The final publication is available at Springer via
https://doi.org/10.1007/978-3-319-69904-2_34

User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment

Norah Alothman, Mehrnaz Zhian, and Sotirios Liaskos

York University, Toronto, ON, M3J 1P3, Canada,
norah@yorku.ca, mehrnaz@yorku.ca, liaskos@yorku.ca

Abstract. Goal models have long been regarded to be an effective way for representing stakeholder goals and how they relate to one another during requirements engineering. One of the ways goals are connected in goal models is contribution relationships, which represent how satisfaction of one goal affects the satisfaction of another. There are several proposals in the literature on how contributions should be modelled and used, but little empirical evidence as to which one is more intuitive for users. We experimentally explore how users interpret numeric contribution labels in goal models. Experimental participants are exposed to a number of pre-constructed goal models and are asked what they believe the satisfaction degree of a goal is given the satisfaction degree of other goals in the model. We find that users tend to prefer specific aggregation rules over others, depending, also, on specific factors.

Keywords: Goal Models, Model Comprehension, Decision Support

1 Introduction

Capturing and modeling stakeholder high-level objectives is an important part of the requirements analysis process. Prior to making any solution decisions analysts need to understand the general and vaguely defined goals that stakeholders consider important and use them as criteria for evaluating alternative solutions. Such high-level goals can be many, with various degrees of importance and interacting in various ways.

Goal models [1, 25, 20] have been suggested to be an effective way to represent goals and the complex interactions between them. Such models consist of various kinds of intentional elements and relationships between them. A particularly interesting type of intentional element used in many goal modeling languages is a goal for which there is no clear-cut criterion for deciding if it is satisfied or not [21, 25]. Examples of such goals are "Happy Customer", "Improve Patient's Experience" or "Ensure Scheduling Fairness". Such goals have traditionally been referred to as *soft-goals* or *quality goals* [16]. As analysts compare solution ideas for the elicited stakeholder problems, these goals serve as criteria to assess the fitness of various possibilities, the latter affecting the former in different degrees.

In goal modeling languages, *contribution* relationships are used to show exactly how satisfaction of one such goal is believed to affect satisfaction of another.

Several approaches exist for modeling contribution links, both qualitative and quantitative. When devising an approach, language designers are confronted with the problem of defining what exactly the contribution links mean and how they can be used, most often in combinations, in order to calculate satisfaction of goals given the satisfaction status of other goals. Different such semantics have been proposed in the literature based on different satisfaction propagation and aggregation rules and techniques. However, given also the abstract nature of the subject matter that these models are meant to represent, how can one evaluate which one is best for adoption in practice?

In this paper, we focus on the intuitiveness of choices of contribution link semantics, understood here as the match between the intended meaning of the language, devised by its designers, and the meaning that the users of the language assign to it. We focus on numeric contribution links and distil from the literature four (4) different theories for contribution link semantics. Then, we perform an experiment with the following goals: (a) understand whether model users who are ignorant to any of the theories perceive contribution semantics in a way that tends to agree (or disagree) to one or more of the theories and (b) identify potential model- or user-related factors that affect such tenancies.

Specifically, we construct a number of goal models containing quality goals connected using numeric contribution links, fixing also the satisfaction level for some of the goals. We present the models to a number of experimental participants and ask them what they think is the most appropriate satisfaction value for a specific goal in the model whose satisfaction level is initially unknown. These are different numbers depending on what contribution semantics one adopts. We present the choices to the users and ask them which one they think is the most appropriate. We observe if there is any concentration of responses to any of the theories and, as such, whether the hypothesis that some semantics match user expectation better than others, is at all plausible. We do find such effects as well as some early indications of factors that can affect participant choices.

The paper is organized as follows. Section 2 presents goal models, contribution links and semantic possibilities thereof in more detail. In Sections 3 and 4 we present the design and results of our experiment. Then in Section 5 we present related work and in Section 6 we offer our concluding remarks.

