Comparing the comprehensibility of numeric versus symbolic contribution labels in goal models: an experimental design

Sotirios Liaskos
School of Information Technology
York University
Toronto, Canada
liaskos@yorku.ca

Wisal Tambosi
School of Information Technology
York University
Toronto, Canada
w.tambosi@gmail.com

***Abstract*—Goal models have been suggested to be an effective way to support decision making in early requirements engineering. Such models are capable of representing a large number of alternative ways to solve stakeholder problems and comparing them against each other with respect to higher level objectives. Core to the realization of such analysis is the concept of the contribution link that represents how satisfaction of one goal affects satisfaction of another. Many ways for representing and assigning precise meaning to contribution links have been proposed, each with different properties and advantages. But which one agrees more with user preferences on how such links should be used? In this paper, we present an experimental design for comparing two ways for representing contribution links, symbolic versus numeric, with respect to how accurately and quickly users identify optimal decisions using each representation format. Apart from comparing the two representation techniques and advising the modeling practice accordingly, the study aims at showing how a quality construct we call intuitiveness can be added to the range of criteria a modeling language designer has at her disposal for evaluating her language design decisions.**

I. Introduction

Goal models [1]–[3] have long been proposed as an effective means for representing intentional structures and their relationship to decision problems in early requirements engineering [4]–[6]. Using such models, business analysts can capture the variety of ways by which stakeholders can solve their business problems and compare them with one another with respect to set criteria.

Many representational and semantic frameworks have been proposed within the goal modeling community to allow such analysis [5]–[8] ([9] for a survey). One of the fundamental constituents of goal models that allow such analysis is the concept of the *contribution link*, which is a representation of a relationship between two goals signifying how satisfaction of one affects the satisfaction of the other. Different goal modeling and analysis frameworks propose different ways to visually represent and assign meaning to the contribution concept. The traditional/de-facto representation choice is qualitative (symbolic) labels signifying the quality of contribution (positive or negative) and crudely characterizing the size of the contribution. However, the use of quantitative (numeric) values has also been proposed, whereby, e.g., sign and absolute value are used to represent quality and size of contribution. These representational options have been studied from a theoretical point of view and different formal semantics have been proposed, each showing how the representations allow inference of satisfaction status of one goal from that of other goals.

However, limited work has been done in terms of how users of the models perceive what the symbols and/or numbers mean and how they expect to use them in order to make inferences pertinent to decision making. It is particularly useful to understand how users *intuitively* assign meaning to signifiers within the language, when no prior training and/or experience with the language can be assumed for them. Knowing what untrained user's intuition is, language designers can settle for representations and semantics that are closer to the user's expectations and, as such, easier to learn and more accurate to use.

In this paper we present an experimental design aimed at comparing the intuitiveness of qualitative versus quantitative contribution labels in goal models, having assumed specific semantics for each. Our design aims at showing which of the two visualization-meaning pairs leads to more accurate decisions in the least amount of time.

The rest of the paper is organized as follows. In Section II we offer some background on goal models, contribution links and their semantics. In Section III we describe the experimental design and in Section IV we summarize and review some of the related work.

II. Background

A. Goal Models and Contribution Links

The goal models we consider in this study look like the ones in Figure 1. The nodes (ovals and clouds) are *goals* that describe states of the world that the actor in question (circular shape) has within their scope (large shaded dashed circle) and want to achieve or maintain. The ovals describe *hard-goals*, which are goals that come with a clear way to decide when they are satisfied, while *soft-goals* (the clouds) are goals for which this is not the case.

Goal modeling languages define a variety of relationships between goals and allow for great structural freedom [10]. However, in our study we restrict our focus to goal models that have specific structural characteristics. Thus, through *means-ends* and *decomposition* links, hard-goals form an AND/OR decomposition tree whose solutions describe alternative ways by which the root hard-goal can be satisfied. Soft-goals on the other hand form their own hierarchy using *contribution links*, the curved directed lines. Similar lines connect some hard-goals with some soft-goals.

A contribution link shows in what way satisfaction (or not) of the origin of the link affects satisfaction (or not) of the destination of the link. This way of affecting the other goal is described through the label of the contribution link. Typically the label will show whether the effect is positive or negative and/or how large it is. Nevertheless, there are more than one ways to represent contribution labels and, for each, multiple ways to define their semantics.

The original and seemingly most popular approach to modeling contribution labels is through *symbols* (diagram on the left in Figure 1). Thus “+”, “++”, “−” and “−−” denote respectively positive (“helps”) very positive (“makes”), negative (“hurts”) and very negative contribution (“breaks”). Alternatively *numbers* can be used to convey this information (diagram on the right in Figure 1). Two distinct numeric approaches have been introduced in the literature. The approach by Giorgini et al. [8], [11] assigns a number in the real interval [0.0,1.0] to represent size of contribution and a sign to represent positive or negative contribution[1]. The AHP-inspired “linear” interpretation [12] also adopted by URN [7] simply assigns a number in the real interval [0.0,1.0] denoting the share of contribution of the origin goal to the destination goal.

