Towards a Framework for Empirical Measurement of Conceptualization Qualities

Sotirios Liaskos(✉) and Ibrahim Jaouhar

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

Abstract. Conceptualization development is central in modeling language design. As one of their first design steps, language designers need to decide on a set of concepts on which the language will be based and which can be understood and used by a population of modelers for characterizing and representing relevant domain information. Thus, exposing candidate concept sets to future users may offer insights on how well the concepts of choice are understood and distinguished from each other by those who will be called to actually use the language. We propose an empirical measurement framework to allow just that. The framework consists of an instrumentation approach whereby participants sampled from the user population classify domain expressions to the corresponding concepts, and a set of measurement constructs for translating participant observed data into design insights. A small case study is conducted to explore the feasibility and limitations of the proposed approach.

Keywords: Conceptual modeling · Conceptualization quality · Experimental study · Goal models

1 Introduction

Developing conceptualizations lies at the heart of conceptual modeling language design. Such conceptualizations are sets of concepts and their definitions that the language designers think are useful for modeling a domain [12,26]. Once the main conceptualization is decided, designers can proceed with the definition of syntax, notation, modeling and reasoning procedures and mechanisms and other components needed to develop a fully-fledged modeling language. However, deciding why a particular candidate set of concepts is better than a competing one, all else being equal, seems to remain an art rather than a science. Designers seem to have little to rely on for knowing if and how end users would understand candidate conceptualizations that emerge in the design process.

We propose an empirical measurement framework to be used for assisting the evaluation of qualities of conceptualizations in the context of a language design

G. Dobbie et al. (Eds.): ER 2020, LNCS 12400, pp. 512–522, 2020.
https://doi.org/10.1007/978-3-030-62522-1_38

effort. The framework consists of a set of empirical constructs, operationalized on the basis of exposing untrained experimental participants to sets of concepts and then asking participants to classify expressions of the domain under the concept that best describes each. The resulting metrics reveal levels and patterns by which participants agree on how expressions should be classified, both within themselves and with the language designers. Subsequent descriptive analyses offer designers insight of how their conceptualization proposals are understood by prospective modelers. In an empirical study, we investigate whether consistencies exist among the metrics and between the metrics and our intuition.

The rest of the paper is organized as follows. In Sect. 2 we present our motivation and the experimental constructs we propose. In Sect. 3 we describe the empirical study we performed and in Sects. 4 and 5 we discuss our findings, limitations, future and related work.

2 Conceptualizations and Their Quality

2.1 Conceptual Modeling Languages and Conceptualizations

Conceptual modeling languages are based on the definition of a core set of concepts that modelers are to use in order to develop models according to the language. Such sets are called *conceptualizations* (henceforth also: *concept sets*), that is, *"concepts used to articulate abstractions of state[s] of affairs in a given domain"* [12]. Based on this foundation of concepts, a modeling language complete with syntax, notation, modeling procedures and mechanisms can be developed [4,17]. To facilitate the discussion that follows, we distinguish between the *concern domain* (henceforth *domain*), which describes the aspects of reality that we wish to focus on in our modeling (e.g. intention/motivation, process, structure, function etc.) and *application domain*, the actual problem that we wish to model (e.g. a travel agency or a flight booking system).

It is easy to observe that for the same or similar concern domain, different concept sets can emerge. Taking the intention domain for example, several *goal modeling* languages have been introduced: KAOS [10], *i** [33], URN/GRL [2,34], Tropos [30], iStar 2.0 [9] and their variants, as well as Archimate and its a motivation aspect [31]. These languages have similar but not the exact same concept sets. For instance, KAOS models intention using a set that includes *"agents"*, *"goals"*, *"constraints"* and *"actions"* [10], while iStar 2.0 includes *"actors"*, *"goals"*, *"qualities"* and *"tasks"* (the latter referred to as *"plans"* in some Tropos conceptualizations [30]) and Archimate's motivation aspect has *"goals"*, *"outcomes"*, *"drivers"* and *"requirements"* [31].