2 Background

2.1 Goal Models and Contribution Links

A goal model of the kind we consider in this research can be seen in Figure 1 – adapted from Mylopoulos et al. [20]. The model represents a decision problem in the Meeting Scheduling domain. Design alternatives are represented through an AND/OR decomposition hierarchy of hard-goals (ovals), rooted in goal *Schedule Meeting*. The cloud-shaped elements represent *quality goals*, i.e., goals whose

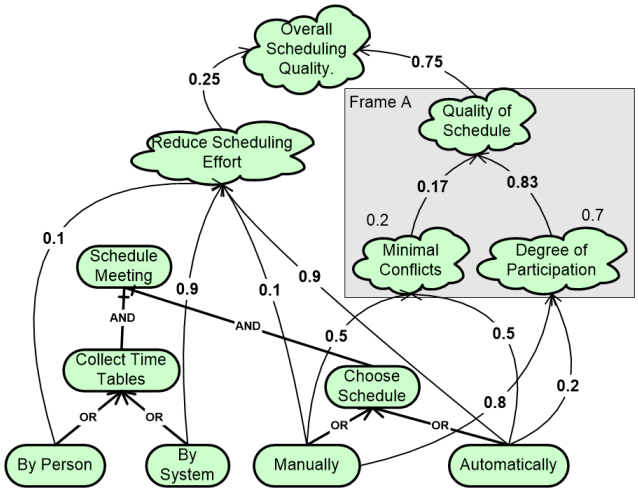


Fig. 1. A Goal Model Represented as a Diagram

satisfaction is generally not defined in a clear-cut manner. Quality goals, written here in an unstructured way, form a separate hierarchy that acts as decision criteria: each alternative of the AND/OR decomposition implies different levels of satisfaction for each of the criteria. Modeling the level and quality of this satisfaction is possible through contribution links that originate from hard goals or quality goals and target (other) quality goals.

Different approaches can be found in the literature on how contribution links can be labelled, and what such labels would mean. Most common are qualitative contribution labels, in which an ordinal scale such as {"– –", "–", "+", "++"} is used [25, 2]. Elsewhere it is proposed that contribution labels can be values from a real interval such as [0,1] (as in Figure 1) or [-100,100] [17, 19, 1, 9]. Most of these proposals come with concrete semantics as to how the contribution label is to be interpreted and used to infer satisfaction of goals from the satisfaction of other goals in the model. In this paper, we focus on quantitative contribution labels and different proposals for their semantics.

2.2 Quantitative Propagation Semantics

An established approach for modeling and reasoning about quantitative contribution links is offered by Giorgini et al. [9]. The framework they propose first assumes that each goal is associated with two variables, each representing the amount of evidence that the goal is satisfied or denied, respectively. The variables take values from the real interval [0,1], 1.0 denoting maximum possible evidence and 0.0 denoting absence of evidence. When two goals are connected through a contribution link, the label of the link describes how the evidence of satisfaction and/or denial of the origin goal affects our belief of satisfaction and/or denial of the destination goal. Specifically the label is a number in the

interval [0,1], which denotes the degree of contribution, a subscript *S, D* or both (denoted through absence of subscript), denoting which of the two variables is considered and a sign "+" or "–" denoting that the contribution is positive or negative with respect to the involved variable.

A second approach to quantitative contribution has been proposed in the context of URN [1] as well as in efforts to combine reasoning about contributions with the Analytic Hierarchy Process (AHP) [17, 19]. In these approaches each goal has one satisfaction value. The label of each contribution link that points toward that goal, denotes the degree by which the satisfaction value of the origin of the link is interpreted into satisfaction of the destination. When AHP semantics are considered, where both contribution labels and satisfaction values can only be positive, the label indicates the *share* of satisfaction influence of each origin goal in calculating the satisfaction of the destination goal. In the Amyot et al. proposal, though, labels can be negative allowing satisfaction of origin goal to actually subtract from the satisfaction of the destination [1].

Given a goal model with numbers such as that of Figure 1, the above proposals can lead to different conclusions as to how satisfaction propagates from one goal to the other. We look into these differences in more detail below.

2.3 Four alternative theories

To allow for a comparison among the contribution modeling frameworks for our purposes here, we make certain assumptions and simplifications. Firstly, we consider simple acyclic hierarchies of quality goals such as the one seen in Figure 1. Secondly, labels are a real number in the interval [0,1] without any subscripts and signs (so assumed to be positive), rounded to two decimal places. Thirdly, only initial satisfaction values are considered, keeping initial denial values zero, when denial variables are defined by the theory at all. These restrictions take away much of the expressiveness of the examined frameworks, but make them comparable with respect to their core semantics, which is our interest here.

We may, thus, attempt a common formulation of satisfaction propagation, which will, in turn allow us to perform a comparison. Thus, let G be the set of all quality goals in the diagram and $s : G \mapsto [0, 1]$ denote the satisfaction value for each of them. Let further O_g be the set of goals g' for which there exists a contribution link from g' to g. Let also $S_g = \{s(g') : g' \in O_g\}$ be the set of satisfaction values of all these quality goals and $W_g = \{w : g' \stackrel{w}{\rightarrow} g; g' \in O_g\}$ the set of all labels of the corresponding contributions links. Then, the satisfaction of goal g is a function f of these two sets: $s(g) = f(S_g, W_g)$.