B. Contribution Semantics

Informal descriptions such as the above about the meaning of the contribution link allow a model reader/user (henceforth simply *user*) perform some very basic *inferences* by looking at the model. For example, she can compare two contributions with respect to which one is larger or she can even choose between alternatives in the hard-goal decomposition with respect to a soft-goal of interest. For example, in the symbolic model on the left side of Figure 1, if to *Reduce Scheduling Effort* is an important soft-goal, then we know that *(Choose Schedule) Automatically* is preferable than doing so *Manually*, by simply looking at the contribution labels and without knowing precisely what they mean. However, more detailed semantics need to be given in order to perform more complex inferences such as deciding on the satisfaction status of a goal that receives multiple incoming contribution links, or, as we will see below, deciding the optimal alternative by considering all contribution links in the structure.

Label	Effect	Label	Effect
++	FS → FS	−−	FS → FD
	PS → PS		PS → PD
	PD → PD		PD → PS
	FD → FD		FD → FS
+	FS → PS	−	FS → PD
	PS → PS		PS → PD
	PD → PD		PD → PS
	FD → PD		FD → PS

TABLE I
SYMBOLIC CONTRIBUTION SEMANTICS

Giorigini et al. have developed the most expressive semantics for both symbolic and numeric links [8], [11]. According to their framework each goal in the diagram can be associated with two variables: one that measures satisfaction and one that measures denial. In the *qualitative (symbolic)* framework each of these variables can take one of three values: Full evidence (denoted with prefix **F**), Partial Evidence (**P**) and No Evidence (**N**) – of, respectively satisfaction (suffix **S**) or denial (**D**). For example, for a goal we may have partial evidence of satisfaction and no evidence of denial (denoted {**PS,ND**}) and, for another, full evidence of satisfaction and partial evidence of denial ({**FS,PD**}); the inconsistency is perfectly acceptable and the framework's ability to represent it is one if its strengths. A set of rules, seen in Table I, combine the satisfaction and denial values of the origin goal with the contribution label to decide the satisfaction and denial values of the destination. Returning to Figure 1 (qualitative model on the left), if we know that satisfaction and denial values of *Minimal Conflicts* are {**FS**,**PD**} then based on the rules of Table I *Quality of Schedule* must be {**PS**,**PD**} – assuming no other influence.

In the *quantitative (numeric)* framework the rules are replaced by algebraic formulae. The researchers allude to three possible ways by which this formula can be structured, seen in the top three rows of Table II; in practice their framework is open to the adoption of many other ways. Given a set of goals $g' \in O_g$, each with satisfaction value $s(g') \in [0.0, 1.0]$ targeting goal g with contribution links weighted as $w(g', g)$, the satisfaction value of goal g is expected to be $s(g)$ as defined in each of the formulae. In all the proposed formulae (*“Bayesian”*, *“Min-Max”* and *“Serial-Parallel”*) aggregation is implemented through maximization. Note that in this semantic framework, users are supposed to understand the numbers of the contribution links as absolute contribution values potentially elicited and understood in isolation from the other ones.

A different interpretation of numeric contributions, which is of particular interest here, is the de-facto approach followed by URN [7] which has been studied by Liaskos et al. [12]. According to that interpretation, a unique numeric satisfaction value is assigned to each goal with values in the real interval [0.0,1.0] – so no distinct satisfaction and denial values. Then, the number on the contribution link denotes the share of contribution of the satisfaction of the origin goal to the

[1] Giorgini et al.'s expressive framework also includes a subscript representing what is being contributed between satisfaction and denial; both their quantitative and qualitative version includes this dimension. Presentation of this dimension is outside our scope.

