Factors affecting comprehension of contribution links in goal models: an experiment.

Sotirios Liaskos and Wisal Tambosi

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

Abstract. Goal models have long been regarded to be useful instruments for visualizing and analysing decision problems. Key to using goal models for the purpose is the concept of satisfaction contribution between goals. Several proposals have been offered in the literature for representing contributions and performing inferences therewith. Theoretical arguments and demonstrative examples are typically used to support the usefulness and soundness of such proposals. However, the degree to which users of goal models intuitively understand the meaning of a specific contribution representation and use it for making valid inferences constitutes an additional measure of the appropriateness of the representation. We report on an experimental study to compare the intuitiveness of two alternative contribution representation approaches via measuring the degree to which untrained users perform inferences compliant with the semantics defined by the language designers. We further explore the role of individual differences such as cognitive style and attitude and ability with arithmetic in establishing and applying the right semantics. We find significant differences between the representations under comparison as well as effects of various qualities and levels with regards to individual factors. The results inspire further research on the specific matter of contribution links and support the overall soundness and operationalizability of the intuitiveness construct.

Keywords: Conceptual Modelling · Goal Models · Model Comprehension · Experimental Study.

1 Introduction

For more than two decades, goal models [39,4] have been extensively studied as an instrument for capturing and communicating intentional structures for a variety of purposes within information technology. One of the strengths of such models is their ability to represent alternative ways by which stakeholder goals can be materialized into design solutions [34,26,27]. Using goal models business/systems analysts can reason about and communicate the advantages and disadvantages of alternative solutions with respect to their impact to higher level business objectives. Multiple proposals for doing such analysis have been proposed in the literature [3,15,26,27] ([21] for a survey).

To make such analysis possible, goal models employ a concept commonly referred to as *contribution* to represent how satisfaction of one goal affects the satisfaction of another. There is variety with regards to how different goal modelling frameworks treat the representation and meaning of contributions. The traditional approach for representing contributions is through symbolic labels (e.g."+', "–") [39,15,20] or words ("help","break") [9] expressing the quality (positive or negative) and the size of contribution in high-level terms. The use of numeric values in various ways has also been proposed [25,30,3], whereby, e.g., sign and absolute value are used to represent quality and size of contribution. The approaches vary with regards to both representation and underlying semantics. Theoretical analyses and demonstrations are usually employed to support the soundness and usefulness of each approach. However, an additional indication of the quality of the chosen representation and semantics could be the extent to which untrained users of the model can *intuitively* understand the meaning of the representation and use it to make inferences in a way that complies with the semantics intended by the modelling language designers.

In this paper, we experimentally explore the intuitiveness of two choices for representing contribution links in goal models, one symbolic and one numeric. At the core of the experiment, a series of decision problems modelled in either of the two ways are presented to untrained users who are asked to use the contributions to perform inferences and make decisions. We measure the extent to which their inferences comply with the semantics of each representation. We further explore how individual differences pertaining to cognitive style, attitude and ability with mathematics and mental arithmetic as well as overall working approach taken by the participants affect the degree of success in performing compliant inferences. Among other things, we find that numeric models evoke much more compliant responses, especially among participants who claim to have followed a methodical rather than an intuitive working approach.

The rest of the paper is organized as follows. In Section 2 we offer background on goal models, contribution links and their semantics as well as the concept of intuitiveness and individual differences that may affect its manifestation. In Sections 3 and 4 we describe the experimental design and the results and in Sections 5 and 6 we review some of the related work and offer concluding remarks.

2 Background

2.1 Goal Models and Contribution links

The type of goal modelling notation we use in this research is akin to the *i** family of goal modelling notations [39,4]. Two examples can be seen in Figure 1. The oval- and cloud-shaped nodes represent actor *goals* (states of the world the actor wants to hold in the future), the ovals describing *hard-goals* and the cloud-shaped ones *soft-goals*. As per their standard meaning [39], soft-goals – as opposed to hard-goals – do not have a precise satisfiability criterion. Further, the goal models we study follow a specific structural pattern. Specifically, using *means-ends* and *decomposition* links, hard-goals form a decomposition that shows