2.2 Conceptualization Quality

Given two candidate concept sets for a domain, examples can be devised in which it is obvious for some observer that one conceptualization is a better fit for the domain than the other. For example, the above mentioned iStar 2.0

concepts seem to be more suitable for modeling intention than the concepts *"account"*, *"credit"* and *"debit"*, which are probably better suited for, say, modeling economic transactions. However, more rigorous and systematic measures of fit would be useful when the candidate conceptualizations are not as semantically distant, and therefore which one should be preferred is not as "obvious" or a matter of strong agreement among designers.

Our proposed framework focusses on *empirical* measures of fit between a chosen conceptualization and its domain, i.e. measures coming from observing behaviors and attitudes of potential users of the conceptualizations (or proxies of such users). To devise such goodness-of-fit measures, and a theory thereof, we draw inspiration from well established methods from the area of qualitative content analysis [18]. At the heart of content analysis lies the effort to classify defined units of qualitative content (e.g. text, audiovisual) into a system of codes, a "data language". The latter contains data *variables*, each offering a set of semantically mutually exclusive *values* (also: *codes*), each of which is best suited for classifying specific *units* of content.

Our reference to the content analysis tradition is based on an analogy between conceptualizations and such variables. Given a unit of phenomena in the world – known indeed through consumption of content (interview responses, policy documents etc.), modelers choose one of the concepts of the conceptualization to model the unit as such. For example, an iStar 2.0 modeler confronted with the unit "Travel Office" has to choose a concept from the iStar 2.0 concept set we saw above to incorporate the unit within their model. In the iStar 2.0 diagram, the unit will most likely emerge in form of a circular visual element, signifying that the modeler has decided to model it (*code* it) as an *"actor"*, simultaneously excluding the possibility of modeling it as something else (e.g. *"task"*).

Continuing our analogy with content analysis, a variable and its set of codes can be seen as a measurement instrument that detects the presence or absence of specific kinds of meaning within content. For such an instrument to be useful it needs to be *reliable*, i.e. to result in the same coding outcome independent of the coding event and involved person(s) [18]. A similar expectation largely holds in modeling. We would not like "Travel Office" to be modelled as either an *"actor"*, a *"task"* or a *"quality"* in equal frequency depending on who does the modeling and when, the domain information being otherwise the same.

We propose three conditions for a conceptualization to be reliable – which by no means exhaust all such conditions. Firstly, the conceptualization allows for reproducible modeling: a group of different modelers in different times, when exposed to the exact same information about the application domain, they will choose the same concepts to model the same units of content. Thus, "Travel Office" is always modelled using the same concept independent on whom one asks and when. We will henceforth refer to this reliability construct as *agreement*, which can be intra-rater (compare answers of the same person at different times) and inter-rater (compare answers of different persons). Secondly, the way modelers classify domain information into concepts, should agree with standards set by conceptualization designers. We call this, *accuracy*. Thus, when the designers

expect that "Travel Office" should be modelled as *"actor"* within a given application domain description and modelers actually do so in practice, this supports the belief that *"actor"* is a concept that will likely be used as intended. Finally, a conceptualization would tend to be more reliable if its constituent concepts have minimum or no *semantic overlap*, in a way that each concept partitions states of affairs within an application domain instance into classes with no or minimal intersection. For instance, all instances of *"goals"* and all instances of *"actors"* in the application domain are conceptualized as such, respectively, without a large class of instances being equally able to be classified either way based on the same application domain information.

These reliability features need not be seen as a pre-requisite for conceptual modeling language usefulness. However, they can be useful for supporting language design in terms of assessing how the concept set will be understood by the application community.

2.3 Metrics

Operationalizations of the qualities described above is based on observing how a group of human participants S uses a provided concept set O to model a description of a state of affairs. The group of participants is sampled from a population of potential users (modelers) of the language which will use O as its basis, or a proxy when a suitable sample is unattainable. The participants are trained to O using definitions and authoritative examples, such as those that accompany language guides and tutorials. Then the participants are offered a set L of *expressions* of states of affairs within some application domain and are asked to *classify* each to one of the concepts in O. Reliability qualities can then be explored through various aggregations and visualizations of such observational data, on the basis of the constructs discussed above: agreement, accuracy and overlap. We turn our focus to operationalizations of each of these.

