case of experimental research) a situation in which (a) multiple grammars exist that vary in their number of defects, or (b) multiple scripts exist that were created using grammars that vary in their number of defects. Tests of the theory will then involve researchers examining whether this variance in the number of defects in the grammars (or in the scripts) is associated with variance in one of the four outcome variables.

In these types of tests, researchers may be accused of "testing the obvious." For example, assume researchers wish to test the impact of construct deficit on the effort that modelers exert to produce high quality use-case diagrams in the UML grammar (i.e., the type of study in Table 1, cell 6). As we outline in Appendix 3 (Figure A3), UML's use-case grammar is ontologically deficient because it lacks constructs to show how work systems are decomposed. One way to

test whether this type of construct deficit affects representational efficiency would be for researchers to create two versions of the use-case grammar (one with deficit and one without), randomly assign the grammars to a set of modelers (one grammar per modeler), and then ask the modelers to model a domain in which the decomposition of work systems is relevant. Researchers could then compare the effort that it takes modelers to produce scripts that faithfully model the domain. Presumably, it will take modelers less effort to model the domain faithfully if they have the grammar without construct deficit (i.e., the grammar that has constructs for modeling the decomposition of work systems). Some researchers might claim that this result would be "obvious" and thus of little value.

To address this criticism, we believe that researchers who conduct such a study should take the following steps. First, they should demonstrate that the problem they are studying occurs in the practice. For example, researchers might use quotes from practicing modelers who have written about the use-case grammar and who have mentioned that the deficiency in the grammar is problematical. Second, researchers should explain that the test they are undertaking will not produce obvious results or, if they agree the results are obvious, the test is still useful. For example, researchers might explain that the results are not obvious because modelers who receive the more-complete grammar may fail to use the additional construct in the grammar. Alternatively, they may make mistakes when using it because, for example, they are cognitively burdened by the number of constructs to consider in the grammar. As a result, modelers who use the grammar without construct deficit may exert the same level of effort or even exert more effort than modelers who use the deficient grammar. Even if the results emerge as expected, however, researchers might still argue that the test is valuable because (a) it is the first time the prediction has been tested, or (b) it can help to determine the power of the test and the validity of the instrumentation used to conduct the test.

In summary, whenever researchers test the theory of ontological expressiveness, we believe they will have to test for the presence of variance in their variables. In such cases, they may be subject to the criticism that the results are obvious. By taking the steps above (i.e., describing why the test is relevant in practice, why it is not obvious, and why it has value empirically), researchers can explain why this criticism is misplaced.

Testing for the absence of variance. When testing the theory of ontological expressiveness, researchers will typically use tests for the absence of variance as a way to "control" for possible confounds (i.e., threats to internal validity). According to Parsons and Cole (2005), a major confound in some past studies that used the theory of ontological expressiveness was that they failed to ensure the scripts they compared were informationally equivalent.

Given the importance ascribed to the notions of informational equivalence and computational equivalence in Parsons and Cole (2005) and other recent studies (e.g., Siau, 2004, Gemino and Wand, 2004, Aguirre-Urreta and Marakas, 2008, Poels et al., in press), we briefly explain how these concepts might apply to studies that use the theory of ontological expressiveness.

Informational equivalence and computational equivalence are notions that can be used to describe scripts. As a result, ontological evaluations of grammars do not engage these notions directly. Nonetheless, ontological evaluations of grammars have implications for predictions about the informational and computational equivalence of scripts produced using the grammars. By choosing an ontological benchmark to evaluate a grammar, three outcomes can be achieved.

First, the benchmark can be used to predict when alternative scripts that have been prepared to describe a domain are *not* informationally equivalent (at least in a denotational sense). Specifically, if alternative scripts contain different instances of construct overload, excess, and

deficit, they are *not* informationally equivalent according to the ontological benchmark. This outcome occurs because:

- In the case of construct overload, one script will make ontological distinctions that are not present in the other script (e.g., distinctions between things and events).
- In the case of construct excess, one script contains information that does not map to an ontological construct.
- In the case of construct deficit, one script contains less information than the other script.

Second, from a *denotational* perspective, the ontological benchmark can be used to gain insights into the implications of the lack of informational equivalence among the scripts. The nature of the differences among scripts that arise because of construct overload, excess, and deficit foreshadow the types of problems readers are likely to encounter when they try to understand the scripts. Judgments or theory-based predictions can then be made about the likely seriousness of these problems. Such judgments or predictions can be tested empirically.

Third, the ontological benchmark can be used to predict when alternative scripts are not *computationally equivalent*. Specifically, if two scripts are identical except that one has instances of construct redundancy, then the two scripts are informationally equivalent in a *denotational* sense because they reflect the same ontological information. Nonetheless, researchers might predict that readers who are given the script that contains instances of construct redundancy will take longer to interpret the script. The reason is that readers will have to expend cognitive resources to decide whether the different grammatical constructs represent the same ontological construct or different ontological constructs.

These outcomes have important implications for the design of studies that test the theory of ontological expressiveness. Specifically, informational equivalence and computational

equivalence are relevant only when a researcher is testing readers' interpretations of scripts. In such studies, if researchers wish to examine whether the presence of construct *overload*, *excess*, or *deficit* in scripts affects readers' interpretations, the scripts in the study must *not* be informationally equivalent according to the ontological benchmark chosen. Otherwise, the study will not have construct validity (because the scripts will not reflect differences in construct overload, excess, or deficit). On the other hand, if researchers wish to examine whether the presence of construct *redundancy* in scripts affects readers' interpretations, the scripts examined must be informationally equivalent from a denotational perspective according to the ontological benchmark chosen. If the scripts are not informationally equivalent, there will be a lack of construct validity (because construct redundancy has not been manipulated properly) as well as a lack of internal validity (because another variable must have been manipulated to cause differences in the information content of the scripts).

Moreover, in the latter type of study, researchers should not be required to "prove" that the scripts in their study are informationally equivalent, because this standard is impossible to meet. Instead, they should explain the steps they took to *maximize* the extent to which the two scripts were informationally equivalent from a *denotational* perspective and provide evidence that the scripts are indeed maximally equivalent. As we noted earlier, Kim et al. (2000) and Gemino and Wand (2005) provide examples of how these steps might be done.

In both types of studies, researchers should also consider a range of *pragmatic* factors that might lead readers of the scripts to engage different connotational semantics. For example, researchers may propose that differences in denotational semantics caused by construct overload, excess, or deficit will have no significant impact on readers with substantial knowledge of the domain shown in the script. Similarly, researchers may predict that the additional computation caused by construct redundancy will have little impact on readers with substantial knowledge of the domain. Such predictions need to be tested empirically because the effects of

pragmatics often are difficult to predict. It should not be assumed that the hypothesized existence (or lack thereof) of informational or computational equivalence between alternative scripts of a domain will always be manifested in users' performance with the scripts.

SOME GUIDELINES REVISITED

Several studies have offered frameworks (Wand and Weber, 2002, Gemino and Wand, 2004), concepts (Siau, 2004), and guidelines (Parsons and Cole 2005, Aguirre-Urreta and Marakas, 2008) to assist researchers who wish to evaluate a conceptual modeling grammar empirically. In the subsequent sections and their corresponding tables (Tables 2a and 2b), we briefly discuss how our guidelines relate to these prior guidelines.

Parsons and Cole (2005)

Parsons and Cole (2005) propose guidelines for the design of experimental work to evaluate conceptual modelling "techniques." They focus on "read" studies, in which researchers test the ability of individuals to understand the semantics in alternative scripts that represent a domain. Their guidelines are intended to "assist in developing experimental materials that support meaningful tests of domain semantics" (Parsons and Cole, 2005, p. 327). In Table 2a, we summarize their guidelines and note the ways in which theirs guidelines agree with or differ from our guidelines. Rather than discuss each guideline in depth, we focus here on the main spirit of their guidelines. Specifically, we believe a major difference between their guidelines and our guidelines is that their guidelines are designed for a specific type of study in which a researcher aims to:

(a) compare scripts that differ only in syntax (e.g., the symbols used or the arrangement of symbols in a script) rather than semantic or pragmatic characteristics;

(b) compare readers' interpretations only in terms of the denotational semantics they infer from the scripts.

