Experimental practices for measuring the intuitive comprehensibility of modeling constructs: an example design

Sotirios Liaskos, Mehrnaz Zhian, and Ibrahim Jaouhar

School of Information Technology, York University,
4700 Keele St., Toronto, Canada, M3J 1P3
{liaskos,mzhian,jaouhar}@yorku.ca

Abstract. Conceptual model comprehensibility has attracted the interest of many experimental researchers over the past decades. Several studies have employed a variety of definitions and operationalizations of the comprehensibility construct as well as procedures for measuring it on a variety of model types. Intuitive comprehensibility is a specialization of the construct, referring to model or language comprehensibility exhibited by partially trained users. We present an experimental design for measuring the intuitive comprehensibility of a proposed extension to a goal modeling language as a means for reviewing experimental practices we have followed for similar studies in the past. Through such review, we hope to demonstrate the possibility of experimental design and technique reusability and its role as a motivating factor for more experimentation within the conceptual modeling research community.

Keywords: Model Comprehensibility/Understandability · Empirical Conceptual Modelling · Goal Models

1 Introduction

Experimentally evaluating the quality of conceptual models and conceptual modeling languages has enjoyed substantial attention from researchers over the past decades. Various studies have explored how users interact with diagrammatic representations and how they perceive modeling constructs represented in such ways. Often, the subject of investigation is the *comprehensibility* of models, and various interpretations of the meaning of this construct have been utilized both in theory and in empirical measurement [11].

A specialization of comprehensibility has been put forth that is concerned with the level of understanding of information appearing in a diagrammatically presented conceptual model by viewers with limited training in the corresponding conceptual modeling language. The working term *intuitiveness* has been proposed for this construct and a number of studies have been performed by our group for assessing it in requirements goal models [1,14,17,18] and independently elsewhere in process and other diagrams [3,12,22]. Our experiments

focussed on the intuitiveness of a specific language construct, namely contribution links within goal models, and the role thereof in making decisions within such models. Through these experiments we adopted and/or developed a set of methodological practices that we found served the purpose of studying the particular construct and may be applicable to a larger class of studies.

In this paper, we describe these practices and discuss their strengths and weaknesses, via presentation of an experimental design for a future study on the intuitiveness of temporal precedence constructs within goal models. We elaborate on the intuitiveness construct (Section 2), offer an introduction to our example study (Section 3) and describe our proposed design as an opportunity to also reflect on our experimental practices (Section 4). We conclude in Section 5.

2 Comprehensibility and Intuitiveness

Several efforts to empirically study comprehensibility of conceptual models have emerged in the literature, albeit with no clear consensus of what exactly the construct means and how it is to be measured, as reported by Houy et al. [11]. A possible starting point for understanding the construct may be found in SEQUAL, a semiotic framework for organizing conceptual model qualities [13]. There, the concept of *(manual) model activation* is put forth to describe the ability of models to guide the actions of human actors. Comprehensibility of a model is found within the category of *pragmatic quality* of a conceptual model, measured by the appropriateness of the model's activation. In other words, by being exposed to the model and its information, users (i.e., readers) of the model act (perform inferences, respond to questions, organize their work, make decisions etc.) in ways that satisfy the model, according to the designers of the latter. For example, a business process model is comprehensible by process actors, if, once they read it, said actors, organize their work, communicate with co-workers, answer process questions, troubleshoot etc. in ways that are compliant with the model – according to the model developers.

Model comprehensibility is distinct from *comprehensibility appropriateness of language* [9,24] which refers to the ability of the language to be the basis for the building of comprehensible models. From an empirical standpoint, this would, in principle, be measured by means of evaluating the comprehensibility of samples of several models developed in accordance to a language, controlling for factors that may affect model comprehensibility independent of the language, such as representation medium appropriateness, visual/physical quality [21] or, otherwise, language use. For such controlling and sample identification to be tractable, evaluation may take place at the individual construct level (e.g. individual elements, visualizations and relationship types) and/or a specific language feature or structural pattern (e.g. use of models for a specific task).