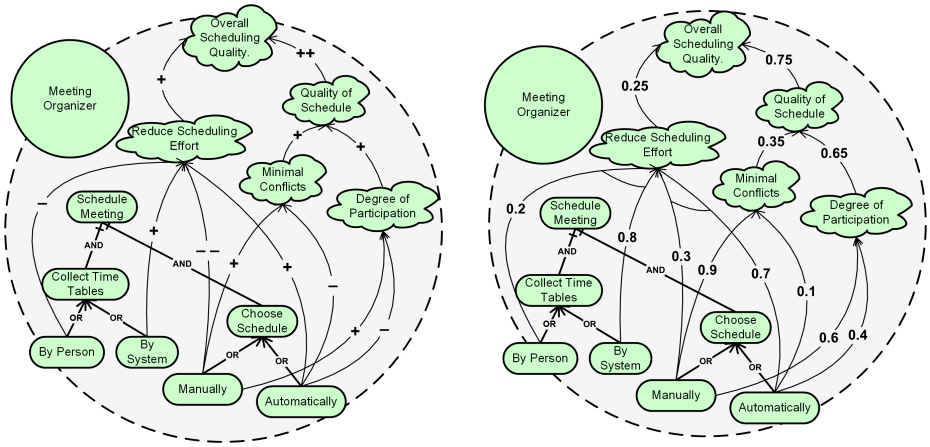

Fig. 1. Goal models with symbolic (left) and numeric (right) contribution links.

how different subsets of low-level goals can enable the achievement of the top-level hard-goal. Soft-goals are recipients of *contribution links*, the curved directed lines. Such links can originate from hard-goals or other soft-goals.

A contribution link from goal A to soft-goal B expresses the hypothesis that evidence of satisfaction or denial of goal A has an effect to our belief about the satisfaction or denial of soft-goal B. The exact quality (positive or negative) and level of effect is expressed using a label on top of the contribution link. The literature offers several proposals for what could be used as a label and what it would mean. The original approach [39,15,4] is to use symbols "+", "++", "−" and "−−" denoting respectively various levels of positive and negative contribution. As of iStar 2.0 [9] words are used ("help", "make", "hurt" and "break") in place of symbols. An alternative approach to symbols and words is numbers: a numerical value in the interval [0.0,1.0] [15,25] or [-100,+100] [3], describes the level of contribution of A to B. Of these various labelling options, the two that are of particular interest here can be seen in Figure 1. They are henceforth referred to as the *symbolic* and the *numeric* representation (mode).

Even without describing the meaning of the contribution links in any more precision, the models in the figure can already be used for performing useful *inferences*. Focussing on the symbolic model on the left side of Figure 1, a user who is only minimally informed to the specifics of the notation and has no knowledge of the precise semantics of "+" and "−−", can probably infer that the goal *(Choose Schedule) Automatically* is preferable to goal *(Choose Schedule) Manually* when we are interested in the goal *Reduce Scheduling Effort*. It is easy to see however that more complex inferences are not possible without an appeal to more formal and precise semantics. Such precise semantics unambiguously define a way for performing inferences. In the absence of such semantics, i.e., without more information about what the labels precisely mean and how they are to be used, in neither model of Figure 1 is it easy to confidently infer optimal decisions

vis-à-vis the root goal *Overall Scheduling Quality*. Various such semantics have been proposed in the literature with both ontological motivation (to clarify what contributions really mean, e.g. [16]) and operational motivation (to suggest how contributions can be used, e.g. [25]). In our study, we pick two proposals of the latter kind, one for each of the representation modes of Figure 1.

The semantic framework for symbolic contributions we consider is due to Giorgini et al. [15]. According to that framework each goal is associated with two variables, each measuring satisfaction and denial of the goal respectively. The variables 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, we may have partial evidence of satisfaction and no evidence of denial for one goal (denoted {**PS,ND**}) and partial evidence of denial and full evidence of satisfaction for another goal ({**FS,PD**}); the inconsistency being perfectly acceptable here and actually one of the strengths of the framework. Given a symbolic contribution link as described thus far, a set of rules, seen in Table 1, defines completely what the satisfaction and denial value of the destination of the link is, given the type of the label ("+","++", etc.) and the corresponding satisfaction and denial values of the origin goal. No evidence (**NS** or **ND**) in the origin is propagated as-is independent of label. Multiple incoming links are treated following a precise evidence maximization principle.