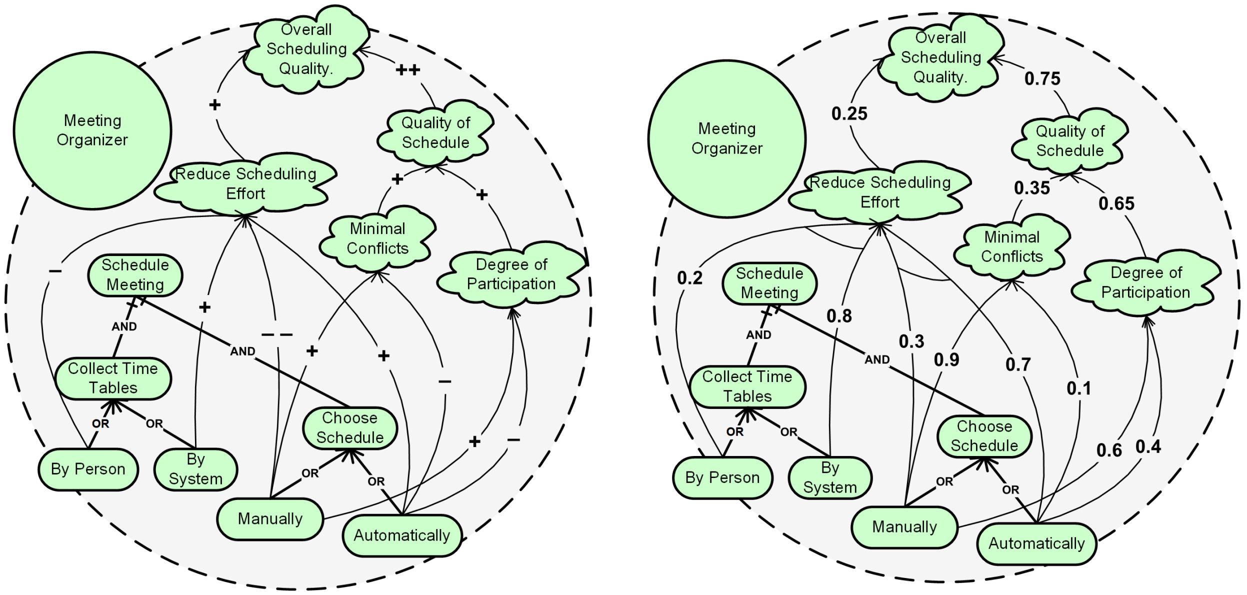

Fig. 1. Goal models with symbolic (left) and numeric (right) contribution links.

satisfaction the destination goal. This implies also a different formula for satisfaction propagation, the last one on Table II; the formula is labeled as *"Linear"* for it calculates the satisfaction of the destination goal through linearly combining the satisfaction value of each goal that influences it, using the numbers on the contribution links as weights for the linear combinations.

Bayesian	$s(g) = \underset{g' \in O_g}{MAX}\{s(g') \times w(g',g)\}$
Min-max	$s(g) = \underset{g' \in O_g}{MAX}\{MIN(s(g'), w(g',g))\}$
Serial-parallel	$s(g) = \underset{g' \in O_g}{MAX}\{\frac{s(g') \times w(g',g)}{s(g') + w(g',g)}\}$
Linear	$s(g) = \sum_{g' \in O_g}\{s(g') \times w(g',g)\}$

TABLE II
NUMERIC CONTRIBUTION SEMANTICS

While the linear interpretation is arguably less expressive and imposes structural limitations to the models (the soft-goal sub-graph must be acyclic) they have been found [12] to be amenable to systematic elicitation through an established decision making technique, the Analytic Hierarchy Process (AHP) [13]. Following AHP, contribution values are not assigned directly but through pairwise comparisons followed by transformation of the output of these comparisons into the final values, controlling also for the consistency of the input, via calculation of a Consistency Ratio (CR). Given this promise of the linear interpretation for practical use, we adopt it as the quantitative interpretation of choice in the study we propose here.

C. (A case for) the Intuitiveness Construct

Given the above options for visually representing and understanding the use of contribution labels for inferring satisfaction propagation, it is natural to ask which one is more "friendly" to users of the models. One aspect of "friendliness" is the level by which the intended meaning and use of the contribution aligns with the users' intuition.

We use the (working) theoretical construct "intuitiveness" of a model construct to describe the ability of untrained users of a conceptual model to readily understand what the construct means and how it should be used to make inferences in the model. The concept is analogous to the idea of an intuitive human-machine interface: the more intuitive an interface is, the more readily first-time users can use it without the need to resort to help, a manual etc. The term is akin to that of *learnability* which is a quality of an interface that allows users to learn how to use it easily and quickly [14]. One can think of intuitiveness as a facilitator of learnability. Design principles such as consistency and compliance to standards [15] are understood here to facilitate intuitiveness: users will likely find intuitive a user interface that uses conventions with which the user is already familiar.

With this user-machine interface analogy in mind, we can reasonably claim that conceptual models are also artifacts to be efficiently used by people, where "use" here is "understanding and communication" [16]. Further, as design artifacts themselves, modeling languages are results of design decisions at two levels: at the level of the concepts they consider (e.g., hard-goals and soft-goals) and at the level of the visualization of those concepts (e.g., ovals and clouds). It appears that there might be better and worse decisions for each of those levels. For example, would we instead of ovals and clouds use animal pictures (e.g. elephants and dolphins) to represent hard-goals and soft-goals? Likewise, are the concepts "upper-goal" and "lower-goal" more successful choices for representing human intention than currently used concepts "hard-goal" and "soft-goal"?