The literature proposals we discussed above, suggest four possible definitions for f. Three of them come directly from the label propagation framework proposed by Giorgini et al. [9]. According to the proposed algorithm the satisfaction/denial value of every node is always calculated by maximizing individual evidence values formed by the satisfaction/denial values of the origin nodes and the corresponding contribution weights. A generic operator $\otimes$ is used to denote that the two values (satisfaction values of origins and contribution link weights)

are combined to produce a candidate value for the satisfaction of the destination: $s(g) = s(g') \otimes w(g', g)$. Note that given our assumptions of zero initial denial values and positive labels, the denial values are always zero and can, thus, be ignored. There are at least three ways to interpret $\otimes$, which will make for our first three possible definitions of f.

The fourth possible definition of f comes from interpreting how other literature [1, 17, 19] addresses combinations of incoming satisfaction evidence. While label propagation maximizes, these approaches sum-up individual incoming evidence, treating thereby contribution aggregation as a linear combination. Thus, our four possible definitions of f are as follows.

Bayesian, assumes that the satisfaction value of the origin is multiplied by the weight of the corresponding contribution link ($p_1 \otimes p_2 =_{def} p_1 \cdot p_2$). The function f is then defined as:

$$f_b(S_g, W_g) = \underset{g' \in O_g}{MAX}\{s(g') \times w(g', g)\}$$

Min-Max, assumes that $\otimes$ denotes the minimum of the satisfaction value of the origin and the weight of the corresponding contribution link ($p_1 \otimes p_2 =_{def} MIN(p_1, p_2)$). The function f is then:

$$f_m(S_g, W_g) = \underset{g' \in O_g}{MAX}\{MIN(s(g'), w(g', g))\}$$

Serial-Parallel, proposes that $\otimes$ combines the satisfaction value and the weight in a serial/parallel resistance model ($p_1 \otimes p_2 =_{def} p_1 \cdot p_2/(p_1 + p_2)$). The function f is then:

$$f_s(S_g, W_g) = \underset{g' \in O_g}{MAX}\{\frac{s(g') \times w(g', g)}{s(g') + w(g', g)}\}$$

Linear, is similar to the Bayesian with the difference that candidate values are not maximized but added up:

$$f_l(S_g, W_g) = \sum_{g' \in O_g} \{s(g') \times w(g', g)\}$$

Given the above four alternatives, it seems now inevitable to ask what criterion one should use to select a theory for a practical purpose.

2.4 Comparing theories

We view visually represented conceptual models, such as goal models, as devices to be used by humans for comprehending and communicating domain knowledge. Designers of conceptual modeling languages have specific meanings in mind for the constructs they introduce, often in the form of formal semantics as in our case. Such semantics define, among other things, what are correct ways to perform inferences using the information represented in the visualized model. Users of the visualized models, however, may have their own way of interpreting the model constructs and perform inferences accordingly. In other words, users may develop a *mental model* on how the visualization device is supposed to be used

[22, 23]. This model can be due to a combination of factors: potentially partial and incomplete training, experience with similar models and tasks, educational or cultural background and, importantly, the way the model is visually represented – the "system image" according to D. Norman's discussion on mental models for user interface designs [22]. While in interface design designers strive to align their intent on how their devices are supposed to work with the corresponding perception that users develop, in our case, modeling language designers might likewise adjust either the semantics or the visual representation of the language so that the latter evokes correct perception of the former.

In our work we use "intuitiveness" as a working term for describing this level of a match between the designer's intended semantics and the user's assumed meaning. While the former can be drawn from the formal definitions above, the latter needs to be observed empirically. Thus, we measure the meaning users assign to contributions by observing how they perform inferences about goal satisfaction. We particularly perform a simple test: if we provide a decomposition such as that of Figure 1, Frame A to (unsuspecting of any theories) users, how would they combine the numbers to decide a missing satisfaction value? The result of such a test is an assessment of users' expectation of how the numbers presented to them should be combined in order to perform inferences and, consequently, what the meaning of the contribution is.

3 Experimental Study

3.1 Study Design

The main objectives of the study are to: (a) assess whether model users who are oblivious to aggregation theories perceive contribution semantics in a way that tends to agree (resp. disagree) with one or more of the theories, supporting the hypothesis that such theories are more (resp. less) intuitive, (b) explore what factors related to the models or the users affect said agreement (resp. disagreement).