Intuitive comprehensibility appropriateness of a language, or part thereof or *language/construct intuitive comprehensibility* or, simply here, *(language/ construct) intuitiveness*, refers to comprehensibility appropriateness exhibited by users who have had limited previous exposure to the modeling language. The

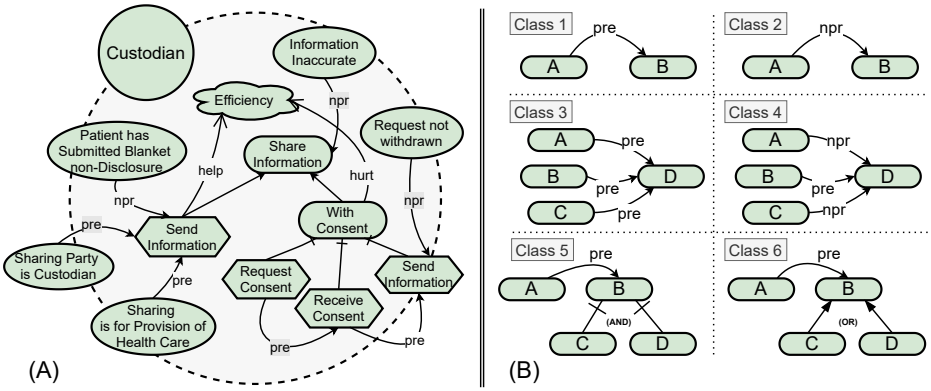

Fig. 1. (A) Goal model with preconditions — (B) Experimental Models

addition of the intuitiveness requirement is motivated by the need to make languages usable by users who would not otherwise dedicate effort to receive training in the language at hand. All else being equal, it is preferable that a language can be effectively used – i.e., allow models that lead to compliant activation – with less required training. Note that the definition of intuitiveness is here distinct from the concept of intuition (versus, e.g., reflection) studied in dual-process cognitive psychology [6,10], in that the former is agnostic to the exact cognitive process employed to interpret and make use of the constructs.

Our work – like much other literature, e.g., [8] – has been focussing on the intuitiveness of the diagrammatic representation choices of the language, i.e., whether shapes and symbols that appear on the diagram allow users to instantly know how to make correct use with the model. In Moody's terms [21] this is *semantic transparency* of the visual constructs, i.e. the ability of notational elements to communicate their meaning. While intuitiveness can be studied at the concept level alone [15] when expressed in natural language and, thus, free from the interference of visualization choices, our discussion here concerns the evaluation of the combination of the concept and its visualization.

3 Example: Preconditions in Diagrammatic Goal Models

To see how a study of intuitiveness appears in the process of language design we consider an example from the goal modeling domain. Goal models have been extensively studied with regards to their ability to represent intentional structures of stakeholders [2,25]. In the latest goal modeling standard, iStar 2.0 [5], elements such as *actors*, their *goals* and ways by which the latter can be decomposed into other goals or *tasks*, through *AND-refinements* and *OR-refinements* are presented. A diagrammatic notation faithful to the tradition of the original *i** language is used to visualize the concepts.

Such goal diagram appears in Figure 1(A), representing the ways by which, according to a jurisdiction, a custodian of health information is allowed to share such information with another agent. The legal requirement is that the custodian

can share health information without the patient's consent only as long as the third party is another custodian, the sharing is for the provision of health care and that the patient has not submitted a blanket non-disclosure statement. If any of these conditions are not met, consent must be acquired prior to sharing.

To model constraints such as the above, suppose that we want to extend the iStar 2.0 language to allow for *precondition* (resp. *negative precondition*) links $\{B \xrightarrow{pre} A\}$ (resp. $\{B \xrightarrow{npr} A\}$). Intuitively, such link shows that a goal/task A cannot be pursued/performed *unless* (resp. *if*) some condition B is met, including that some other goal/task has been achieved/performed or that some state of the world is believed to be true – for representing the latter *beliefs* are added in the diagram. A rigorous semantics of such or similar links is possible [16]. In setting up such semantics, however, designers often have flexibility. For example, what is the rule for combining multiple $\xrightarrow{pre}$ and $\xrightarrow{npr}$ arriving at an element such as *Send Information*; is it a conjunction, a disjunction, or something else? Most designers would probably opt for conjunction, but what if users of the diagram insist to act as if it were a disjunction? Likewise, what does it mean for users that a goal is "pursued"? Given $\{$*Information Inaccurate* $\xrightarrow{npr}$ *Share Information*$\}$ can I allow interpretations in which some *but not all* of the subtasks of goal *Share Information* are performed if *Information Inaccurate* holds, pretending that, e.g., performance of the tasks is for the pursuit of other unmentioned goals? Most designers would probably say no, but what if users act as if that was the correct interpretation? Disagreements between designer intent and user interpretation may imply that either the language features (allowing multiple incoming preconditions and allowing a precondition to a decomposed goal) or their visual representation deserve some reexamination.