Some studies will have these aims. In such studies, Parsons and Cole's guidelines will be relevant. For example, researchers may wish to study the impact of *grammatical syntax* on *interpretational efficiency*. In such a study, to the extent possible researchers should control for semantic and pragmatic factors. Parsons and Cole's guidelines seek to ensure that the semantics in the scripts are equivalent (i.e., only the syntax differs). They focus the experimental tasks as much as possible on the denotational semantics in the scripts to reduce the risk that pragmatic and connotational issues confound the results.

While we agree that Parsons and Cole's guidelines are relevant in some contexts, they will not apply in many other contexts. For example, as noted earlier, researchers who test the theory of ontological expressiveness will often need to create scripts that contain *different* semantics according to the ontological benchmark used. Moreover, they may be interested in a variety of pragmatic factors.

We also disagree with Parsons and Cole's contention that researchers should not test predictions if they appear, at first, to be obvious. We believe this concern was their primary motivation for advising that researchers ensure the scripts they compare are informationally equivalent (Parsons and Cole, 2005, p. 330). Other researchers have also espoused this belief. For example, Gemino and Wand (2004, p. 257) write: "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 ..."

In contrast to these views, we believe that concerns over the *a priori* "obviousness" of results are misplaced. Often in science, the aim is to confirm what we think we know. If a study is

designed well, the results are valuable whether the expectation is confirmed or disconfirmed. Indeed, for a disconfirmation to be truly surprising, a study has to be designed to confirm the expected. Overall, we believe that researchers should strive to test hypotheses that are relevant for practice and that will contribute to research via theory or methodology. Whether the results are surprising is a secondary consideration.

Table 2a: Consideration of Prior Guidelines – Parsons and Cole (2005)

Type of Guidelines: Guidelines for studies that examine readers' interpretations of alternative scripts of a domain.

Guidelines	**Comments**
1. Alternative scripts should be informationally equivalent.	Agree for some studies only: In some studies, it may be desirable to have scripts that are informationally equivalent. On such occasions, researchers should explain the steps they took to maximize the equivalence of their scripts and present evidence regarding their practical equivalence. In other studies, however, informational equivalence will not be a relevant concept.
2. Measure performance based only on semantics in script.	Agree for some studies only: In studies focusing on denotational semantics only, performance measures should focus on the denotational semantics in the script. Researchers should ensure these semantics are salient for participants in the study. In studies focusing on pragmatics and connotational semantics, however, performance should not be based solely on the semantics in the script. For example, such studies may also be interested in the connotational semantics that readers can infer from scripts.
3. Do not use subject matter experts.	Agree for some studies only: In studies focusing on denotational semantics, novices in a domain are desirable participants because they lack domain knowledge. Thus, they are more likely to be influenced by the denotational semantics in the script. For studies focusing on pragmatics and connotational semantics, however, subject matter experts may be

Table 2a: Consideration of Prior Guidelines – Parsons and Cole (2005)	
	desirable participants.
4. Participants should have scripts when they answer questions.	Agree for some studies only: In studies focusing on denotational semantics, it may be useful for participants to have scripts when they answer questions. For studies focusing on pragmatics and connotational semantics, however, it may be useful to remove scripts from participants prior to asking them questions because the intent is not to focus solely on the denotational semantics in the script.

Aguirre-Urreta and Marakas (2008)

Aguirre-Urreta and Marakas's (2008) guidelines for script creation and script interpretation are motivated by the lack of clear results that have been obtained in past research that has compared an entity-relationship grammar with an object-oriented grammar. As noted in Table 2b, we agree with many of their recommendations. Nonetheless, we consider their advice to measure informational equivalence and computational equivalence to be problematical because in many studies informational equivalence and computational equivalence are not applicable. Moreover, we believe they cannot be measured conclusively. Therefore, in studies where researchers need to verify informational equivalence or computational equivalence, they should not be required to "prove" equivalence. Rather, they should explain the steps they took to maximize the equivalence of their scripts and provide evidence to justify the practical or near-equivalence of the scripts.

Table 2b: Consideration of Prior Guidelines – Aguirre-Urreta and Marakas (2008)
Type of Guidelines: Guidelines for studies that compare grammars in terms of their effectiveness and efficiency in supporting script creation and script interpretation.

Guidelines	Comments
1. Conduct comparative analyses of the ontological expressiveness of alternative grammars.	Agree: When evaluating a grammar empirically, the ontological expressiveness of a grammar could serve as a useful predictor or control variable. Other theories could also be used to examine the expressiveness of a grammar.
2. Control for, or directly investigate, the modeling experience of the modelers and readers in the study.	Agree: The modeling experience of the modeler and/or reader can be an important pragmatic factor. Depending on the study, it might be a predictor variable or a control variable.
3. Control for, or directly investigate, the individual differences of the modelers and readers in the study.	Agree: Individual difference variables (such as cognitive ability of the modeler or reader) can be important pragmatic factors. Depending on the study, it might be a predictor variable or a control variable.
4. Measure the informational equivalence and computational equivalence of the scripts created using alternative grammars.	Agree in part (for some studies only): Informational equivalence and computational equivalence cannot be measured conclusively. If they are relevant concepts in a given study, researchers should explain the steps they took to maximize the equivalence of their scripts and present evidence regarding their practical equivalence.
5. Distinguish between the modeling technique used to create a script and the modeling practices used to create a script with that technique.	Agree: When comparing grammars, researchers should clarify whether they are comparing the grammars alone or also the practices that exist for using them. A given grammar can be used in different ways. The ways in which a grammar is used can affect the syntax and semantics presented in scripts using the grammar.

CONCLUSIONS

The short history of empirical work to evaluate theoretical predictions about the merits of alternative conceptual modeling grammars and scripts has shown that researchers face major

challenges if they are to mitigate threats to internal, external, and construct validity (see, e.g., Siau, 2004; Gemino and Wand, 2004). In this regard, the insights and guidelines provided by past researchers are laudable, because they provide an important platform for further debate on how empirical work on conceptual modeling grammars and scripts might be improved.

In this paper, we proposed a set of guidelines to support researchers who wish to evaluate conceptual modeling grammars empirically. The issues addressed in our guidelines (variable selection, theory and design, and operationalization and measurement) are not limited to a particular theory or methodology. Instead, they are designed to support conceptual modeling research in general. To show how they could be used by researchers, we illustrated how they could apply to studies that use the theory of ontological expressiveness to evaluate conceptual modeling grammars. Our guidelines also help to clarify issues that have been unclear in past research. For example, several studies in the past have advised researchers to ensure that the conceptual modeling scripts they compare in their studies are informational equivalent. We explained why this advice is appropriate for some studies but inappropriate for others. For studies in which informational equivalence is a desirable property of scripts, we explained how researchers should address this concept in their work.

Like Parsons and Cole (2005, p. 340), we see our "work as part of an ongoing dialogue." Some researchers may disagree, for example, with our assessment of the need sometimes to test for outcomes that appear, at first glance, to be obvious. Such researchers might explain why our views are misplaced and recommend alternative guidelines in their place. Other researchers might agree with our guidelines but see ways to extend them. Certainly, our guidelines are limited and could be extended in various ways. For example, our guidelines primarily address the internal validity and construct validity of empirical tests. We addressed external validity only in a limited way (in relation to incorporating pragmatic factors in empirical tests) and did not address statistical conclusion validity at all. Future studies could develop a more complete set

of guidelines that address the full range of validities required in empirical research.