Agreement is based on the measuring of the degree to which participants in S classify each item in L using the same concept from O. Given an item $l \in L$ the *agreement per expression* $(\boldsymbol{GpE(l)})$ is any measure of concentration of classifications of l to specific subset of concepts by the participants. From the several available options, we here adopt the Herfindahl-Hirschman index – used in Economics to measure market concentration – normalized to [0,1] [6]. Specifically, let $f(l, o_i)$ be the proportion of classifications in which l is classified by $s \in S$ as $o_i \in O$. The GpE for l is then:

$$GpE(l) = \frac{\sum_{o_i \in O} f(l, o_i)^2 - 1/|O|}{1 - 1/|O|}$$

The closer the index is to 1, the more the concentration of responses to a specific concept, hence the more the agreement on the classification of l. GpE can be used as a building block for aggregated agreement measures such as the *total agreement* $(\boldsymbol{GT})$ which is the average GpEs of all expressions in L. Note that although the above are for inter-rater analysis, analogous constructs can be envisioned for intra-rater agreement.

Accuracy measures are based on calculating the degree to which participants in S classify an item l in L in a way that agrees with how the designers would classify l. Analogously to agreement, *accuracy per expression* $(\boldsymbol{ApE(l)})$ is the proportion of classifications that are in agreement with the authoritative one for item l, normalized from the interval $[q, 1]$ to $[0, 1]$, where $q = 1/|O|$ is the proportion expected by random. Then, *accuracy per concept* $(\boldsymbol{ApC(o)})$ is aggregation of individual ApEs by the authoritative concept to which l is classified by the designers, $ApC(o) = mean_{l \in L_o}(ApE(l))$, where $L_o \subseteq L$ is the set of items that designers think should be classified as o. Finally *accuracy per participant* $(\boldsymbol{ApP(s)})$ measures each participant's $s \in S$ proportion of classifications that agree with the authoritative classification.

Overlap is based on measuring the degree to which participants in S classify each item in L on the same pair of concepts from O. One way to define *observed overlap per expression* $(\boldsymbol{VpE(l, o_1, o_2)})$ for two concepts o_1 and o_2 is:

$$VpE(l, o_1, o_2) = \frac{min\{f(l, o_1), f(l, o_2)\}}{max\{f(l, o_1), f(l, o_2)\}}[f(l, o_1) + f(l, o_2)]$$

where, again, $f(l, o)$ is the proportion of classifications in which l is classified as o. The observed overlap per pair is the average per expression for a specific pair $\boldsymbol{VpI(o_1, o_2)} = mean_{l \in L}(VpE(l, o_1, o_2))$ – noting that the average can be weighted per authoritative classification through the *overlap per concept* $\boldsymbol{VpC(o)} = mean_{o' \in O}(VpP(o, o'))$ metric.

3 Case Study

We now turn to an empirical study we performed to acquire initial feasibility evidence for some of the measures. Our goal is to examine whether acquisition of the measures is possible and whether they are consistent with each other and with intuitions we have about the qualities of the languages we put to test.

Two conceptualizations are studied: one constructed as a subset/derivative of iStar 2.0 and one from a made-up language we call *"intention models"*. The goal modeling language conceptualization contains the concepts {*goal*, *quality*, *task*, *belief*} the former three concepts adopted directly from the iStar 2.0 and *belief* added from GRL [34]. The intention modeling language conceptualization contains the concepts {*goal*, *objective*, *claim*, *assertion*}. The concepts are chosen in a way that the first two and the last two appear to be synonyms, so referring to the same kinds of phenomena in the application domain.