To fulfill these objectives, we first develop a number of goal models. The models consist exclusively of hierarchies of quality goals. We construct a total of nine (9) model structures. The structures are different in a number of ways, including the number of goals they contain, the depth of the hierarchy and the number of contributions they contain. Table 1 describes each structure in detail. As seen in the table, using depth as the primary size measure, we split the goals into three size levels: small, medium and large. The goals of all structures have "dummy" names, A, B, C, etc. For each structure we devise four (4) different concrete models. Each of the four models has a different *number-set*, i.e. set of labels for the contribution links and initial satisfaction values for the leaf level quality goals, the latter presented as an annotation next to the goal. The resulting models look like what is contained in Frame A of Figure 1 (depth = 2, num. of goals = 3, num. of contributions = 2). Given a complete model, one can calculate the satisfaction value of its root using each of the aggregation functions

Table 1. Structure Characteristics

Size	Depth	# Goals	# Contributions
	1	3	2
Small	1	4	3
	1	5	4
	2	5	4
Medium	2	7	6
	2	6	5
	3	7	6
Large	3	9	8
	3	10	9

Table 2. Participant Demographics

	Female	Male	**Total**
Business and Econ.	8	5	**13**
Education	3	3	**6**
Fine Arts	2	3	**5**
Health Sciences	1	1	**2**
Humanities	8	3	**11**
Science and Tech.	3	7	**10**
Social Sciences	6	3	**9**
Total	**31**	**25**	**56**

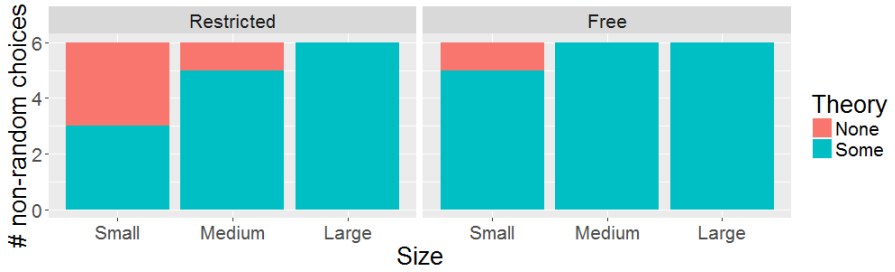
we introduced earlier (f_b,f_m,f_s and f_l), leading to four different corresponding values.

The choice of number-sets deserves further discussion. All values are randomly sampled, under the following conditions. Firstly, for two (2) out of the four (4) number-sets devised for each structure, labels of contributions pointing to the same goal are restricted to necessarily add up to exactly 1.0. For the other two (2) number-sets, such labels need to add up to more than 1.5. We refer to these as the two weighting styles: restricted (to 1.0) and unrestricted or free (to add up to any value above 1.5). Secondly, the four values that result from calculating the satisfaction value of the root goal using each of the four (4) aggregation functions must have a distance of at least 0.08 between each other – the number is the maximum we could achieve across all models. It is important to add that for a given number-set, the satisfaction values that result from applying each theory are ranked almost consistently, due to their mathematical structure. Serial-Parallel in all models gives the smallest number, followed by Bayesian which always is the second smallest. Linear is usually the largest number (∼86% of times in our models) and MinMax is usually (∼86% of times) the second largest.

In all, a total of (9 structures) × (2 weighting styles) × (2 number-sets per style) = 36 distinct models are constructed. The models are used to construct the experimental instrument. The instrument is a sequence of screens/tasks presented to the participants using an on-line survey tool (surveygizmo.com). On each screen the user is presented with one of the 36 models and the four (4) possible satisfaction values for the root goal that result from applying the four different aggregation functions; the values are presented in random order. Participants are asked to choose the "most appropriate satisfaction value" for the root goal. The 36 screens are presented in random order.

Prior to performing the tasks, the participants are also asked to provide demographic information and watch an instructional video. The video introduces goal models, explains what contributions are about and presents the idea that the more the contribution weight or the more the satisfaction of the origin, the more the satisfaction of the destination. It does not, however, provide any information of the precise method to calculate that value in a way that would bias the respondents in the subsequent tasks. Prior to beginning the tasks, participants

Fig. 2. Total Occurrences of Non-Random Preference to a Theory



are also instructed to not use calculator or pen and paper, and try to be quick, i.e. not spend more than half a minute in each screen. The reason for this request is to better simulate natural use of a goal model visualization.

A final question presents the participants with a small sample model and a list of formulae for calculating the satisfaction of the root goal, corresponding to the four theories under investigation. The participants are asked to choose the formula that describes the way they worked in the exercises or describe their own. In a second version of the instrument, this question is replaced with one in which the participants are asked whether they follow a specific calculation method, which they are asked to describe, or whether they *"just used [their] intuition"*.