Despite these limitations, we believe our work has several implications for future research. First, we have shown why researchers need to be circumspect when they rely on the concepts of informational equivalence and computational equivalence. In particular, we have pointed out why researchers ought to take great care when they claim (sometimes dogmatically) that informational equivalence or computational equivalence are needed or exist in empirical studies. Second, we addressed the related issue of testing predictions that appear, *a priori*, to be obvious. We explained why researchers should seek to examine important and relevant problems, even if answers to the problems seem 'obvious' at first glance. Clearly, more theoretical work and more exploratory studies of conceptual modeling in practice are needed to identify important, relevant phenomena. Third, we highlighted the important role that connotational semantics and pragmatics play when users seek to understand conceptual modeling scripts. To date, few studies have investigated conceptual modeling phenomena associated with connotational semantics and pragmatics (e.g., Siau et al., 1997; Khatri et al., 2006). Given the importance of these concepts, more work needs to be done. Likewise, most research that has evaluated conceptual modeling grammars has focused on the *main* effects of syntax, semantics, or pragmatics (see Table 1). Many opportunities exist to extend this research by examining how these factors *interact* during the script creation and script interpretation processes. Finally, we have evaluated and refined guidelines offered in recent research (Parsons and Cole, 2005; Aguirre-Urreta and Marakas, 2008). Hopefully, our work will facilitate the conduct of higher-quality theoretical and empirical research on this important topic.

Acknowledgements: An earlier version of this paper was presented at the *Sixth Annual Symposium on Research in Systems Analysis and Design* (AIS SIG-SAND), Tulsa OK, 2007. We are indebted to Andrew Gemino, Jeff Parsons, Jan Recker, and Iris Vessey for helpful discussions on several of the issues canvassed in this paper, and to Fei Sun for his help coding articles. The paper also benefited from comments from seminar participants at Queensland

University of Technology, the University of British Columbia, and participants at the 2007 SIG-SAND Symposium. The research was supported by funds from the Natural Sciences and Engineering Research Council of Canada to two of the authors.

References

Agarwal, R., De. P., and Sinha, A. "Comprehending object and process models: An empirical study," *IEEE Transactions on Software Engineering* (25:4) 1999, pp. 541-556.

Aguirre-Urreta, M.I. and Marakas, G. "Comparing conceptual modeling techniques: A critical review of the EER vs. OO empirical literature," *The DATA BASE for Advances in Information Systems* (39:2) May 2008, pp. 9-32.

Allen, G. and March, S.T. "The effects of state-based and event-based data representation on user performance in query formulation tasks," *MIS Quarterly* (30:2) 2006, pp. 269-290.

Allen, G. and Parsons, J. "Is query reuse potentially harmful? Anchoring and adjustment in adapting existing database queries," *Information Systems Research* in press 2008.

Angeles, P. A. *Dictionary of Philosophy*. Harper Perennial, New York, NY,: 1981.

Becker, J., Niehaves, B. and Pfeiffer, D. "Ontological evaluation of conceptual models: A linguistic interpretivist approach," *Scandinavian Journal of Information Systems* (20:2) 2008, pp. 83–110.

Berners-Lee, T., Hendler, J., and Lassila, O. "The semantic web," *Scientific American* (284:5) 2001, pp. 34-43.

Bodart, F., Patel, A., Sim, M., and Weber, R. "Should optional properties be used in conceptual modeling? A theory and three empirical tests," *Information Systems Research* (12:4) Dec 2001, pp. 384-405.

Bowen, P.L., O'Farrell, R.A., and Rohde, F.H. "Analysis of competing data structures: Does ontological clarity produce better end user query performance," *Journal of the Association for Information Systems* (7:8) 2006, pp. 514-544.

Burton-Jones, A. and Meso, P. "Conceptualizing systems for understanding: An empirical test of decomposition principles in object-oriented analysis," *Information Systems Research* (17:1) 2006, pp. 38-60.

Burton-Jones, A. and Meso, P. "The effects of decomposition quality and multiple forms of information on novices' understanding of a domain from a conceptual model," *Journal of the Association for Information Systems* (9:12) 2008, pp. 748-802.

Burton-Jones, A., and Weber, R. "Understanding relationships with attributes in entity-relationship diagrams," in *Proceedings of the 20th International Conference on Information Systems*, P. De and J. I. DeGross (eds.), Charlotte, NC, 1999, pp. 214-228.

Corral, K., Schuff, D., and St. Louis, R.D. "The impact of alternative diagrams on the accuracy of recall: A comparison of star-schema diagrams and entity-relationship diagrams," *Decision Support Systems* (42) 2006, pp. 450-468.

Gemino, A., and Wand, Y. "A framework for empirical evaluation of conceptual modeling techniques," *Requirements Engineering* (9:4) 2004, pp. 248-260.

Gemino, A., and Wand Y. "Complexity and clarity in conceptual modeling: Comparison of mandatory and optional properties," *Data & Knowledge Engineering* (55:3) 2005 pp. 301-326.

Gomez-Perez, A., Fernandez-Lopez, M., and Corcho, O. *Ontological engineering*, Springer-Verlag, London, 2004.

Hadar, I., and Soffer, P. "Variations in conceptual modeling: Classification and ontological analysis," *Journal of the Association for Information Systems* (7:8) 2006, pp. 568-592.

Hirschheim, R., Klein, H., and Lyytinen, K. *Information systems development and data modeling: Conceptual foundations and philosophical foundations*. Cambridge, UK: Cambridge University Press, 1995.

Irwin, G., and Turk, D. "An ontological analysis of use case modeling grammar," *Journal of the Association for Information Systems* (6:1) 2005, pp. 1-36.

Kent, W. *Data and reality*, 1st Books Library, 2000 (originally published 1978).

Khatri, V., Vessey, I., Ramesh, V., Clay, P., and Park, S.-J. "Understanding conceptual schemas: Exploring the role of application and IS domain knowledge," *Information Systems Research* (17:1) 2006, pp. 81-99.

Kim, J., Hahn, J., and Hahn, H. "How do we understand a system with (so) many diagrams? Cognitive integration processes in diagrammatic reasoning," *Information Systems Research* (11:3) 2000, pp. 284-303.

Krogstie, J., Sindre, G., and Jørgensen, H. "Process models representing knowledge for action: A revised quality framework," *European Journal of Information Systems* (15) 2006, pp. 91-102.

Kung, C.H., and Sølvberg, A. "Activity modelling and behavior modelling," in: *Information systems design methodologies: Improving the practice,* T.W. Olle, H.G. Sol and A.A. Verrijn-Stuart (eds.), IFIP, Amsterdam, North-Holland, 1986, pp. 145-171.

Larkin, J., and Simon, H. "Why a diagram is (sometimes) worth ten thousand words," *Cognitive Science* (11) 1987, pp. 65-99.

Lee, A.S. and Baskerville, R.L. "Generalizing generalizability in information systems research," *Information Systems Research* (14: 3) September 2003, pp. 221-243.

Lindland, O.I., Sindre, G. and Sølvberg, A. “Understanding quality in conceptual modeling.” *IEEE Software* (11) 1994, pp. 42–49.

Maes, A. and Poels, G. “Evaluating quality of conceptual modeling scripts based on user perceptions,” *Data & Knowledge Engineering* (63) 2007, pp. 701-724.

Moody, D.L. "Dealing with complexity in information systems modeling: Development and empirical validation of a method for representing large data models, *Proceedings of the 24th International Conference on Information Systems*, Seattle WA, 2003, pp. 207-221.

Nordbotten, J.C. and Crosby, M.E. “The effect of graphic style on data model interpretation,” *Information Systems Journal* (9) 1999, pp. 139-155.

Norman, D. “Cognitive Engineering,” in *User Centered Design: New Perspectives on Human Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, 1986, pp. 31-61.

Parker, F. and Riley, K. *Linguistics for non-linguists: A primer with exercises*, 4th ed. Pearson/Allyn and Bacon, Boston, 2005.

Parsons, J. “Effects of local versus global schema diagrams on verification and communication in conceptual data modeling,” *Journal of Management Information Systems* (19:3) 2003, pp. 155-184.

Parsons, J. and Cole, L. “What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques,” *Data & Knowledge Engineering* (55) 2005, pp. 327-342.

Parsons, J. and Wand, Y. “Using cognitive principles to guide classification in information systems modeling,” *MIS Quarterly* (32:4) 2008, pp. 839-868.