Intuitiveness, as we conceptualize and apply it here, measures the entire package of a concept and its visualization:

the visualization evokes a meaning, which, in turn, is used to make inferences. When a user is exposed to a visualization and ends up performing an inference that is not intended by the designers, a sub-optimal decision may be claimed at any of the levels: either the users did not map the visualization to the right concept (e.g. confused a "goal" for an "event", both otherwise being clearly understood concepts), or they did so correctly but did not understand the concept as the language designers intended them to (e.g., they correctly mapped a symbol to an "upper-goal" but didn't know what to do with the latter). While training may arguably establish correct bridging between visualization and inference in the long term, intuitiveness is exhibited when limited such training is necessary.

In the context of contribution links in goal models, the inference we are interested in is how users assign satisfaction to goals given satisfaction of other goals based on their own interpretation of what contribution labels seem to mean. Reversely, their observed inferences reveal their perceived meaning of the links, and, as such, the former can be used to develop empirical operationalizations of the latter. In the experiment we describe below, we ask the users to make decisions using goal models. To do so, they need to adopt a way of using the contribution link and, implicitly, a semantics for those links. The alignment of the semantics implied by how users use the models with the designed semantics (i.e. the semantics intended by the designers), as exhibited by whether the results of the inference match, is, we claim, a possible indication of the intuitiveness of the designed semantics.

III. Experimental Design

A. *Overview and Research Question*

In the proposed study we pick two approaches for modeling and assigning meaning to contribution links and compare them with regards to measures of intuitiveness and efficiency. We specifically compare the symbolic against the numeric approach, the latter under the linear interpretation. There is one main research question we wish to address:

RQ. Which of the two methods for modeling contribution links is the most (a) intuitive and (b) efficient for the task of identifying optimal alternatives in goal models?

We address the above through a controlled experiment with human participants.

B. *Experimental Tasks and Measurements*

1) Measures: The two constructs we are considering are intuitiveness and efficiency. We theoretically defined intuitiveness as the degree by which untrained users can make accurate inferences with models they are exposed to. Operationally, we will measure intuitiveness by exposing the experimental participants to a sample of goal models and asking them to perform an inference, which we then compare with the "correct" inference as dictated by the adopted contribution modeling approach. Perception of intuitiveness is also included as a possible measure via self-reporting of participants' confidence about the aforementioned inferences they perform. Efficiency, will be, in this context, measured as the total time it takes for participants to perform this inference, independent of correctness.

2) Experimental Units: We develop a number of goal models such as those in Figure 1. We specifically develop two (2) sets of models: qualitative, in which contribution lables are symbolic, and quantitative, where contribution labels are numeric following the "linear" semantics. All models contain one OR-decomposition of hard-goals (so one decision) together with a hierarchy of soft-goals that are used as criteria for choosing the optimal alternative within the decomposition. By having a unique root goal in the soft-goal hierarchy the goal model implies that, generally, one of the depicted alternatives is optimal compared to the others.

To show how this is possible let us go back to Figure 1 and consider the decomposition *Manually* versus *Automatically*. We can assume that whenever we pick one of the alternatives the corresponding hard-goal is assigned maximum satisfaction and, if applicable minimum denial value. Thus, to choose the alternative *Manually* we assign to it maximum satisfaction values {**FS**, **ND**} (qualitative case) or *s(Manually) = 1* (quantitative case), and to all other alternatives (in our case only *Automatically*) values {**NS**, **ND**} or *s(·) = 0*. We then perform recursive bottom-up application of the propagation rules of Tables I and II (depending on case), in order to calculate the satisfaction of the root goal *Overall Scheduling Quality*. For the quantitative models specifically we follow the linear interpretation of the last row of Table II. Different choices of alternative will result in different satisfaction level for the root goal. The alternative that results to the highest satisfaction value for the root goal is the optimal.

In the quantitative case, satisfaction is a unique value and the comparison straightforward. In the example of Figure 1 (model on the right), *Manually* causes satisfaction of *Overall Scheduling Quality* by approx. 0.6 compared to approx. 0.4 implied by selection of *Automatically*. Thus, *Manually* is the optimal alternative[2].

In the qualitative case, calculation is less straightforward in that there are two variables to consider, satisfaction and denial. To make different satisfaction levels comparable we aggregate the two values into one, the *aggregated satisfaction value*. To calculate the aggregated satisfaction values, we firstly associate qualitative satisfaction labels {N, P, F} with numeric values 0,1,2, respectively. We denote the resulting numeric satisfaction and denial of a goal g as $sat(g)$ and $den(g)$, respectively. The aggregated satisfaction value is then $sat(g) - den(g)$ which results to an integer in [-2,2]. Thus, the aggregated satisfaction value of a goal g_1 with {**PS**, **FD**} is $sat(g_1) - den(g_1) = 1 - 2 = -1$ and of a goal g_2 with {**FS**, **ND**}, $sat(g_2) - den(g_2) = 2 - 0 = 2$. For the qualitative

[2] To simulate the experience of our experimental participants the reader can look at the diagram and verify if the assertion that *Manually* is optimal can be inferred intuitively, by roughly comparing the numbers and without performing precise calculations.

model on the left of Figure 1, it can be verified that *Overall Scheduling Quality* is {**PS**, **PD**} for *Manually* and also {**PS**, **PD**} for *Automatically*. Hence, both alternatives lead to the same aggregated satisfaction value for the root goal, that is 0, and as such they are equally optimal.