In what follows, we use the problem of evaluating this hypothetical language extension to review our past experimental practices and experiences we acquired by applying such practices in similar problems.

4 Experimental Strategy

Our experimental approach consists of: (a) developing models that exemplify the construct or feature that we want evaluated and differ based on factors of interest, (b) identifying a participant sample that can be seen as representative of a user population, (c) partially training participants, (d) exposing participants to the models, observing inferences they perform therewith and comparing them with the ones language designers consider correct. We address these for our example problem.

4.1 Model Sampling

Model Format. When evaluation targets a specific construct or feature, the sampled models are constructed to exactly exemplify use of the construct or feature and abstract away other extraneous elements that may interfere with the measurement. In most of our earlier work [1,14,18], for example, we studied the

intuitiveness of various representations of contribution links in goal models for the purpose of identifying optimal decisions in such models. Given such narrow focus, sample models were structurally constrained: one OR-decomposition and a soft-goal-only sub-graph connected with contributions in a restricted way. No elements of the language that were extraneous to the research question were included – e.g. actors or AND-decompositions. Elsewhere [17], our samples were even simpler, containing two goals and one contribution link.

The advantage of such focused manufacturing of experimental units are (a) better experimental control and definition of factors (see below), (b) reduced need for training about unrelated modeling constructs. The disadvantages are: (a) the generalizability argument relies on showing that the manufactured models capture the "essence" of the language construct and its use under evaluation, (b) influential (but unknown) factors that exist in real-world models are absent.

An additional variable is how much context should be added to the example models. This can come in the form of: (a) real element descriptions in place of symbolic variables, (b) scenarios that create an even more elaborate context. In our example, we can use uninterpreted literals as in $\{B \xrightarrow{pre} A\}$, refer to specific goals as in Figure 1(A), or, do the latter and also add introductory material on the health information sharing case. While one can argue that such context information supports external validity by making the model samples more similar to the respective generalization class (real models), they have the potential of disturbing internal validity by switching the focus from the modeling construct to the content. For example, in $\{\textit{Request Consent} \xrightarrow{pre} \textit{Receive Consent}\}$, the precondition relationship is so obvious from the content, that measurement of the influence of $\xrightarrow{pre}$ in conveying such relationship is confounded.

Sample Models and Factors. When manufacturing sample models rather than sampling them in the wild, we have the benefit of introducing model-related factors of interest with more control. Such factors reflect properties, kinds or structural patterns of models, as per the research question. Our experience has shown that such factors are better treated in a *within-subjects* manner: the same participant is sequentially exposed to different classes of model structures, each such class being (part of) a level of the factor. In comparative studies, a *between-subjects* factor often emerges as well. In our past studies the comparison of various ways to represent a construct (e.g. in [14,17]) was arranged in such between-subjects fashion. The need for different training for each level of the factor in question is one of the main motivators of the between-subjects choice.

Example Design. In our example study, we would devise various examples of precedence links using abstract literals *A, B, C, ...* for the origin and destination goals, as seen in Figure 1(B). Several examples of each of the six presented classes can be considered, noticing that: (a) Classes 1 and 3 versus Classes 2 and 4 constitute the two levels of a $\xrightarrow{npr}$ presence/absence factor, (b) Classes 1 and 2 versus Classes 3 and 4 constitute the two levels of a "complexity" factor. These factors are crossed allowing the study of interactions. Classes 5 and 6 can further be compared with each other and with Class 1 as baseline; noting that larger samples will need to be acquired to allow for meaningful statistical analysis.

Thus, any or all of three within-subjects factors – negative precondition presence vs. absence, complex vs. simple and AND-decomposed vs. OR-decomposed vs. non-decomposed – can be studied. A between-subjects factor could be considered if we were to compare alternative ways to visualize $\xrightarrow{pre}$ and $\xrightarrow{npr}$, including adding comprehension aids, e.g. an AND arc to signify conjunction of $\xrightarrow{pre}$ links.