Patel, V.L., Allen, V.G., Arocha, J.F., and Shortliffe, E.H. "Representing clinical guidelines in GLIF: Individual and collaborative expertise," *Journal of the American Medical Informatics Association* (5:5) 1998, pp. 467-483.

Poels, G., Maes, A., Gailly, F., and Paemeleire, R. "The pragmatic quality of Resources-Events-Agents diagrams: An experimental evaluation," *Information Systems Journal*, in press, Published online: Nov 13 2007, 27 pp.

Recker, J. *Understanding Process Modelling Grammar Continuance: A Study of the Consequences of Representational Capabilities*, Unpublished Doctoral Dissertation, Queensland University of Technology, Australia, 2008.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy. F., and Lorensen, W. *Object-oriented modeling and design*, Prentice Hall, Englewood Cliffs NJ, 1991.

Rumbaugh, J., Jacobson, I., and Booch, G. *Unified Modeling Language reference manual*, 2nd ed., Addison-Wesley, Boston, 2005.

Scheer, A.W. *ARIS – Business Process Frameworks*, 3rd Edition, Springer-Verlag, 1999.

Shanks, G., Tansley, E., Nuredini, J., Tobin, D., and Weber, R. "Representing part-whole relations in conceptual modeling: An empirical evaluation," *MIS Quarterly* (32:3) 2008, pp. 553-573.

Siau, K. "Informational and computational equivalence in comparing conceptual modeling methods," *Journal of Database Management* (15:1) 2004, pp. 73-86.

Siau, K. and Rossi, M. "Evaluation techniques for systems analysis and design modelling methods – A review and comparative analysis," *Information Systems Journal*, in press (Published Online: Dec 21 2007), 20 pp.

Siau, K. and Tian, Y. "A semiotic analysis of unified modeling language graphical notations," *Requirements Engineering* (14) 2009, pp. 15-26.

Siau, K., Wand, Y., and Benbasat, I. "The relative importance of structural constraints and surface semantics in information modeling," *Information Systems* (22:2&3) 1997, pp. 155-170.

Soffer, P. and Hadar, I. "Applying ontology-based rules to conceptual modeling: A reflection on modeling decision making," *European Journal of Information Systems* (16:5) 2007, pp. 599-611.

Vessey, I. Cognitive fit: A theory-based analysis of the graphs versus tables literature, *Decision Sciences* (22:2) 1991, pp. 219-240.

Wand, Y. and Weber, R. "An ontological model of an information system," *IEEE Transactions on Software Engineering* (16:11) 1990, pp. 1282-1292.

Wand, Y. and Weber, R. "On the ontological expressiveness of information systems analysis and design grammars," *Journal of Information Systems* (3) 1993, pp. 217–237.

Wand, Y. and Weber R. "Information systems and conceptual modeling - A research agenda," *Information Systems Research* (13:4) 2002, pp. 363-376.

Yourdon, E. *Modern structured analysis*, Prentice Hall, Englewood Cliffs, NJ, 1989.

APPENDIX 1: CLASSIFYING CONCEPTUAL MODELING RESEARCH

We populated Table 1 in the following way. First, an independent coder who had completed an M.Sc. thesis on conceptual modeling was asked to scan the titles, abstracts, and contents of each paper in the sample provided and to identify all papers that related to conceptual modeling. He was then asked to identify the subset of these papers that evaluated a conceptual modeling grammar. To ensure none were missed, we asked the coder to perform both steps as liberally as possible. He was to include papers in each set even when they related only tenuously to the topic. Of the 1602 papers in the sample, he classified 35 papers as relating to conceptual modeling, 13 of which he then designated as having evaluated a modeling grammar.

Two of the authors and the independent coder then read the 13 studies and mapped them to Table 1. Because there were some differences in our classifications, we devised heuristics to improve the reliability of our coding. The two authors and the independent coder then reclassified the papers using the heuristics. The classifications between the authors and the independent coder were reliable; they were identical for 11 of 13 articles (85 percent agreement) and differed only slightly for the other two articles. These minor differences in coding were then resolved through discussion. We ultimately concluded that only seven articles in our sample empirically evaluated a conceptual modeling grammar. Table A1 summarizes how we classified these seven papers according to the dimensions of our framework.

We describe the heuristics that we used to code articles, together with examples, below. We also provide a table (Table A1) that explains our coding of each article. We provide these details to ensure that our coding process is transparent for the reader and to provide heuristics that other researchers might find useful when classifying or reading conceptual modeling work.

Coding Heuristics and Examples

We used the first two heuristics to help us classify the content of a paper:

1. *Author objectives versus study details*: We coded papers based on our reading of the study, rather than according to the objectives stated by the author. For example, we coded the study by Khatri et al. (2006) as an evaluation of a conceptual modeling grammar even though the authors did not state explicitly that this was an objective of their study.

2. *Major issues versus minor issues*: When coding papers, we considered the apparent significance of issues described in the paper. We used three levels of significance: major, minor, and very minor. We coded an article as having examined a factor if it did so in a major or minor way, but not if it only examined it in a very minor way. For example, the main issue examined in the study by Bodart et al. (2001) was the effect of semantics. In one of three experiments in that study, however, the authors also manipulated a pragmatic factor – task complexity. (Task complexity is a pragmatic factor because it could affect the cognitive process undertaken by a reader of the script.) In their statistical tests, the authors tested for *both* the main effect of task complexity and the interaction effect between task complexity and the effect of semantics. In their description of their experiment, however, they did not explain the nature of the interaction between these factors. Moreover, in their results section, they only focused on the *main* effect of task complexity (p. 396). Therefore, for this study, we coded the main effect of semantics as the major issue, the main effect of pragmatics as a minor issue, and the interaction between semantics and pragmatics as a very minor issue that was not counted in our classification. A similar decision was made when classifying Soffer and Hadar (2007).

We used the next heuristic to assess whether a paper was a 'conceptual modeling' paper:

3. *Conceptual modeling versus data modeling*: We coded papers as 'conceptual modeling' if the models in the paper were models of a real-world domain (whether physical or social) or if the empirical tests in the paper focused on whether individuals could obtain an understanding of a real-world domain from the model. We coded papers as 'data modeling' if the models in the paper were models of a database or database view or if the empirical tests focused on whether individuals could derive an understanding of the database or database view from the model. For example, based on these heuristics, we coded Nordbotten and Crosby (1999), Allen and March (2005), and Bowen et al. (2006) as 'data modeling' articles.

Heuristic 4 helped us to assess whether a study empirically evaluated a conceptual modeling grammar:

4. *Grammars versus methods*: Some papers examined readers' abilities to interpret scripts. In these papers, if the differences in the scripts stemmed from differences in one or more characteristics of a grammar (syntax, semantics, or pragmatics), we coded the study as an evaluation of a grammar. If the differences stemmed from issues not prescribed in the grammar, however, we coded the paper as not being an empirical evaluation of a grammar. More specifically, we identified several cases where the differences could be attributed to the method of using the grammar rather than to the characteristics of the grammar. For example, Bodart et al. (2001) examined readers' abilities to interpret scripts that either did or did not have optional properties. We viewed this as a comparison of two grammars: a grammar that advocated optional properties and a grammar that proscribed them. Accordingly, we coded the study as an evaluation of a grammar. In contrast, Parsons (2003) compared readers' interpretation of scripts that reflected portions of a domain (local schemas) with readers' interpretation of scripts that reflected an entire domain (global schemas). The differences in the scripts did not stem from differences in the syntax, semantics, or pragmatics of the grammar.

Rather, they implicitly stemmed from differences in methods used to create scripts. For instance, they were methods that advised modelers to produce a script of the entire domain, or they were methods that advised modelers to produce scripts of portions of the domain. As a result, we did not code this study as an evaluation of a grammar. We coded the studies by Burton-Jones and Meso (2006, 2008) in a similar fashion.