3) Model Sampling: To develop the samples of goal models that we need, we pick a goal structure (more below) and randomly choose contribution link labels, such that the distance in satisfaction value of the best alternative (i.e., optimal with respect to the root soft-goal) compared to the second best alternative is controlled to not exceed or be less than a fixed value. Thus, we ensure that the distance is neither too large so that the task of identifying the optimal alternative is trivial in all cases, nor too small to constitute an unimportant distance in terms of decision making and also be impossible to detect even by some of the participants.

Specifically, in **qualitative** models contribution labels are assigned randomly one of the labels "++", "+", "−−", and "−", such that the first alternative has a distance from the second alternative of two (2) levels of satisfaction, based on the aggregated satisfaction value of the root soft-goal that each alternative results in.

Thus, a goal model in which the best alternative, when chosen, makes the root goal {**FS**, **ND**}, hence aggregated value $2 - 0 = 2$ and the second best makes the root goal {**PS**, **PD**}, hence aggregated value $1 - 1 = 0$, qualifies for inclusion to our sample as the distance of the two top alternatives is 2. A goal model, on the other hand, in which the top two alternatives are both {**FS**, **PD**} have both an aggregated value of 1 and hence distance of zero; so they do not qualify.

In **quantitative** models we also randomly sample while ensuring that the first alternative has a distance of 0.4 from the second; again, in terms of the satisfaction they imply for the root goal. For example a set of weights that gives satisfaction value 0.7 to the first alternative and 0.3 to the second qualifies for inclusion to our sample. The model of Figure 1 (right), focussing on the *Choose Schedule* decision, does not qualify as the distance is $0.6 - 0.4 = 0.2$

The choice of 0.4 is made to match the corresponding choice in the qualitative models. Observe that in qualitative models the maximum distance between alternatives is 4 ({FS, ND} versus {NS, FD} so 2 - (-2)). The distance we demand is 2, thus half of this space. Respectively in the quantitative models the maximum theoretical distance is 1.0, so half the space would be 0.5. However we end-up to 0.4 – biasing slightly against numeric models – as for some of our structures there does not seem to exist combinations of numeric labels that yield a distance of exactly 0.5.

To remain consistent with the claim that linear interpretation is chosen due to the systematic elicitation approach that is afforded by it, namely AHP, all numeric sampling is done through simulated AHP pair comparison processes and subsequent profile calculations, such that the consistency ratio (CR) is less than 0.1.

It is worthwhile to note that, while it is understood that the two representation approaches have different precision levels, the numeric one being understood as more precise, this difference does not seem to threaten our comparison effort but rather offer us a possible explanatory view to a potential result. If, for example, a difference is discovered in favor of the numeric format, it might be due to a number of reasons including precision but also, e.g., familiarity of the participants with numerical reasoning and assessment of proportions. Identifying those precise reasons – assuming the effect is eventually observed – is a matter for future investigation.

4) Instrument and Tasks: Using the sampling procedure described above we develop a total of twelve (12) *quantitative* models. The goal structures refer to three (3) different domains describing intentional structures in the context of decisions: Choosing an Apartment, Choosing a Course, and Choosing a Means of Transportation. We develop the models based on specific domains, rather than using dummy names (A, B, C etc.) for the purpose of making the tasks more realistic. This introduces the threat that participants may use their own opinion of how goals are related to each other ignoring the information provided in the contribution link. To avoid this bias, participants are told that the structures represent decision problems of a third party and that their task is to help that party make the decision based on the priorities of that party as these priorities are represented in the goal structure.

For each domain we develop two (2) structures (one with two and one with three alternatives) and for each structure we sample two (2) *labels-sets* (i.e., sets of labels for the contributions) sampled as described above. To produce *qualitative* counterparts we simply copy the twelve (12) quantitative structures and replace the numbers with randomly sampled symbolic labels – again, as described above.

We then present the resulting twelve (12) models of each type (qualitative and quantitative) to the participants one after the other asking each time what they believe the optimal alternative is. Domains are presented in random order and models within the domains in random order as well. Three video presentations precede these tasks: one describes decision problems in general, another introduces goal models and a third one introduces the three domains. The second video specifically, describes the intuition behind the contribution links of each type carefully without getting into the mechanics of satisfaction propagation. The videos are scripted and are the same in the two cases (qualitative and quantitative) except obviously for the places where the numbers or symbols are presented.