4.2 Training

Prior to being exposed to the models, participants are partially trained to the notation just enough so that the language's purpose and function is understood but the solutions to the experimental tasks do not directly follow from the training. We have extensively used short video presentations for such training. The benefits of video presentations over live lectures are manifold. Firstly, the exact training offered to participants is reviewable and reproducible. Secondly, in cases in which different language/construct versions need to be compared in a between-subjects manner, careful scripting and editing of the videos allows uniformity of training between groups. In our past experiments, videos have been fully recorded from script with only components that differ between groups appropriately video-edited. Thirdly, video presentations allow remote participation and consideration of on-line participant pools (more below).

As in any training, the threat in preparing video presentation remains that researcher bias can affect participant training in a way that skews the results towards one or the other direction. A possible way to address this is third-party evaluations or even development of training material. Despite such measures being practically difficult, video instead of in-classroom training removes many obstacles for such validation efforts.

Example Design. A video presentation can be developed to explain relevant goal modeling elements (goals, tasks, decomposition links) and the informal meaning of the $\xrightarrow{pre}$ and $\xrightarrow{npr}$ constructs, but would generally not describe specific uses of the construct for which we want to measure intuitiveness. For example, the video would not discuss how multiple $\xrightarrow{pre}$ links targeting the goal should be interpreted or elaborate on how pursuit of a goal is defined. When comparison with a baseline is desired such details can however be given in a separate control group or, less practically, in a within subjects pre-post manner [23].

4.3 Tasks

Experimental tasks are geared towards triggering and measuring model activation i.e. prompting, observing and recording *inferences* participants make with the displayed models. Parts of the theory of such inferences may need to be explained during training. In our study on assessing the intuitive comprehension of satisfaction propagation rules [17], the notions of partial and full goal satisfaction and denial had to be described in the videos. For the tasks, participants pick an inference that they think valid based on the model and their training. In our decision assessment studies [1,14,18], the task was a choice of goal alternative,

while in our propagation rule study [17] it was the specification or choice of the satisfaction level of a recipient of a contribution link.

Example Design. In our example, the notion of a situation (i.e. a state in which goals have been achieved, tasks have been performed, or beliefs are held) satisfying or not the model, needs to be part of the video training. Then, each model is accompanied by descriptions of situations and participants are asked if the model satisfies the situations. For example, a model of Class 1 (Figure 1(B)) can include the question whether $\{A, \neg B\}$ and $\{\neg A, \neg B\}$ are situations satisfying the link – which test whether participants perceive precondition as also a trigger condition, or whether they think that the presence of a link alone necessitates some satisfaction, as we actually observed with contribution links [17]. Notice how factors of interest can also be thus identified at the task level.

4.4 Operationalizations of Language Intuitiveness

The operationalization of the intuitive comprehension construct follows its theoretical definition (Section 2). The main measure we have used in the past is that of the level of agreement between participant responses to the experimental tasks (the model activation) and the normative answers to the questions (the language designer expectations), which agreement we refer to as *accuracy*. The accurate responses are then tallied up into an accuracy score used for the analysis. Calculation of *inter-respondent agreement* is also possible in the absence of a normative response. However, with such measures being aggregates of all participant responses, statistical inference possibilities are limited.

In some of our experiments we also asked the participants to rate their *confidence* to their response, using a Likert-type scale. Confidence can also be offered for the overall task, e.g., through one question in the end [18], which saves from execution time and perhaps allows for a more thoughtful response, but prevents analysis over the within-subjects factors. Whenever applicable, *response time* can also be relevant to understanding intuitive comprehension. However, both response time and confidence *alone* are not indicators of intuitive comprehension, in that participants may quickly and confidently provide inaccurate answers in the tasks. Nevertheless, following Jošt et al. [12], the ratio of accuracy over response time can also be an effective utilization of response time data.

Finally, in some of our experiments we invited participants to type-up a description of the method they used to make inferences and provide a response, as a proxy for a debriefing session. We have found that while some participants' textual descriptions can be usefully coded, they are often difficult to read and comprehend in any useful way. Note that both a debriefing sessions and response time measurements usually necessitate in-person administration.