For most studies that we coded as having empirically evaluated a conceptual modeling grammar, it was easy to map them into Table 1. Nonetheless, for some studies the mapping was still unclear. For these studies, we used the last two heuristics:

5. *Grammatical rules versus modeling rules*: Some papers examined rules for creating conceptual modeling scripts. We coded these papers as an evaluation of a conceptual modeling grammar if the rules related closely to one or more elements of the grammar (syntax, semantics, or pragmatics). For example, Soffer and Hadar (2007) examined rules for creating conceptual modeling scripts. The rules suggested how to map specific phenomena into specific grammatical constructs (that is, the rules prescribed semantics for the grammar). As a result, their study essentially compared two grammars: a grammar that offered prescribed semantics, and a grammar that did not offer prescribed semantics. Accordingly, we coded their paper as an empirical evaluation of a conceptual modeling grammar. We coded Hadar and Soffer (2006) and Parsons and Wand (2008) in a similar fashion.

6. *Interpretation fidelity versus a combination of representation fidelity and interpretation fidelity*: Researchers might examine readers' interpretation of a *script* or readers' interpretation of a *domain* shown in a script. Variations in readers' interpretations of a *script* reflect variation in interpretation fidelity, but variations in readers' interpretation of a *domain* could reflect variations in representation fidelity (if the scripts vary in how well they represent the domain) and/or

variations in interpretation fidelity (if readers interpret the same script differently). We accounted for both types of interpretations. For example, in Bodart et al. (2001), one dependent measure was readers' ability to recall elements of a conceptual modeling script. Variations on this measure reflected variations in interpretation fidelity. Another dependent measure in that study, however, was readers' ability to infer information about the domain shown in the script (assessed via readers' answers to inferential problem-solving questions). Variations on this measure could reflect variations in both representation fidelity and interpretation fidelity.

Table A1: Coding of Research Studies				
Study	***Conceptual modeling?***	***Evaluation of a grammar?***	***Predictor variables?***	***Outcome variables?***
1. Nordbotten and Crosby (1999)	No. The models in the study were data models, which the authors described as models used in database design "to specify the information objects, their interrelationships, and the constraints required by the application system" (p. 140).	NA	NA	NA
2. Kim et al. (2000)	Yes. The models in the study were models of a business domain.	Yes. The grammars were evaluated implicitly in terms of syntax and semantics.	Main effects of syntax and semantics. Participants in their experiment received sets of diagrams that varied in the similarity of their syntax (e.g., using nodes and arcs) and that also varied in semantics. The authors argued that the diagrams were informationally equivalent, but our reading of the diagrams in their paper suggests that their semantics differed.	Representation fidelity, interpretation fidelity, and interpretational efficiency. Participants answered problem-solving questions about the domain shown in the scripts. Because the scripts differed in how well they represented the domain, this test measured both representational fidelity and interpretational fidelity. The authors also tested participants' difficulty in reading the diagrams (interpretational efficiency).
3. Bodart et al. (2001)	Yes. The models in the study were models of a business domain.	Yes. Two alternative grammars were evaluated in terms of readers' ability to interpret scripts created using those grammars.	Major factor: Main effect of semantics. Participants in their experiment received diagrams that were produced with an ER grammar that used optional properties or an ER grammar that proscribed optional properties. Minor factor: Main effect of pragmatics. In one of three experiments, the	Major outcomes: Representation fidelity and interpretation fidelity. The problem-solving questions given to participants required them to understand the domain shown in the diagrams. Because the diagrams could differ in how completely they represented the domain, this test examined both representational fidelity and interpretational fidelity. Minor outcome: Interpretational

Table A1: Coding of Research Studies				
Study	***Conceptual modeling?***	***Evaluation of a grammar?***	***Predictor variables?***	***Outcome variables?***
			authors controlled for the complexity of the interpretation task. This is a pragmatic factor because it affects the reader's cognitive process when interpreting the models.	efficiency. In one of three experiments, the authors measured the time taken to complete the task as a measure of interpretational efficiency.
4. Parsons (2003)	Yes. The models in the study were models of an imagined domain.	No. The paper evaluated alternative scripts of a domain. It could also be viewed as an implicit evaluation of two methods for creating scripts of a domain: a method that produced global scripts, and a method that produced local scripts.	NA	NA
5. Allen and March (2005)	No. The models in the study were database views ("logical level constructs that provide … users with… conceptualizations of [a] database" p. 270). The paper studied how such views "affect a user's ability to understand the database" (p. 269).	NA	NA	NA
6. Bowen et al. (2006)	No. The authors set out to study whether findings from research on "conceptual models" apply to research on "implementation (logical) data models" (p. 514).	NA	NA	NA
7. Burton-Jones and Meso (2006)	Yes. The models in the study were models of a domain.	No. The paper evaluated alternative scripts of a domain. It could also be viewed as an implicit evaluation	NA	NA

Table A1: Coding of Research Studies				
Study	***Conceptual modeling?***	***Evaluation of a grammar?***	***Predictor variables?***	***Outcome variables?***
		of two methods for creating scripts that varied in the extent to which they produced scripts that manifested a good decomposition of the domain.		
8. Hadar and Soffer (2006)	Yes. The models in the study were models of a domain.	Yes. The paper implicitly evaluated the UML class diagram grammar. The authors argued that if the constructs in a grammar are not well defined, modelers might use these constructs differently when modeling a domain.	Main effect of semantics. The UML grammar does not specify mappings between grammatical constructs and real world constructs. The authors examined whether the lack of a prescribed mapping could lead to variations in scripts of a domain.	Representational fidelity. Variations among scripts reflect differences in the completeness, accuracy, or coverage of the domain being modeled.
9. Khatri et al. (2006)	Yes. The models in the study were models of a domain.	Yes. The paper implicitly evaluated the ER and EER grammars. The evaluation focused on whether readers' prior knowledge affected their ability to understand scripts created in these grammars.	Main effect of pragmatics. Readers' background knowledge is a pragmatic factor. The authors showed that background knowledge affected readers' interpretations of scripts created in the ER and EER grammars.	Interpretational fidelity. The authors measured readers' understanding of ER and EER scripts by giving them schema-based problem-solving questions that checked how well the readers understood the scripts.
10. Soffer and Hadar (2007)	Yes. The models in the study were models of a domain.	Yes. The paper implicitly evaluated two grammars: a grammar with prescribed mapping rules and a grammar without prescribed mapping rules.	Major factor: Main effect of semantics. The paper examined whether giving modelers grammars with prescribed mappings would reduce variation among modelers' scripts of a domain. Minor factor: Main effect of pragmatics. The authors controlled for modelers' knowledge of the domain and interviewed participants to determine its possible effect.	Representational fidelity. Variations among scripts reflect differences in the completeness, accuracy, or coverage of the domain being modeled.

Table A1: Coding of Research Studies				
Study	***Conceptual modeling?***	***Evaluation of a grammar?***	***Predictor variables?***	***Outcome variables?***
11. Burton-Jones and Meso (2008)	Yes. The models in the study were models of a domain.	No. The paper evaluated alternative scripts of a domain. It could also be viewed as an implicit evaluation of two methods for creating scripts that varied in the extent to which they produced scripts that manifested a good decomposition of the domain.	NA	NA
12. Parsons and Wand (2008)	Yes. The models in the study were models of a domain.	Yes. The paper evaluated whether scripts created with a grammar with semantic mapping rules would be better than scripts created with a grammar without semantic mapping rules.	Main effect of semantics. The mapping rules examined in the paper concerned the meaning of the "class" construct in conceptual modeling grammars.	Representational fidelity. The authors examined whether a script created with a grammar that followed their prescribed mapping rules would provide a better representation of a domain than a script created with a grammar without these rules.
13. Shanks et al. (2008)	Yes. The models in the study were models of a domain.	Yes. Two alternative grammars were evaluated in terms of readers' ability to interpret scripts created using those grammars.	Main effect of semantics. Participants in their study received scripts that were produced either with an ER grammar that showed parts and wholes as entities or an ER grammar that showed parts and wholes via relationships among entities.	Major outcomes: Representation fidelity and interpretation fidelity. Participants answered problem-solving questions about the domain shown in the scripts. Because the scripts differed in how well they reflected the domain, this test measured a combination of representational fidelity and interpretational fidelity. Minor outcome: Interpretational efficiency. The authors also tested for differences in the time taken to understand the scripts and the difficulties they experienced in interpreting the scripts.