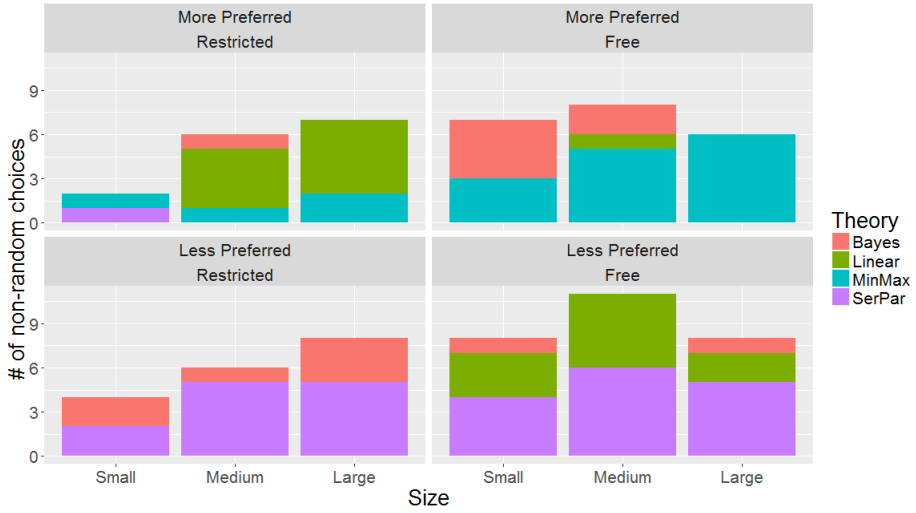
Sixty (60) participants are recruited from Amazon's Mechanical Turk (AMT), an online crowdsourcing platform. In AMT the experiment is posted as a Human Intelligence Task (HIT) for members in the platform. Participants are screened to have at least a bachelor's degree and respond from North America. Half of them use the original instrument and the other half the instrument with the last question changed and at a later time. Data from a total of 56 participants are analyzed – four (4) are excluded for not passing a reliability test. Participants demographics can be seen in Table 2.

3.2 Results

More Preferred and Less Preferred Theories. As a first step of our analysis we test whether the participant responses deviate from the uniform distribution in each of the models. Thus, for each of the 36 models we collect all 56 responses. If, for a given model, participants pick each of the four theories randomly, we expect that the four choices will appear with equal likelihood in each of the 56 ratings. Reversely, if we observe substantial preference (or lack thereof) to one or more of the four categories, then we can suspect that participants do not respond randomly but exhibit preference toward (or against) one or more theories.

Running binomial tests for each model gives us this evidence. Figure 2 shows for how many of the 36 models there was at least one theory choice that was atypically high or low in preference; atypically meaning so high or low that the likelihood of it being due to a uniformly random process is very small $p < 0.05$. The figure organizes those numbers by model size and weighting style. In all

Fig. 3. More Preferred and Less Preferred Theories

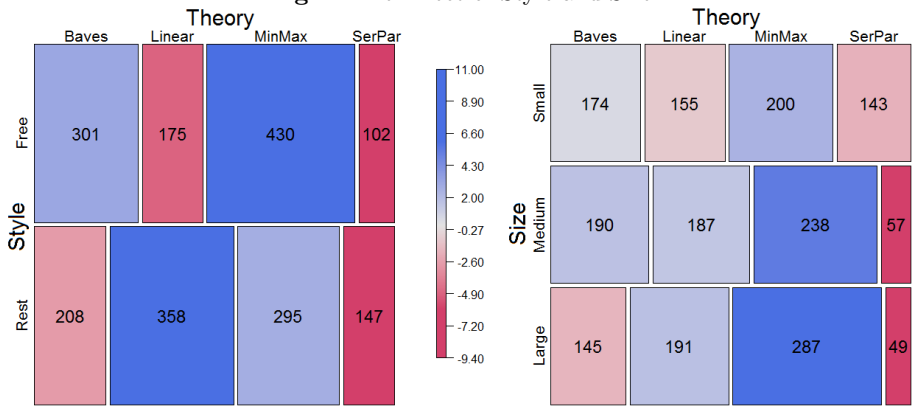


cases, half or more of the models exhibit some deviation from the random and the effect is more pronounced for larger models. Figure 3 further shows for each of those factor configurations, how many times was each theory preferred more or less than uniformly randomly expected. For example, in large models with free weighting, there were six (6) models in which the MinMax theory was chosen more frequently than expected under the uniform assumption (upper right chart), and five (5) models in which the Serial-Parallel theory was chosen less frequently than expected (lower right chart). Note that more than one occurrence of such statistically significant deviations may occur in one model.