A number of sets of expressions are also prepared for each language: one featuring only a list of such without any additional context, one based on the main example from the iStar 2.0 guide contextualized within a description of a fictional character with goals, tasks, qualities copied as-is from the guide [9] and beliefs constructed from scratch, and a third constructed in the same way from Archimate's authoritative examples on motivation structures; the expressions are transferred as-is from concept instances in exemplar models [31]. For

the intention models, an additional set of expressions and context description concerning a hypothetical grocery store owner are constructed from scratch.

The experimental units are initially placed in two separate instruments, one for goal models and one for the intention models. Each instrument starts with an instructional video presenting the concepts through the authoritative definitions and authoritative examples. Then, each expression set is presented in a separate screen, with its context description, wherever applicable, and participants are asked to classify each expression to one of the four concepts of the corresponding concept set. A total of 41 participants from the Mechanical Turk pool [7], 13 female and 28 male, ages 23 to 69 (median 40), majority (34) in Science, Technology and Engineering are recruited.

Results. For a first glimpse of how the two languages compare, we use a heatmap style visualization we call *concept overlap maps* to visually explore overlaps between concepts, as in Fig. 1. Starting from intention models, the categories within intentions and statements exhibit substantial overlap compared to other pairs, as strongly expected. Also in agreement with expectation, goal models show that *goals* overlap with *tasks* and, less so with *qualities*.

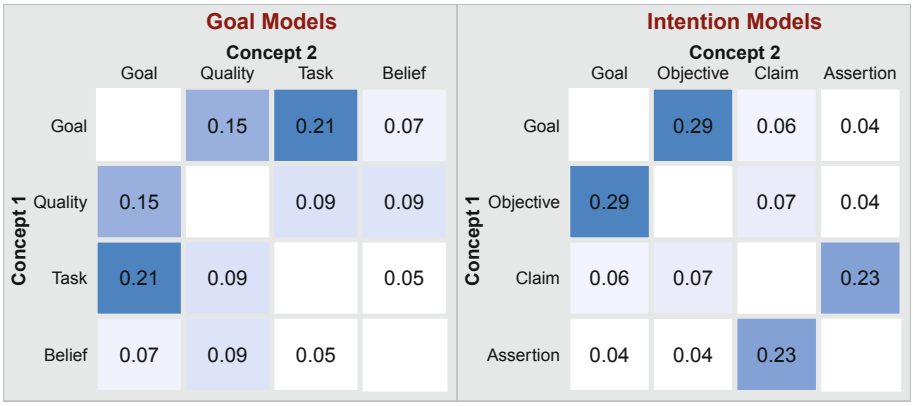

Fig. 1. Concept Overlap Maps

As a second exploratory step, we compare GpE with the p values of simple multinomial tests for each expression against the hypothesis that participants classify randomly, i.e. choose one of the four choices as if rolling a dice. Histograms of the results are seen in Fig. 2(A). For the analysis, intention models are considered in three modes: as introduced (*flat mode*, "Int. M. Flat"), with each pair of overlapping concepts (*goal-objective* and *claim-assertion*) merged to one (*between mode*, "Int. M. Between") and, conversely, focussing on the dominant overlapping pair in each expression and treating it as if it were a two-concept language (*within mode*, "Int. M. Within"). As expected, more frequent high levels of GpE are observed in the between mode of intention models. Goal

models offer similar distributions of GpE as flat mode intention models. Comparison with the binomial result, however, which offers an indication of overall randomness, shows that patterns of agreement may exist within seemingly low GpEs; from Fig. 1 we see that in goal models this is probably due to the distance of belief from the three other concepts.