APPENDIX 2: EXAMPLES OF POSSIBLE RESEARCH STUDIES

In Table A2, we provide examples of possible research questions and studies that could be conducted to evaluate conceptual modeling grammars empirically.

Table A2: Examples of Possible Research Studies[1, 2]

		Process and Performance Criteria			
		Script Creation		Script Interpretation	
		Representational fidelity as outcome	**Representational efficiency as outcome**	**Interpretational fidelity as outcome**	**Interpretational efficiency as outcome**
Effect of grammar character-istics	**Effect of syntax only**	1. Can a difference in the syntax (only) of two grammars result in alternative scripts of a domain that differ in representational fidelity? For example, modelers may make fewer errors when they construct scripts using grammars that contain simpler syntax than with grammars that contain complicated syntax.	2. Can a difference in the syntax (only) of two grammars make it simpler or quicker to construct a script of a domain with representational fidelity? For example, grammars that contain construct redundancy may lead a modeler to spend more time deciding which symbol to use to represent the required denotational semantics.	3. If two scripts have the same denotational semantics, can a difference in syntax (only) affect readers' interpretational fidelity? For example, construct redundancy in a script may lead readers to believe that the differences in syntax imply different semantics.	4. If two scripts have the same denotational semantics, can a difference in syntax (only) lead readers to consume more effort or time to achieve interpretational fidelity? For example, construct redundancy in a script may cause readers to spend time trying to determine whether the differences in syntax imply different semantics.
	Effect of semantics only	5. Can a difference in the denotational semantics of two grammars result in alternative scripts of a domain that differ in representational fidelity?	6. Can a difference in the denotational semantics of two grammars lead a modeler to consume more time or effort to achieve representational	7. Will a difference in the denotational semantics shown in alternative scripts of a domain affect readers' interpretational fidelity?	8. Will a difference in the denotational semantics shown in alternative scripts of a domain affect the time/effort readers need to achieve interpretational

Table A2: Examples of Possible Research Studies[1, 2]

		Process and Performance Criteria			
		Script Creation		Script Interpretation	
		Representational fidelity as outcome	**Representational efficiency as outcome**	**Interpretational fidelity as outcome**	**Interpretational efficiency as outcome**
		For example, if one grammar contains construct deficit, a modeler may be unable to construct a model with that grammar that faithfully represents the domain.	fidelity? For example, if one grammar contains construct excess, the modeler may consume time or effort deciding not to use the excess constructs.	For example, a reader given a script that exhibits construct overload may gain a different interpretation of the script from that intended.	fidelity? For example, if a reader is given a script that exhibits construct excess, he or she may realize only after some time or effort that the excess constructs can be ignored.
	Effect of pragmatics only	9. Depending on the context in which the script is created, can modelers create alternative scripts of a domain that differ in representational fidelity? For example, if the modeler knows the reader will have little time to read a model, he or she may show only those semantics that are most critical rather than showing all semantics in the domain.	10. Depending on the context in which the script is created, can modelers consume a different amount of effort/time to create scripts that exhibit representational fidelity? For example, experienced modelers may be able to construct an accurate (high-fidelity) script more quickly or more easily than inexperienced modelers.	11. Depending on the context in which the script is read, can readers' understanding of a script differ in interpretational fidelity? For example, readers with knowledge of the domain may infer more from certain semantics in the script (i.e., gain additional correct or incorrect connotational semantics) than other readers.	12. Depending on the context in which the script is read, can it take readers a different amount of effort/time to achieve interpretational fidelity? For example, readers with knowledge of the domain shown in the script may interpret the semantics more easily than other readers.
	Interaction effect of syntax and semantics	13. Does representational fidelity depend on the syntax *and* denotational semantics available in the grammar? For example, the	14. Does the amount of effort or time that modelers consume to faithfully model a domain depend on the syntax *and* denotational semantics available	15. Does interpretational fidelity depend on the denotational semantics and the syntax in the script? For example, readers may be	16. Does the amount of effort or time that readers consume to interpret the semantics of a script depend on the syntax *and* the denotational semantics in the

Table A2: Examples of Possible Research Studies[1, 2]

		Process and Performance Criteria			
		Script Creation		Script Interpretation	
		Representational fidelity as outcome	**Representational efficiency as outcome**	**Interpretational fidelity as outcome**	**Interpretational efficiency as outcome**
		presence of construct deficit in a grammar may not matter if the grammar includes syntax that allows a modeler to annotate the script with text that describes the missing semantics.	in the grammar? For example, modelers may take more time to construct a faithful script of a domain when the grammar has construct overload (e.g., by trying to use syntax consistently), but this negative effect may be alleviated if modelers can use textual annotations to clarify the way they are using the overloaded constructs.	more able to ignore excess constructs in a script if the syntax enables the reader to clearly identify the excess constructs (e.g., through the use of color or the arrangement of excess constructs vis-à-vis other constructs in the script).	script? For example, readers may take less time or effort to realize that they can ignore excess constructs in a script if the syntax enables the reader to clearly identify the excess constructs (e.g., through the use of color or the arrangement of constructs in the script).
	Interaction effect of syntax and pragmatics	17. Does representational fidelity depend on the syntax of the grammars *and* the context in which the scripts are created? For example, while experienced modelers may be able to use simple syntax and complicated syntax equally well, inexperienced modelers may make more errors when using complicated syntax.	18. Does the amount of effort or time that modelers consume to faithfully model a domain depend on the syntax available in the grammar *and* the context in which the script is created? For example, grammars that contain construct redundancy may lead inexperienced modelers to spend time deciding which symbol to use to represent the required	19. Does interpretational fidelity depend on the syntax used in a script *and* the context in which the script is read? For example, construct redundancy in a script may lead inexperienced readers to believe that the differences in syntax imply different semantics, but readers with extensive knowledge of the domain shown in the script may realize that the	20. Does the amount of effort that readers consume to interpret the semantics of a script depend on the syntax used *and* the context in which it is read? For example, construct redundancy in a script may cause inexperienced readers to spend time trying to determine whether the differences in syntax imply different semantics, but readers with

Table A2: Examples of Possible Research Studies[1, 2]

		Process and Performance Criteria			
		Script Creation		Script Interpretation	
		Representational fidelity as outcome	**Representational efficiency as outcome**	**Interpretational fidelity as outcome**	**Interpretational efficiency as outcome**
			denotational semantics, but redundancy may not cause a problem for experienced modelers because they may simply ignore the redundant constructs.	different symbols are just different syntactic ways to represent the same type of phenomenon.	extensive knowledge of the domain shown in the script may take no time to determine that the different symbols are just different syntactic ways to represent the same phenomenon.
	Interaction effect of semantics and pragmatics	21. Does representational fidelity depend on the denotational semantics in the grammar *and* the context in which the scripts are created? For example, faced with construct excess in a grammar, inexperienced modelers may be more inclined than experienced modelers to include the excess constructs in the scripts that they create.	22. Does the amount of effort or time that modelers consume to faithfully model a domain depend on the denotational semantics in the grammar *and* the context in which the script is created? For example, faced with construct excess in a grammar, inexperienced modelers may consume more time or effort deciding not to use the excess constructs in the script.	23. Does interpretational fidelity depend on the denotational semantics in the script *and* the context in which it is read? For example, if a script contains construct overload, readers may be able to infer the correct semantics (connotationally) if they have background knowledge of the domain shown in the script.	24. Does the amount of effort or time that readers consume to interpret the semantics of a script depend on the denotational semantics used *and* the context in which it is read? For example, if a script contains construct overload, readers may be able to infer the correct semantics from it more easily or more quickly if they have background knowledge of the domain shown in the script.