Example Design. In the example experiment, we could measure accuracy, individual response confidence and, when possible, response time. Soliciting textual descriptions of how participants worked would not be a priority.

4.5 Participant Sampling

The appropriateness of using students as experimental participants is still debated in software engineering [7], where tasks are often specialized and require some technical ability. We believe that in conceptual modeling, user populations are wider and more diverse. Goal models, for instance, are to be used by any person whose intentions and decisions matter, and such persons can be of arbitrary backgrounds and abilities. Furthermore, intuitive comprehension of and distinction between concepts such as intentions, processes, events etc., is something that most senior college/University are expected to be able to perform. Following the same argument, in several instances we have also utilized on-line work platforms and particularly Mechanical Turk (MT). Such platforms have been found to be remarkably reliable for psychological experiments [4]. Assuming a commitment that the prospective users of the language under investigation is not limited to e.g. IT or management backgrounds, for certain simple tasks in conceptual modeling (e.g. discriminating among common concepts, associating notational symbols with concepts, making a decision on a daily-life problem via models) the MT samples appear to be suitable. Future correlation studies similar to the one performed by Crump et al. [4] would shed more light on the strength of this assumption.

Example Design. For our example design a mixture of University students and Mechanical Turk workers can be invited to participate.

4.6 Analysis

A likely approach for analysing data coming from designs such as the above is analysis of variance (ANOVA) [20]. In our most complex past cases such analysis included one between-subjects factor (e.g. a comparison of three visualizations) and one or two repeated-measures factors (e.g. model complexity and type). One problem we have faced with accuracy measures specifically is that, being integer values in the interval $[1..N]$, N the number of participant responses, they often violate normality assumptions – especially for small N, necessitating robust and/or non-parametric testing.

Further, looking at effect sizes is meaningful in our context. We have found that looking at a simple difference between means offers an intuitive picture. For example, that one group scores 1.5 (vs., e.g., 5) out of 20 accuracy points more compared to another is very informative vis-a-vis the practical importance of the effect, irrespective of statistical significance.

Finally, the generalization class needs to be carefully considered when performing inferences. With instruments such as the ones we described, the simplest generalization statement concerns the performance that participants in the entire population would demonstrate. An analytical step, however, needs to be taken to extend this generalization to the population of models, given that the sample models are manufactured specifically to find an effect rather than randomly sampled. Similar non-empirical arguments apply to generalizing to different kinds of activities with the models or in different contexts.

Example Design. In our example, parametric methods for repeated measures would be applied [20], accompanied with an equivalent robust test, if suspicion of violation of assumptions presents itself. Presence of a between-subjects factor entails a split-plot ("mixed") ANOVA whose likely deviation from assumptions, however, requires resorting to a complex range of countermeasures from transformations and bootstrapping to robust tests [19].

5 Concluding Remarks

We presented an experimental design for measuring the intuitiveness of a proposed extension to a conceptual modeling language, as a means to also review experimental practices we have been following to answer similar research questions in the past. Researchers in conceptual modeling often appear to dread the time, effort and risk associated with performing such studies. Our long term goal is to help develop standardized practices, patterns, techniques and tools that allow systematic, quick and efficient design and conduct of comprehensibility studies for use by researchers who otherwise could not afford the effort. Our focus so far has been decisively narrow within the constellation of phenomena that surround the development and use of conceptual models and their languages. However, it has promisingly allowed us to develop re-usable patterns and ideas within the sampling, measuring, training and analysis aspects that have substantially reduced required effort to set-up, run and analyse an experiment. Community-wide sharing, acceptance and continuous improvement of such cost-effective experimental practices, may make conceptual modeling researchers more eager to incorporate empirical investigation in their research.