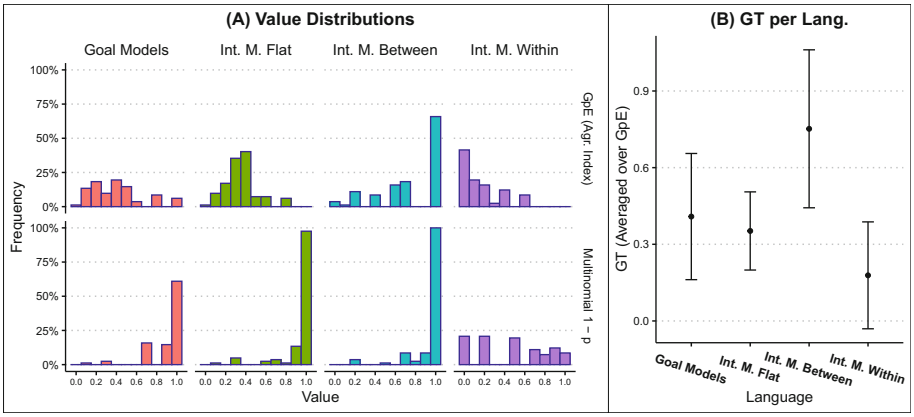

Fig. 2. Randomness and Agreement

Finally, accuracy measures can be meaningfully compared between goal models and intention models in between mode, as there is no authoritative response in the within and flat modes. The ApC for goals, qualities, tasks and beliefs is respectively 0.3, 0.63, 0.47 and 0.64. The two lowest levels are consistent with where overlaps occur as per Fig. 1. In intention models, expressions authoritatively designated as goals or objectives and claims or assertions exhibit ApCs of 0.78 and 0.93, respectively. This is in agreement with our expectation: the two pairs do not have as much of a conceptual overlap between them as *goals, tasks* and *qualities* do in goal models.

4 Validity Concerns and Future Work

We now discuss important validity threats and pitfalls that one must be mindful of when considering the proposed measurement approach. In terms of *external validity*, generalization of the findings is sensitive to the choice of expressions, the domain of origin thereof, and the participant sample. For an independent investigator, a first check can include expressions taken or derived from the authoritative examples most often provided by designers in language-defining publications, tutorials and guides. Such expressions can be assumed to be the best samples of: (a) expressions describing phenomena the designers destine their language to be used for, (b) associations between expressions and their authoritative classifications. The participant sample, on the other hand, is meant to

be taken from a modeler population, i.e., persons who could be using the language in practice. From an *internal validity* viewpoint one can further observe that both the collected expressions and the training procedure can interfere with conclusions with respect to the conceptualization qualities. For example, an otherwise well-chosen conceptualization may yield low agreement measures or strange overlaps due to bad training or badly written expressions. This can be mitigated by observing the behavior of the measures over repeated studies on the same conceptualization whereby training and expression choices and formats vary. An additional *construct validity* threat is whether and how the way a participant classifies an expression to a concept is biased by the way it is written. For example, in iStar 2.0, the examples in the language guide [9] train modelers that if, e.g., bill payment is a task it is written as *"Pay Bills"* but if it is a goal it is written as *"Have Bills Paid"*. Using such cues, participants may accurately classify expressions according to language style rather than the domain information, revealed e.g. in the description context. Avoiding such effect is on the investigator's hands and interests, who can choose to tailor both training and expressions to specific needs.

Finally, the agreement, accuracy and overlap measures themselves are subject for further study and refinement from a variety of angles. One is their ability to compare concept sets of different sizes. While the proposed normalizations allow for rough qualitative comparisons, a theory of such comparisons is yet be developed. It would be specifically relevant to know if decrease in conceptual granularity (cf. [14]) is always (as a law) accompanied with increase in accuracy and agreement, and, if yes, how we control for this increase for a fair comparison. A second concern is the identification of statistical properties of the measures so to allow inferences to populations, when random sampling has been assumed.

Thirdly, a connection of these measures with existing conceptualizations of language quality can be investigated. Relevant here are the analytical constructs of *lucidity*, *laconicity*, *soundness* and *completeness* [32] as used for ontological analysis of conceptualization quality [12]. The constructs presented here appear to be coarser and do not clearly indicate the specific pathology of the conceptualization in those terms. For example, low agreement – they way we defined it – may not be an exclusive symptom of *construct redundancy* as it can also be caused by, e.g., *incompleteness*. It appears, nevertheless, that refinements of our constructs are possible to allow for some commensurability if not direct operationalization relationships with the four quality constructs. Regardless, empirical investigation does not compete with the need for ontological analysis. Likewise, more work will be required to position such metrics within established language quality attributes [19,23,25]. For example, *modeler appropriateness* and *participant appropriateness* [19] refer to the correspondence between the language constructs and the way producers and users of models perceive reality. From an empirical standpoint, however, any measurement of comprehensibility or domain appropriateness (i.e., lucidity, laconicity etc. [12,32]) is likely based on modeler and/or participant samples, requiring care in clarifying the precise object of measurement and the relevant influencing factors.