The videos are chosen as the instruction method for three reasons (a) allow for repeatability of the procedure, (b) control for biases in training, and (c) allow for remote administration or administration by non-experts.

A simple demographics questionnaire (age, sex, education, prior knowledge of goal models) precedes the main test. Participants are unlikely to be familiar with goal models, and input coming from those who actually are will be discarded. However, if familiarity to goal modes turns out to be more prevalent, treating familiarity as a covariant is another option.

5) Participant Sample: We plan to consider the University student pool as the population to opportunistically sample from, specifically intermediate/senior students from various disciplines. We claim that this does not harm the generalizability of the particular study. Firstly, having a valid noteworthy intuition about how the particular conceptual modeling construct works does not seem to require experience and skill in any specific field: goal models refer to concepts (goals and their fulfillment) that should be accessible to anyone who has successfully entered post-secondary education – compared to, for instance, component diagrams describing software designs. Secondly, it seems to be the implicit ambition of goal modeling language designers that goal models are artifacts that not only analysts but also stakeholders are able to comprehend and use to their benefit [17]. If this is the case, then the population we should be drawing participants from is, roughly speaking, the population of all people who might serve as decision making stakeholders in a systems development project. While there is no authoritative data about the characteristics of this population, we believe that the breadth of educational and skill profiles in it can be credibly approximated by a sample of intermediate/senior University undergraduate students.

6) Variables and Analysis Approach: It becomes obvious from the above that the experiment is a simple comparison between two levels (qualitative vs. quantitative) of one independent variable/factor (contribution link representation method) arranged in a *between-subjects* fashion. Dependent variables are the *accuracy* measured as the number of correct responses per participant, hence a number in [0,12], as well as *response time* which is the average time participants need in order to provide a response.

It is also possible to measure confidence in each participant's response as a measure of perceived intuitiveness. In earlier studies [18], [19], we augmented each exercise with a 5-level Likert-style question "how confident are you of your answer above", with possible answers Very Unconfident, Confident, Neutral, etc. The higher the confidence the higher the perceived intuitiveness, i.e., how intuitive the participants think the representation is. However, this additional question increases experimental time and fatigue. Addition of this variable would depend on our ability to keep the instrument short, i.e., around 30 to 40 minutes.

Simple comparisons between means appear to be sufficient as a statistical procedure, with the expected deviation from normality kept in view – the scale [0,12] is particularly inviting for ceiling effects.

IV. Summary and Related Work

We presented an experimental design for comparing the intuitiveness of symbolic versus numeric contribution links in goal models. We use intuitiveness as our main comparison construct, defined as the ability of novice users of the notation to correctly understand how they can use it. We operationalize intuitiveness by measuring agreement between authoritative inferences and inferences participants make, as well as the time it takes for the latter to take place. We also include the option of measuring perceived intuitiveness via self-reporting the confidence of participants on their inferences. Our design relies on random sampling of a number of goal structures depicting a decision problem and asking participants to choose the optimal of the available choices, thereby making intuitive inferences about contribution links. The decision problems are carefully sampled to allow for a controlled distance between the optimal and second optimal choice.

Empirically evaluating the effectiveness of diagrammatic notations has been widely studied in the literature. Much of the research in the field has been dedicated towards understanding the comprehensibility of (various aspects of) UML and ER diagrams – e.g., [20]–[24] – or process models [25]–[28]. Although *understandability* is a popular construct of study, it has been argued that there is little agreement on how this is to be measured. Indeed, in their survey, Houy et al. [29] find variability in how understandability is operationalized in the literature. The concept of intuitiveness, as a specialization of understandability, is less frequently being focused on explicitly as in work by Jošt et al., for example, where the *intuitive understandability* of various methods for modeling processes are empirically compared [30].

Work that relates to understanding the comprehensibility of goal models specifically is more limited. Horkoff et al. evaluate an interactive evaluation technique for goal models [31]. The way various concepts within goal models are visualized has also been the matter of investigation and empirical evaluation. Moody et al. offer an assessment of the *i** visual syntax based on established rules ("Physics of Notations") [32]. An empirical analysis was followed by Caire et al. [33] in which experimental participants evaluate visualization choices of the language's primitives. Elsewhere, Hadar et al. [34] compare goal diagrams with use case diagrams on a variety of user tasks. Measures include text-model mapping, model reading (extracting information from the model), and model modification (performing targeted modifications to models). Carvallo and Franch have also studied, in the context of a case study, how non-technical stakeholders performed in developing strategic dependency *i** diagrams [17].

Compared to the above, our work is more targeted to a specific construct of goal models, that is contribution links. In earlier work on the subject [18] we set out to investigate the intuitiveness of the rules in Table I. In that experiment, we presented to experimental participants a series of contribution links each connecting two goals in which the satisfaction value of the origin is know. As we also propose here, we operationalized intuitiveness by asking participants what their "hunch" is with respect to the satisfaction of the destination goal and comparing their input to the authoritative one of Table I. Among our findings were that rules involving positive labels and goal satisfaction are more intuitive to ones with negative labels and goal denial.