References

1. Alothman, N., Zhian, M., Liaskos, S.: User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment. In: Proc. of the 36th Int. Conf. on Conceptual Modeling (ER'17). pp. 451–465. Valencia, Spain (2017)
2. Amyot, D., Mussbacher, G.: User Requirements Notation: The First Ten Years, The Next Ten Years (Invited Paper). Journal of Software **6**(5), 747–768 (2011)
3. Bork, D., Schrüffer, C., Karagiannis, D.: Intuitive Understanding of Domain-Specific Modeling Languages: Proposition and Application of an Evaluation Technique. In: Proceedings of the 38th International Conference on Conceptual Modeling (ER 2019). pp. 525–539. Salvador, Brazil (2019)
4. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's Mech. Turk as a Tool for Experimental Behavioral Research. PLoS ONE **8**(3), 1–18 (2013)
5. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. The Computing Research Repository (CoRR) **abs/1605.0** (2016)
6. Evans, J.S.B.T.: Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. Annual Review of Psychology **59**(1), 255–278 (2008)
7. Falessi, D., Juristo, N., Wohlin, C., Turhan, B., Münch, J., Jedlitschka, A., Oivo, M.: Empirical software engineering experts on the use of students and professionals in experiments. Empirical Software Engineering **23**(1), 452–489 (2018)

8. Gonçalves, E., Almendra, C., Goulão, M., Araújo, J., Castro, J.: Using empirical studies to mitigate symbol overload in iStar extensions. Software and Systems Modeling **19**(3), 763–784 (2020)
9. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. Ph.D. thesis, University of Twente (2005)
10. Hadar, I.: When intuition and logic clash: The case of the object-oriented paradigm. Science of Computer Programming **78**(9), 1407–1426 (2013)
11. Houy, C., Fettke, P., Loos, P.: Understanding understandability of conceptual models - What are we actually talking about? In: Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012). pp. 64–77. Florence, Italy (2012)
12. Jošt, G., Huber, J., Heričko, M., Polančič, G.: An empirical investigation of intuitive understandability of process diagrams. Computer Standards and Interfaces **48**, 90–111 (2016)
13. Krogstie, J., Sindre, G., Jørgensen, H.: Process models representing knowledge for action: a revised quality framework. European Journal of Information Systems **15**(1), 91–102 (2006)
14. Liaskos, S., Dundjerovic, T., Gabriel, G.: Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment. In: Proc. of the 33rd ACM Symp. on Applied Computing (SAC'18). pp. 1272–1281. Pau, France (2018)
15. Liaskos, S., Jaouhar, I.: Towards a framework for empirical measurement of conceptualization qualities. In: Proceedings of the 39th International Conference on Conceptual Modeling. Vienna, Austria (2020)
16. Liaskos, S., Khan, S.M., Soutchanski, M., Mylopoulos, J.: Modeling and Reasoning with Decision-Theoretic Goals. In: Proceedings of the 32th International Conference on Conceptual Modeling, (ER 2013). pp. 19–32. Hong-Kong, China (2013)
17. Liaskos, S., Ronse, A., Zhian, M.: Assessing the Intuitiveness of Qualitative Contribution Relationships in Goal Models: an Exploratory Experiment. In: Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'17). pp. 466–471. Toronto, Ontario (2017)
18. Liaskos, S., Tambosi, W.: Factors Affecting Comprehension of Contribution Links in Goal Models: An Experiment. In: Proceedings of the 38th International Conference on Conceptual Modeling (ER 2019). pp. 525–539. Salvador, Brazil (2019)
19. Mair, P., Wilcox, R.: Robust statistical methods in R using the WRS2 package. Behavior Research Methods **52**(2), 464–488 (2020)
20. Maxwell, S.E., Delaney, H.D.: Designing Experiments and Analyzing Data. Taylor and Francis Group, LLC, New York, MA, USA, 2 edn. (2004)
21. Moody, D.L.: The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Transactions on Software Engineering **35**(6), 756–779 (2009)
22. Roelens, B., Bork, D.: An Evaluation of the Intuitiveness of the PGA Modeling Language Notation. In: Proceedings of the 25th International Conference on Evaluation and Modeling Methods for Systems Analysis and Development (EMMSAD 2020). pp. 395–410. Grenoble, France (2020)
23. Rosnow, R.L., Rosenthal, R.: Beginning Behavioral Research: A Conceptual Primer. Pearson Prentice Hall, NJ, USA, 6 edn. (2008)
24. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal **3**(4), 217–237 (1993)
25. Yu, E.S.K.: Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering. In: Proceedings of the 3rd IEEE International Symposium on Requirements Engineering (RE'97). pp. 226–235. Annapolis, MD (1997)