We can apply the same logic within the responses of each participant to investigate whether each of them tends to "adopt" a specific theory by choosing it more frequently than expected – under a uniform randomness hypothesis. Indeed, out of the 56 participants only seven (7) seem to respond uniformly across the four (4) theories – i.e. they could be just selecting at random. All the other participant responses tend to concentrate on one or more theories. Thus, for 27, 13, 5 and 2 participants, there is a significant (Binomial test $p < 0.05$) concentration of their choices to MinMax, Linear, Bayes and the Serial-Parallel model, respectively; while for 35, 12, 9 and 4 participants Serial-Parallel, Linear, Bayes and MinMax theories were, respectively, significantly not chosen.

Relevant Factors. Let us now have a closer look into some of the factors that affect choice. We descriptively show these effects (or lack thereof) through mosaic displays [8]. Such displays are clusters of bars, the height of each in the vertical dimension show the relative frequency of the corresponding (y) variable, while the width of each sector within a bar shows the conditional frequency of the horizontal (x) variable. The color of the sector represents deviation from expected frequencies measured through Pearson's residuals $r_i = (n_i - m_i)/\sqrt{m_i}$, where n_i is the observed count and m_i the expected count, again, in our case, of the uniform random case. The darker blue the color the higher the residual,

Fig. 4. The Effect of Style and Size



meaning that the observed count was higher than expected; the darker red the color the lower the residual, hence the observed count is lower than expected.

In Figure 4 we see two such mosaic plots, displaying the distribution of theory choices per weighting style (left) and model size (right). The label of each sector is the number of data points (participant ratings) associated with the sector. We clearly observe differences between the styles of weightings. Thus, unrestricted weighting seems to induce a concentration of choices in the MinMax category to a larger extend than restricted weighting. Importantly while the unrestricted models attract less choice of the Linear model than expected, the restricted models do more so. The reverse is observed for Bayes' models. There is therefore room for a hypothesis that weights adding up to 1.0 evoke a Linear interpretation.

On the right side of Figure 4 the effect of model size appears to be less pronounced, yet notable. The Serial-Parallel theory, in particular, becomes less and less preferred as model size increases. Meanwhile MinMax is slightly more preferred in larger models. Recalling that the Serial-Parallel interpretation is always the smallest number and the MinMax is the second largest, one can as well hypothesize that the larger the model, hence the more the numbers that appear on the graph, the more users will tend to inflate satisfaction values.

A possible suspicion that mathematically-intensive academic background may affect the choice does not seem to be supported by these data. We omit a display for the interest of space.

Self Reporting. The results of the last question, in which participants self-report the method they think they used, strongly indicate that participants are not completely aware of the method they use. Only 20% of 35 participants who specified a concrete calculation method either forcedly (version 1 of instrument) or voluntarily (version 2), state that they use a method that also happens to be the one they actually used with statistically significant consistency. A higher 26% claim that they follow a theory which, in fact, they used unusually less in the exercises (chiefly Serial/Parallel). Importantly, of the participants who were asked if they used their intuition to respond (version 2), 81% states that they did, i.e., they did not use a specific calculation method.

4 Consequences and Validity Threats

Consequences. The general impression we get from the result is that untrained users of quantitative goal models may not come without expectations as to how numbers are supposed to be combined to infer goal satisfaction, and that such expectations may depend on aspects of the model. More specifically, we believe that our data seem to support further corroboration of at least four hypotheses. Firstly, for visual goal models as constructed in this experiment, participants tend to favour certain ways of inferring satisfaction of goals over others, particularly MinMax, Linear and, to a lesser extent, Bayesian. Secondly, the amount to which the weights of incoming contribution links to a goal sum up can affect the choice of interpretation of satisfaction propagation semantics; if the sum is 1.0 the Linear model becomes more popular. Thirdly, the larger the model is the more inflated the assessment of goal satisfaction appears to be. Finally, users do not appear to consciously follow a specific aggregation method but instead work intuitively.

There are, further, some important experimental validity points that deserve a closer look, particularly on construct and external validity.