We also endeavored to compare the quantitative rules of Table II [35]. In that work we simply presented to participants hierarchies of soft-goals with known satisfaction values at the leaf level and asked them to choose the satisfaction of the root goal from a set of four values, each representing one of the possibilities of Table II. We found that the serial-parallel method was not preferred while the most preferred depended on whether the contribution weights added up to 1.0, in which case a linear interpretation was evoked. In general, our fundamental null hypothesis that the answers would be uniformly random was rejected, indicating that more research should be done on the matter.

Finally, in a different effort [19] and in a vain somewhat similar to that of Caire et al. [33] we focused on the visualization of contribution measures that is alternative to diagrammatic. We specifically employed bar-charts, pie-charts and tree-maps to represent quantitative goal diagrams such as those of Figure 1 – following again the "linear" interpretation. Exactly as we propose here, we presented users with decision problems and asked them to pick the optimal alternative using each of the visualizations under comparison. We found that the combination of pie-charts and bar-charts lead to more accurate identification of the optimal alternative and that diagrams were not better in none of the tests or measures.

The difference of the above effort [19] and the current work is that, while in that paper the semantics are assumed and the visualization is in question, in the study proposed here, both the visualization and its meaning are under comparison. The result can thus be interpretable at either level. For the future, we are interested in exploring theoretical and methodological approaches through which these two aspects can be separately evaluated. The endeavour is not a simple one, as understanding of any communication of a concept can be argued to be affected by the way it is communicated – through words, visualizations or other methods. Thus it may prove difficult to measure comprehension of a concept as a "pure" abstraction. Such a problematic demonstrates how empirical investigation, even at the conceptualization stage, forces us to think more deeply into the substance of the process of conceptual modeling and the nature of its artifacts.

References

[1] E. S. K. Yu, "Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering," in *Proceedings of the 3rd IEEE International Symposium on Requirements Engineering (RE'97)*, Annapolis, MD, 1997, pp. 226–235.

[2] D. Amyot and G. Mussbacher, "User Requirements Notation: The First Ten Years, The Next Ten Years (Invited Paper)," *Journal of Software (JSW)*, vol. 6, no. 5, pp. 747–768, 2011.

[3] A. Dardenne, A. van Lamsweerde, and S. Fickas, "Goal-Directed Requirements Acquisition," *Science of Computer Programming*, vol. 20, no. 1-2, pp. 3–50, 1993.

[4] J. Mylopoulos, L. Chung, S. Liao, H. Wang, and E. Yu, "Exploring Alternatives During Requirements Analysis," *IEEE Software*, vol. 18, no. 1, pp. 92–96, 2001.

[5] S. Liaskos, S. M. Khan, M. Soutchanski, and J. Mylopoulos, "Modeling and Reasoning with Decision-Theoretic Goals," in *Proceedings of the 32th International Conference on Conceptual Modeling, (ER'13)*, Hong-Kong, China, 2013, pp. 19–32.

[6] S. Liaskos, S. McIlraith, S. Sohrabi, and J. Mylopoulos, "Representing and reasoning about preferences in requirements engineering," *Requirements Engineering Journal (REJ)*, vol. 16, no. 3, pp. 227–249, 2011.

[7] D. Amyot, S. Ghanavati, J. Horkoff, G. Mussbacher, L. Peyton, and E. S. K. Yu, "Evaluating goal models within the goal-oriented requirement language," *International Journal of Intelligent Systems*, vol. 25, no. 8, pp. 841–877, 2010.

[8] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani, "Reasoning with Goal Models," in *Proceedings of the 21st International Conference on Conceptual Modeling (ER'02)*, London, UK, 2002, pp. 167–181.

[9] J. Horkoff and E. Yu, "Comparison and evaluation of goal-oriented satisfaction analysis techniques," *Requirements Engineering (REJ)*, pp. 1–24, 2011.

[10] E. S. Yu, "GRL - Goal-oriented Requirement Language." [Online]. Available: http://www.cs.toronto.edu/km/GRL/

[11] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani, *Formal Reasoning Techniques for Goal Models*. Springer Berlin Heidelberg, 2003, pp. 1–20. [Online]. Available: https://doi.org/10.1007/978-3-540-39733-5_1

[12] S. Liaskos, R. Jalman, and J. Aranda, "On Eliciting Preference and Contribution Measures in Goal Models," in *Proceedings of the 20th International Requirements Engineering Conference (RE'12)*, Chicago, IL, 2012, pp. 221–230.

[13] T. L. Saaty, "Decision making with the analytic hierarchy process," *International Journal of Services Sciences (IJSSCI)*, vol. 1, no. 1, pp. 83 – 98, 2008.