1. **We only show main effects and two-way interaction effects in this table. As shown in Table 1, more complex three-way interactions are also possible, but we leave these out of this table for simplicity. Likewise, in Table 1 and in this table, we also leave out research that could investigate interactions between factors within each cell (such as the interaction of two pragmatic factors).**
2. **Shaded cells reflect studies that could be undertaken of the theory of ontological expressiveness (see also Appendix 3).**

APPENDIX 3: THE THEORY OF ONTOLOGICAL EXPRESSIVENESS

In the discipline of philosophy, ontological theories articulate a set of constructs and relationships among the constructs to describe phenomena in the real world (Berners-Lee et al., 2001; Angeles, 1981). In the context of conceptual modeling, a number of researchers have argued that such theories can be used as benchmarks to evaluate whether (a) a conceptual modeling grammar is capable of generating scripts that provide a faithful description of some real-world domain, and (b) a specific conceptual modeling script provides a faithful description of some real-world domain (Allen and March, 2006; Wand and Weber, 1993, 2002). For example, Wand and Weber (1993) argue that a conceptual modeling grammar is more "expressive" if it contains fewer of the following defects:

- *Construct overload*: A single grammatical construct maps to two or more ontological constructs. For example, an entity construct is used to reflect both events and things in a domain.

- *Construct redundancy*: Two or more grammatical constructs map to the same ontological construct. For example, an entity construct and an attribute construct are both used to represent classes of things in a domain.

- *Construct excess*: A grammatical construct does not map to any ontological construct. For example, the grammar might include constructs to model implementation-related details.

- *Construct deficit*: The grammar does not offer a construct to represent one or more ontological constructs. For example, a process modeling grammar might not contain any constructs to represent events or goals.

Conclusions about construct overload, redundancy, excess, and deficit are *theory dependent*.

In other words, they depend on the ontological theory chosen as the benchmark. A grammar deemed to have construct overload, redundancy, excess, or deficit when evaluated against one ontological theory might not be deemed to have these defects when evaluated against another ontological theory. Ideally, researchers would examine multiple ontological theories (Hadar and Soffer, 2006). Researchers might examine ontological theories from those published in the literature (such as those published in the field of philosophy). Alternatively, they might attempt to examine the lay or "commonsense" ontological theories that exist in the minds of practitioners who create or interpret conceptual models. Based upon the defects found in an ontological evaluation, a researcher can make predictions about how people use the grammar or how people use scripts created using the grammar. These predictions can then be tested empirically. Even if a grammar has theoretical deficiencies, researchers cannot know whether these deficiencies matter in practice unless they conduct empirical tests of the predictions.

Predictions about Grammars

Ontological predictions about a conceptual modeling grammar most likely will focus on how modelers use a grammar either by itself or in conjunction with other grammars to produce scripts. For instance, researchers might focus on the existence, adoption, or usefulness of strategies that can be used to avoid creating scripts that contain instances of construct overload, redundancy, excess, or deficit. Their research could be guided by social science principles (e.g., investigating the effectiveness of strategies adopted by practitioners to enhance ontological expressiveness when grammars are defective), design science principles (e.g., testing the effectiveness of strategies developed by researchers to enhance ontological expressiveness when grammars are defective), or a combination of both. For instance:

- If a grammar has construct overload, one prediction might be that experienced modelers would devise extra-grammar constructs or textual annotation so that the mapping from grammatical to ontological constructs is one-one.

- If a grammar has construct redundancy, one prediction might be that experienced modelers would devise extra-grammar rules (e.g., non-use of one of the redundant grammatical constructs) so that the mapping from grammatical to ontological constructs is one-one.

- If a grammar has construct excess, one prediction might be that experienced modelers would avoid using the excess construct because it undermines the real-world representational fidelity of the scripts they construct.

- If a grammar has construct deficit, one prediction might be that experienced modelers would devise extra-grammar constructs to cover the deficit, employ the grammar in conjunction with another one that covers the missing construct, or rely on textual annotation to "specialize" existing grammatical constructs.

Predictions about Scripts

If a conceptual modeling grammar has construct overload, redundancy, excess, or deficit (and if a modeler cannot overcome these defects), then scripts generated using the grammar may have *instances* of these defects. Where such instances exist, a theoretical prediction is that readers of the scripts will be unable to accurately, completely, and expeditiously elicit the semantics of the real-world domains represented via the scripts (Wand and Weber, 1993).

Figures A1-A4 illustrate each type of defect. Figure A1 shows two scripts that convey information about the assignment of keys to employees. As Allen and March (2006) explain, some ontological theories distinguish between events and things. From the perspective of

these theories, the script shown in Figure A1B contains *construct overload* because it uses one grammatical construct (an entity type) to represent things (keys) and events (being assigned a key). In this light, researchers might predict that some readers will find the semantics of the script in Figure A1A to be clearer than the semantics of the script in Figure A1B (because Figure A1A distinguishes between things and events). Specifically, if two readers knew little about the domain represented by the scripts, researchers might predict that the reader given the script in Figure A1A would be able explain what the term "assign" means more effectively than the reader given the script in Figure A1B. Nonetheless, if the reader shown the script in Figure A1B had good knowledge of the domain represented by the script, researchers might predict that the reader would have little difficulty explaining what the term "assign" means, because the reader could use his/her background knowledge to interpret the script.

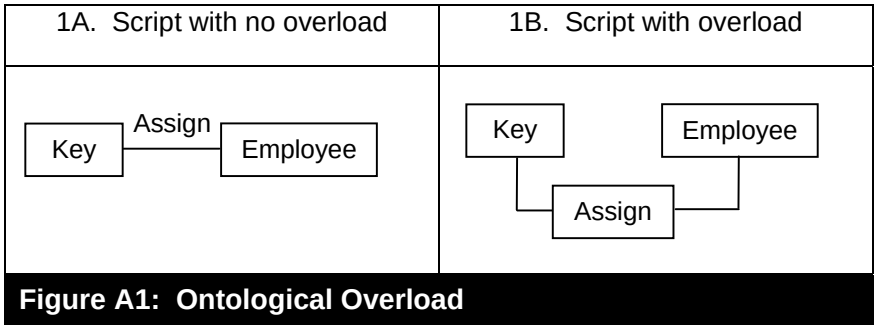


Figure A1: Ontological Overload

Note: Example adapted from Allen and March, 2006, p. 271.

Figure A2 shows an excerpt from a business process model in ARIS (Architecture of Integrated Information Systems), which is a widely used enterprise modeling approach. Note that Figure A2 has grammatical constructs that represent an abstract business process–such as a triggering event, function, and resulting event–as well as constructs that represent the implementation of the process–such as the computer hardware (workstation and CPU), machine resource (machine), and software (PPC system). Although some ontological theories

contain constructs that can be used to model implementation details (e.g., Gomez-Perez et al., 2004), others preclude them because they are deemed undesirable in conceptual models (e.g., Wand and Weber, 1990; Yourdon, 1989).

In the context of these latter theories, Figure A2 contains *construct excess* (associated with representing implementation details). Having construct excess adds denotational semantics. The implementation details are excess constructs because they would not likely map to constructs in ontological theories. In this regard, philosophical ontologies generally do not include constructs related to computer implementations. Moreover, even the commonsense ontologies used by practitioners often exclude implementation details because practitioners are generally taught to create conceptual models in an implementation-independent way (Yourdon, 1989). Nonetheless, researchers might propose that the impact of such additional information depends on the user reading the script. For novices, they might predict that the additional information will impair their ability to understand the business process, because novices may believe mistakenly that the abstract process is constrained by the particular implementation shown in the script. For experts, the researchers might predict that the additional information has no effect on their ability to understand the business process, because experts simply ignore the implementation details when reading the script.