5 Related Work and Conclusions

Empirically studying conceptual modeling languages is not a new enterprise, with many efforts having been dedicated on firstly understanding the basic empirical constructs of quality [19,23,25] and then engaging in experimental or other empirical activity. A plethora of studies have been conducted focussing on various understandability conceptions of conceptual models in general. Houy et al. offer a comprehensive survey aimed at organizing our understanding of understandability [16]. Much of the work has focussed on process and entity or other domain structure models, e.g. [8,24]. Goal models, our example focus here, have also been the subject of empirical investigation in various instances, e.g., Hadar et al. [13], Horkoff and Yu [15], Santos et al. [27], Estrada et al. [11] or Liaskos et al. [1,20–22]. A strong appeal to the consensus of user populations has been put forth by Caire et al. [5], which we also espouse as a principle. Naturally, this line of work is complemented by several analytical and ontology-based efforts to explore qualities of intention conceptualizations, e.g., Bernabè et al. [3].

Our work is inspired by a vision of measurement *standardization* for systematizing empirical evaluation, as is commonly done in other disciplines. By using standard, reproducible and comparable quality assessment instruments, language designers are better equipped in their effort to demonstrate the qualify of their designs and increase the appeal of such to practitioners.

References

1. Alothman, Norah., Zhian, Mehrnaz, Liaskos, Sotirios: User perception of numeric contribution semantics for goal models: an exploratory experiment. In: Mayr, Heinrich C., Guizzardi, Giancarlo, Ma, Hui, Pastor, Oscar (eds.) ER 2017. LNCS, vol. 10650, pp. 451–465. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69904-2_34
2. Amyot, D., Mussbacher, G.: User requirements notation: the first ten years, the next ten years (Invited Paper). J. Software **6**(5), 747–768 (2011)
3. Bernabé, César Henrique., Silva Souza, Vítor E., Almeida Falbo, Ricardo de., Guizzardi, Renata S.S., Silva, Carla: GORO 2.0: evolving an ontology for goal-oriented requirements engineering. In: Guizzardi, Giancarlo, Gailly, Frederik, Suzana Pitangueira Maciel, Rita (eds.) ER 2019. LNCS, vol. 11787, pp. 169–179. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34146-6_15
4. Bork, D., Karagiannis, D., Pittl, B.: How are metamodels specified in practice? empirical insights and recommendations. In: Proceedings of the 24th Americas Conference on Information Systems (AMCIS 2018). New Orleans, LA (2018)
5. Caire, P., Genon, N., Heymans, P., Moody, D.L.: Visual notation design 2.0: towards user comprehensible requirements engineering notations. In: Proceedings of the 21st IEEE International Requirements Engineering Conference (RE 2013), pp. 115–124 (2013)
6. Cracau, D., Lima, J.E.D.: On the normalized herfindahl-hirschman index: a technical note. Int. J. Food Syst. Dyn. **4**(7), 382–386 (2016)
7. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. PLoS ONE **8**(3), e57410 (2013)