Construct Validity. As we saw, to measure which theory users prefer we mainly rely on inference from how they use the models rather than on directly asking them (e.g., *"how would you combine these numbers?"*). This emphasis was in part due to practical reasons – on-line administration prevents meaningful open-ended interaction – but also due to our low confidence that users can provide valid data. The limited self-reporting we solicited (last questions of instrument) indeed revealed that participants have limited awareness of the process they themselves follow. Moreover, the input of those who volunteered to describe the method they followed in their own words proved difficult to interpret and was often plain incomprehensible. Thus, we remain unconvinced that there is a trivial interviewing protocol that can conclusively explain why participants work the way they do. It is, however, subject for future research.

Looking now at the observational measure, the substantial deviation from uniform randomness begs an explanation and supports, we believe, the validity of the endeavour: why are some theories preferred and some others avoided? One explanation is that the participants were asked to choose from a fixed set of values and, thus, naturally leaned toward those that were not extreme, choosing completely randomly one of them. This could explain why they avoided, for example, Serial-Parallel. But it would not completely explain why Linear was not avoided to the same extend, and why there was still concentration to MinMax versus the Bayesian theory – which both give values which are, generally, in the middle of the ranking. Future designs could allow for a more solid picture of the above by asking participants to freely specify satisfaction value that they find more appropriate, instead of offering a predefined inventory.

External Validity. We treat this study as exploratory, with no intent of making strong generalization arguments, about e.g. the universal suitability of a specific theory, our goal being to see if there is *any* effect. Keeping this in mind, in appreciating generalizability of the findings one should consider both the chosen

participants and the chosen models. The former are users of MTurk, who are known to be a good enough proxy for random population samples [5], and might offer more variability than e.g. University students, especially when the latter are drawn from a specific department or course. More important is, we find, the level of representativeness of the models: different sizes, visual properties and goal contents (e.g. real domain concepts vs. A, B, C etc.) might certainly affect participants' reaction to them. Recall also that to enable comparability of the frameworks certain simplifications were made, such as for example not using negative contribution measures or not fully utilizing notions of satisfaction and denial values as defined in the Giorgini et al. framework. Generalizations should be predicated on these restrictions.

5 Related Work

There is a wealth of proposals for modeling partial goal satisfaction and influence thereof between goals in the literature, the semantics of which vary in intuitive meaning and their mathematical/algorithmic treatment (e.g. [15, 10, 14, 3, 7, 18] in addition to ones discussed earlier; [12] for further survey). Such proposals are typically evaluated based on expressiveness standards, amenability to interesting and efficient reasoning or hypothesized ease of label acquisition.

Nonetheless, the idea of also empirically investigating the way a diagram elicits by its viewers a certain way of understanding its subject matter is not new to the conceptual modeling community, either. Several studies, for example, investigate the comprehensibility of diagrammatic notations such as UML state diagrams or ER diagrams [6, 24]. Similar work has been done with goal models. Horkoff et al., for example, propose and evaluate an interactive evaluation technique for goal models [13]. The visual properties of goal modeling languages such as i^* vis-a-vis model comprehensibility have been the target of investigation as well. Moody et al. offer an assessment of the i^* visual syntax based on established rules ("Physics of Notations"). An empirical analysis was followed by Caire et al. [4] in which experimental participants evaluate visualization choices of the language's primitives. Elsewhere, Hadar et al. [11] compare goal diagrams with use case diagrams on a variety of user tasks, including reading and modifying.

The above efforts tell us that there is interest in the community in understanding how users interact with diagrams, and even have users define their visual properties of such, as e.g. Caire et al. demonstrate. Having users go beyond the visuals and evaluate the semantics of notations seems to be a natural next step. On that matter, although we could not find work in which interpretation of satisfaction contribution in goal models is empirically investigated the way we do here, we believe there is potential for much more research.

6 Concluding Remarks

We presented an exploratory experimental study aimed at assessing the intuitiveness of four theories of satisfaction propagation, operationalized through

measuring the frequency by which inferences untrained users perform with the model match inferences that the theory prescribes. The results suggest that participants do not choose at random and tend to favour some theories over others. The way numbers are chosen as well as the size of the model also seem to affect selection of theory, a process which, moreover, appears to take place heuristically rather than through performance of precise calculations.

More investigation will be needed to fully understand how such results may affect the practice of goal modelers and goal modeling language designers. It is important to first consider that the results concern a specific diagrammatic way of visualizing goal models and the kinds of inferences the specific visualization evokes. If a modeler has compelling theoretical reasons to choose an "unintuitive" (vis-à-vis the visualization) propagation theory, e.g. Serial/Parallel if it eventually proves to be such, use of traditional goal diagrams may be problematic, as users will likely make goal satisfaction inferences that contravene the normative values, perhaps even if the latter are explicated in the diagram for the purpose of exactly preventing erroneous user inferences. We intuitively consider such situation sub-optimal compared to a situation in which the visualization and the theory are in alignment. Nevertheless, the impact of misalignment in practical model use needs to be explored in realistic model use scenarios (e.g. decision making), prior to elevating intuitiveness measurement to a priority for language designers.