[14] Y. R. Preece, H. Sharp, and Jennifer, *Interaction Design: beyond human-computer interaction*. Wiley, 2011.

[15] J. Nielsen, "Ten Usability Heuristics." [Online]. Available: https://tfa.stanford.edu/download/TenUsabilityHeuristics.pdf

[16] J. Mylopoulos., "Conceptual Modeling and Telos," in *Conceptual Modelling, Databases and CASE: An Integrated View of Information Systems Development*. Wiley, 1992.

[17] J. P. Carvallo and X. Franch, "An empirical study on the use of i* by non-technical stakeholders: the case of strategic dependency diagrams," *Requirements Engineering (REJ)*, pp. 1–27, 2018.

[18] S. Liaskos, A. Ronse, and M. Zhian, "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)*, 2017, pp. 466–471. [Online]. Available: http://www.yorku.ca/liaskos/Docs/ESEM17.pdf

[19] S. Liaskos, T. Dundjerovic, and G. Gabriel, "Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18)*, Pau, France, 2018, pp. 1272–1281. [Online]. Available: http://www.yorku.ca/liaskos/Papers/SAC2018/Visualizations/SAC2018.pdf

[20] J. A. Cruz-Lemus, M. Genero, M. E. Manso, S. Morasca, and M. Piattini, "Assessing the understandability of UML statechart diagrams with composite states—A family of empirical studies," *Empirical Software Engineering*, vol. 14, no. 6, pp. 685–719, 2009.

[21] H. C. Purchase, R. Welland, M. McGill, and L. Colpoys, "Comprehension of diagram syntax: an empirical study of entity relationship notations," *International Journal of Human-Computer Studies*, vol. 61, no. 2, pp. 187–203, 2004.

[22] P. Shoval and I. Frumermann, "OO and EER Conceptual Schemas: A Comparison of User Comprehension," *Journal of Database Management (JDM)*, vol. 5, no. 4, pp. 28–38, 1994.

[23] A. De Lucia, C. Gravino, R. Oliveto, and G. Tortora, "Data model comprehension an empirical comparison of ER and UML class diagrams," *Proceedings of the 16th IEEE International Conference on Program Comprehension (ICPC 2008)*, pp. 93–102, 2008.

[24] M. Genero, G. Poels, and M. Piattini, "Defining and validating metrics for assessing the understandability of entity-relationship diagrams," *Data and Knowledge Engineering*, vol. 64, no. 3, pp. 534–557, 2008.

[25] D. Q. Birkmeier, S. Klockner, and S. Overhage, "An Empirical Comparison of the Usability of BPMN and UML Activity Diagrams for Business Users," in *Proceedings of the 18th European Conference on Information Systems (ECIS'10)*, 2010, pp. 51–62.

[26] K. Figl, J. Recker, and J. Mendling, "A study on the effects of routing symbol design on process model comprehension," *Decision Support Systems*, vol. 54, no. 2, pp. 1104–1118, 2013.

[27] K. Figl and R. Laue, "Cognitive Complexity in Business Process Modeling," in *Proceedings of the 23rd International Conference on Advanced Information Systems Engineering (CAiSE 2011) London, UK, June 20-24, 2011. Proceedings*, 2011, pp. 452–466.

[28] J. Mendling and M. Strembeck, "Influence Factors of Understanding Business Process Models," *11th International Conference on Business Information Systems*, pp. 142–153, 2008.

[29] C. Houy, P. Fettke, and P. Loos, “Understanding understandability of conceptual models - What are we actually talking about?” in *Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012)*, vol. (LNCS 7532), 2012, pp. 64–77.

[30] G. Jošt, J. Huber, M. Heričko, and G. Polančič, “An empirical investigation of intuitive understandability of process diagrams,” *Computer Standards and Interfaces*, vol. 48, pp. 90–111, 2016.

[31] J. Horkoff and E. S. K. Yu, “Interactive goal model analysis for early requirements engineering,” *Requirements Engineering*, vol. 21, no. 1, pp. 29–61, 2016.

[32] D. L. Moody, P. Heymans, and R. Matulevičius, “Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation,” *Requirements Engineering*, vol. 15, no. 2, pp. 141–175, 2010.

[33] P. Caire, N. Genon, P. Heymans, and D. L. Moody, “Visual notation design 2.0: Towards user comprehensible requirements engineering notations,” in *Proceedigns of the 21st IEEE International Requirements Engineering Conference (RE’13)*, jul 2013, pp. 115–124.

[34] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi, “Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments,” *Information and Software Technology*, vol. 55, no. 10, pp. 1823–1843, 2013.

[35] N. Alothman, M. Zhian, and S. Liaskos, “User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment,” in *Proceedings of the 36th International Conference on Conceptual Modeling (ER’17)*, 2017, pp. 451–465. [Online]. Available: http://www.yorku.ca/liaskos/Docs/ER17.pdf