8. Cruz-Lemus, J.A., Genero, M., Manso, M.E., Morasca, S., Piattini, M.: Assessing the understandability of UML statechart diagrams with composite states–a family of empirical studies. Empirical Software Eng. **14**(6), 685–719 (2009)
9. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. The Computing Research Repository (CoRR) (2016). http://arxiv.org/abs/1605.07767
10. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. Sci. Comput. Program. **20**, 3–50 (1993)
11. Estrada, H., Rebollar, A.M., Pastor, O., Mylopoulos, J.: An empirical evaluation of the i* framework in a model-based software generation environment. In: Proceedings of the 18th International Conference on Advanced Information Systems Engineering (CAiSE 2006), pp. 513–527. Luxembourg (2006)
12. Guizzardi, G.: Ontological foundations for structural conceptual models. Ph.D. thesis, University of Twente (2005)
13. Hadar, I., Reinhartz-Berger, I., Kuflik, T., Perini, A., Ricca, F., Susi, A.: Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: results from a family of experiments. Inf. Software Technol. **55**(10), 1823–1843 (2013)
14. Henderson-Sellers, B., Gonzalez-Perez, C.: Granularity in conceptual modelling: application to metamodels. In: Proceedings of the 29th International Conference on Conceptual Modeling (ER 2010), pp. 219–232. Vancouver, Canada (2010)
15. Horkoff, Jennifer, Yu, Eric: Finding solutions in goal models: an interactive backward reasoning approach. In: Parsons, Jeffrey, Saeki, Motoshi, Shoval, Peretz, Woo, Carson, Wand, Yair (eds.) ER 2010. LNCS, vol. 6412, pp. 59–75. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16373-9_5
16. 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)
17. Karagiannis, D., Khün, H.: Metamodelling platforms. In: Proceedings of the 3rd International Conference on E-commerce and Web Technology, pp. 182–197. France (2002)
18. Krippendorff, K.: Content Analysis: An Introduction to its Methodology. SAGE (2004)
19. Krogstie, J.: Model-Based Development and Evolution of Information Systems: A Quality Approach. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4471-2936-3
20. Liaskos, S., Dundjerovic, T., Gabriel, G.: Comparing alternative goal model visualizations for decision making: an exploratory experiment. In: Proceedings of the 33rd ACM Symposium on Applied Computing (SAC 2018). pp. 1272–1281. PAU, France (2018)
21. 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 2017), pp. 466–471. Toronto, Ontario (2017)
22. Liaskos, Sotirios, Tambosi, Wisal: Factors affecting comprehension of contribution links in goal models: an experiment. In: Laender, Alberto H.F., Pernici, Barbara, Lim, Ee-Peng, de Oliveira, José Palazzo M. (eds.) ER 2019. LNCS, vol. 11788, pp. 525–539. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33223-5_43
23. Lindland, O.I., Sindre, G., Solvberg, A.: Understanding quality in conceptual modeling. IEEE Software **11**(2), 42–49 (1994)

24. Mendling, J., Strembeck, M.: Influence factors of understanding business process models. In: Proceedings of the 11th International Conference on Business Information Systems, pp. 142–153. Innsbruck, Austria (2008)
25. Nelson, H.J., Poels, G., Genero, M., Piattini, M.: A conceptual modeling quality framework. Softw. Quality J. **20**, 201–228 (2012)
26. Olivé, A.: Conceptual Modeling of Information Systems. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72677-7
27. Santos, M., Gralha, C., Goulão, M., Araújo, J.: Increasing the semantic transparency of the KAOS goal model concrete syntax. In: Proceedings of the 37th International Conference on Conceptual Modeling (ER 2018), pp. 424–439. Xi'an, China (2018)
28. Stoet, G.: PsyToolkit: a software package for programming psychological experiments using Linux. Behav. Res. Methods **42**(4), 1096–1104 (2010)
29. Stoet, G.: PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. Teach. Psych. **44**(1), 24–31 (2017)
30. Susi, A., Perini, A., Mylopoulos, J.: The tropos metamodel and its use. Informatica **29**, 401–408 (2005)
31. The Open Group: ArchiMate® 3.1 Specification. Technical report (2019)
32. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Inf. Syst. J. **3**(4), 217–237 (1993)
33. 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 1997). pp. 226–235. Annapolis, MD (1997)
34. Yu, E.S.: GRL - Goal-oriented Requirement Language. http://www.cs.toronto.edu/km/GRL/