References

1. Amyot, D., Ghanavati, S., Horkoff, J., Mussbacher, G., Peyton, L., Yu, E.S.K.: Evaluating goal models within the goal-oriented requirement language. International Journal of Intelligent Systems 25(8), 841–877 (2010)
2. Amyot, D., Mussbacher, G.: User Requirements Notation: The first ten years, the next ten years. Journal of Software (JSW) 6(5), 747–768 (2011)
3. Baresi, L., Pasquale, L., Spoletini, P.: Fuzzy goals for requirements-driven adaptation. In: Proceedings of the 18th IEEE International Requirements Engineering (RE'10). pp. 125–134. Sydney, Australia (2010)
4. 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'13). pp. 115–124 (2013)
5. 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), 1–18 (2013)
6. 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 Engineering 14(6), 685–719 (2009)
7. Elahi, G., Yu, E.S.K.: Requirements trade-offs analysis in the absence of quantitative measures: a heuristic method. In: Proceedings of the 2011 ACM Symposium on Applied Computing (SAC'11). pp. 651–658. TaiChung, Taiwan (2011)
8. Friendly, M., Meyer, D.: Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data. Chapman Hall, NY, USA (2015)

9. Giorgini, P., Mylopoulos, J., Nicchiarelli, E., Sebastiani, R.: Formal Reasoning Techniques for Goal Models. Journal on Data Semantics, LNCS vol. 2800, 1–20. (2003)
10. Giorgini, P., Mylopoulos, J., Sebastiani, R.: Goal-oriented requirements analysis and reasoning in the Tropos methodology. Engineering Applications of Artificial Intelligence 18(2), 159–171 (2005)
11. 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. Information and Software Technology 55(10), 1823 – 1843 (2013)
12. Horkoff, J., Yu, E.: Analyzing goal models: different approaches and how to choose among them. In: Proceedings of the 2011 ACM Symposium on Applied Computing (SAC'11). pp. 675–682. TaiChung, Taiwan (2011)
13. Horkoff, J., Yu, E.S.K.: Interactive goal model analysis for early requirements engineering. Requirements Engineering 21(1), 29–61 (2016)
14. van Lamsweerde, A.: Reasoning about alternative requirements options. In: Conceptual Modeling: Foundations and Applications, LNCS vol. 5600, 380–397 (2009)
15. Letier, E., van Lamsweerde, A.: Reasoning about partial goal satisfaction for requirements and design engineering. In: Proceedings of the 12th International Symposium on the Foundation of Software Engineering FSE-04. pp. 53–62. (2004)
16. Li, F.L., Horkoff. J., Mylopoulos. J., Guizzardi, R.S.S., Guizzardi, G., Borgida, A., Liu, L.: Non-functional requirements as qualities, with a spice of ontology. In: Proceedings of the 22nd International Requirements Engineering Conference (RE'14). pp. 293–302. Karlskrona, Sweden (2014)
17. Liaskos, S., Jalman, R., Aranda, J.: On eliciting preference and contribution measures in goal models. In: Proceedings of the 20th International Requirements Engineering Conference (RE'12). pp. 221–230. Chicago, IL (2012)
18. 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'13). pp 19–32. Hong-Kong, China (2013)
19. Maiden, N., Pavan, P., Gizikis, A., Clause, O., Kim, H., Zhu, X.: Making decisions with requirements: Integrating i* goal modelling and the AHP. In: Proceedings of the 8th International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ'02). Essen, Germany (2002)
20. Mylopoulos, J., Chung, L., Liao, S., Wang, H., Yu, E.: Exploring alternatives during requirements analysis. IEEE Software 18(1), 92–96 (2001)
21. Mylopoulos, J., Chung, L., Nixon, B.: Representing and using nonfunctional requirements: A process-oriented approach. IEEE Transactions on Software Engineering 18(6), 483–497 (1992)
22. Norman, D.: The Design of Everyday Things. Basic Books, NY, USA (2013)
23. Payne, S.J.: A descriptive study of mental models. Behaviour & Information Technology 10(1), 3–21 (1991)
24. Purchase, H.C., Welland, R., McGill, M., Colpoys, L.: Comprehension of diagram syntax: an empirical study of entity relationship notations. International Journal of Human-Computer Studies 61(2), 187 – 203 (2004)
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)