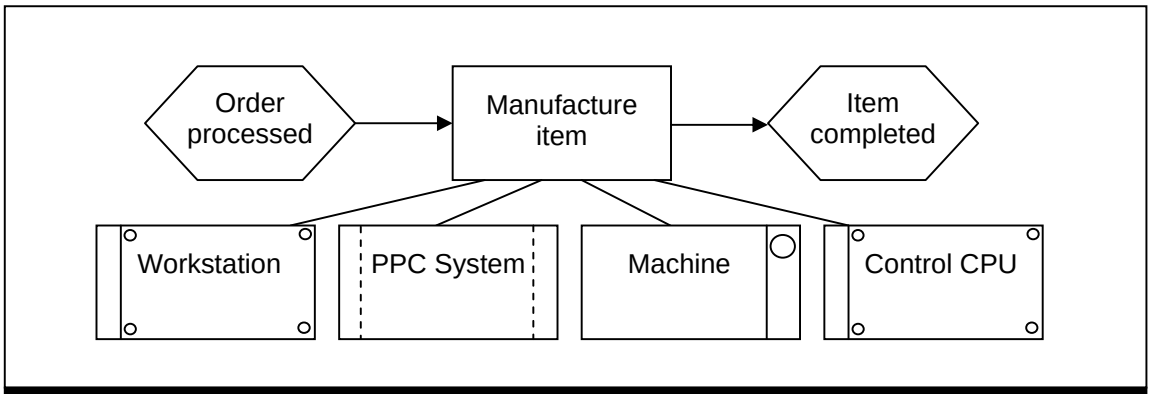


Figure A2: Ontological Excess

Note: Example adapted from Scheer, 1999, p. 19.

Figure A3 shows a script that conveys information about the processes involved in booking medical appointments. The script shows how (a) a computer system (the appointment application) operates within a work system (the medical office), and (b) actors operate within the work system (such as the receptionist) and outside it (such as the patient). Irwin and Turk (2005) explain that analysts might *wish* to show such phenomena using the use-case grammar (part of the Unified Modeling Language) (Rumbaugh et al., 2005). They cannot do so, however, because the use-case grammar lacks sufficient constructs to show how systems are decomposed. As Irwin and Turk (2005) explain, Figure A3 illustrates how use-case scripts can be *deficient* ontologically because all the information shown in Figure A3 cannot be shown in a "pure" use-case diagram. Even so, researchers might argue that the effect of construct deficit may depend on connotational and pragmatic factors. For example, if readers are experienced medical practitioners, they may consider the distinction between functions performed in the work system and those performed in the application to be self-evident. For novices, including such a distinction may be necessary if they are to understand the domain properly.

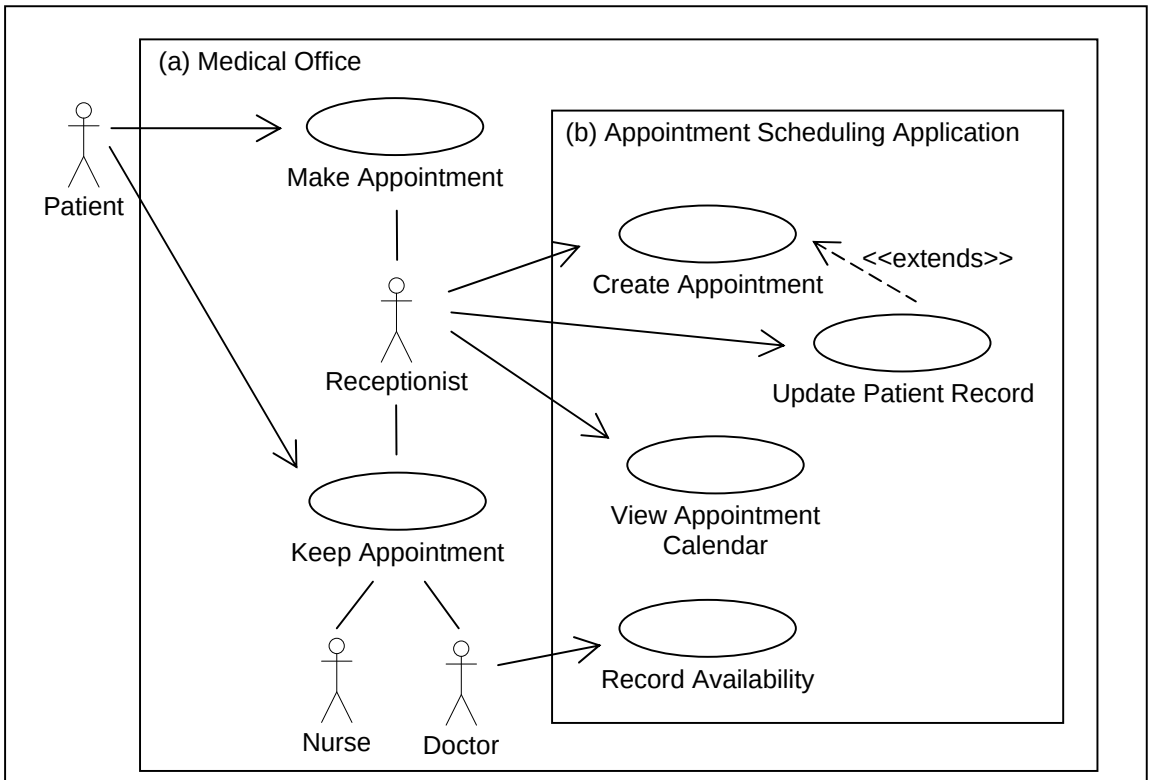


Figure A3: Ontological Deficit

Note: Example from Irwin and Turk, 2005, p. 9.

Figure A4 shows four conceptual modeling scripts that represent users' access permissions in a directory. Figures A4A-A4B have been created using the Object Modeling Technique (OMT) (Rumbaugh et al., 1991), while Figures A4C-A4D use a slightly amended syntax. All four scripts show users' access permissions via a grammatical construct that OMT refers to as a "link attribute." From the perspective of some ontological theories, link attributes often reflect the ontological construct of a "mutual property" (Burton-Jones and Weber, 1999). For example, the access permission in these figures can be viewed as a property associated with the interaction between the user and a file (Figures A4A, A4C) or between a user who employs a particular application to access a file (Figures A4B, A4D). The key point is that OMT contains *construct redundancy*, because it offers two different ways to show one phenomenon: mutual properties

connected to a line in binary associations (Figure A4A), and mutual properties connected to a diamond in ternary associations (Figure A4B).

Researchers might predict that some readers will become confused by the use of two symbols (a line and a diamond) to represent the same phenomenon. As a result, these readers may expend cognitive resources determining whether the two symbols have different meanings. Figures A4C-A4D show two ways to eliminate this redundancy: by always using a line (as in Figures A4A and A4D), or by always using a diamond (as in Figures A4B and A4C). Thus, researchers might predict that readers will expend fewer cognitive resources if link attributes are always shown using the same symbol. Once again, however, researchers might also predict that the outcome depends on the reader's level of expertise. The presence of redundant syntax may have little effect on readers who have extensive knowledge of OMT.

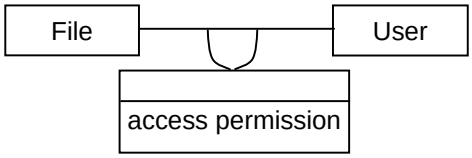
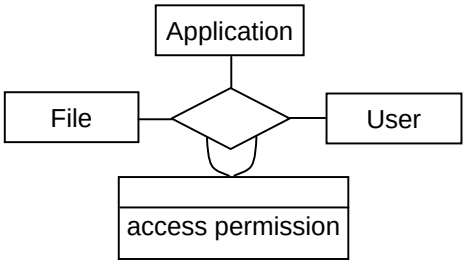
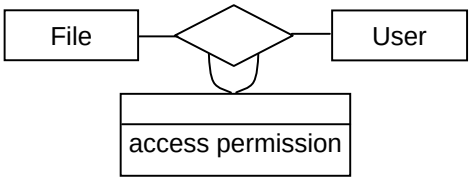
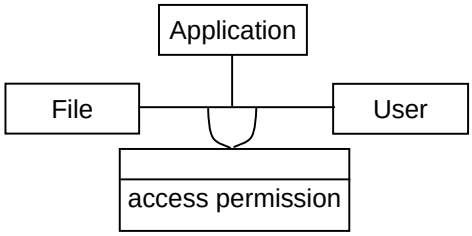
A4A. Link attribute for a binary association in OMT	A4B. Link attribute for a ternary association in OMT
	
A4C. Alternative link attribute for a binary association in OMT	A4D. Alternative link attribute for a ternary association in OMT
	

Figure A4: Ontological Redundancy

Note: Example from Rumbaugh et al., 1991, pp. 32-33.