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

Table 1. Symbolic Contribution Semantics

While Giorgini et al. offer an equally comprehensive numeric version of their satisfaction propagation framework we here focus on one used (directly or by implication) by Maiden et al. [30] and Liaskos et al. [25], following the same logic as the one followed by the Unified Requirements Notation (URN) [3]. According to this interpretation each goal has a unique satisfaction value in the real interval [0.0,1.0]. The numeric label on the contribution link represents the share of influence of the satisfaction of the origin goal to the satisfaction of the destination goal. Thus, when a soft-goal is targeted by one or more contribution links, its satisfaction is a linear combination of the satisfaction values of the origin goals weighted by the labels of the corresponding contribution links, as in:

$$s(g) = \sum_{g' \in O_g} \{s(g') \times w(g', g)\} \quad (1)$$

where g is the soft-goal targeted by the links, O_g the set of goals g' from which the contribution links originate, $w(g', g)$ the numeric weights of those links, and $s(g)$ the satisfaction value of a goal g.

2.2 Intuitiveness and Individual Differences

The intent of a developer of visualized conceptual models like the above box-and-line goal models is to evoke a *mental model* of how the visualization is

supposed to be understood and used to make inferences about the domain. Our research question here is whether and to what degree the mental model that is actually evoked within the reader's mind is indeed consistent with the designers' intent, hence promoting "correct" inferences. We use the term *intuitiveness* to furthermore refer to attainment of such consistency with limited or no training. The intuitiveness construct is akin to the concept of *semantic transparency* as per Moody's framework for principled visual design of modeling languages [32]: an intuitive visualized conceptual modelling language is one that allows its users recognize and understand the meaning of the language's constructs based on the visual appearance of the constructs, thus without reference to additional training or explanatory material.

Moreover, when users of a modelling notation are asked to guess the meaning of shapes/symbols and perform inferences therewith, we can hypothesise that *individual differences* in terms of skills, attitudes and styles may affect their choices. One question is whether users attempt to develop a complete and precise theory of how the notation works and make conscious inferences with it or make rough gut-feeling ones based on intuition. A construct that attempts to formalize this distinction is *cognitive style* [1]. According to that construct the approach that decision makers take in solving a judgement problem lies in a *cognitive continuum* [18] between analytical and intuitive cognitive work. While the former describes conscious, controlled, systematic, detail-oriented work towards making an inference, the latter describes quick, approximate, holistic, synthetic and less conscious approach. While Hammond et al. support that cognitive style is largely induced by the task at hand [18], Hayes et al. have shown that decision makers may have a tendency towards one or the other extreme as a personality trait and have developed the CSI (Cognitive Style Index) to measure it [1]. At the same time, simple ability and comfort with mental arithmetic can be a predictor of successful performance of symbol-intensive inferences within a model. Likewise, *math anxiety*, i.e. the presence of feelings of fear, tension, and apprehension with mathematics [19], may affect both how the mathematical/symbolic (e.g. contribution labels) are interpreted and used.

3 Experimental Design

Overview and Research Questions. The goals of our experimental study are to (a) compare the intuitiveness of alternative contribution link representations in the context of assessing optimal decisions within goal models and (b) assess the role of individual differences to the enablement of intuitiveness in the said task. Specifically, the experiment has a confirmatory and an exploratory aspect. We first want to compare the two modes of representation, symbolic (Figure 1 left, Table 1) vs. numeric (Figure 1 right, Equation 1), with regards to their intuitiveness, testing the hypothesis that numeric models are bound to be more intuitive for the purpose of detecting optimal solutions (**RQ1**). The hypothesis is based on the belief that the specific numeric representation utilizes participants' familiarity with numbers and proportions, commonly used in their

daily lives. We further want to explore whether individual differences and ways of working, specifically ability and attitude towards math, cognitive style as well as followed approach, affect intuitiveness (**RQ2**). In the absence of earlier experience, no explicit hypotheses are made with regards to RQ2. The experimental design is an extension/revision of an earlier one presented elsewhere [29].

Constructs and Measures. Our central construct is intuitiveness as discussed above. To measure it, we expose experimental participants to a set of models and ask them to perform inferences based on the information in the model. The participants have only basic awareness of the language and the abstract meaning of its constructs but no knowledge of precise semantics. Intuitiveness is measured primarily via *accuracy* of the participant inferences, i.e., the number of inferences that match the ones that the language semantics dictate. Wherever applicable, we also measure *efficiency*, which is the number of accurate (matching) responses divided by the time it took to make the necessary inferences as well as self-reported *confidence* levels of the method followed to make the inferences (*method confidence*) as well as confidence in the inferences themselves (*response confidence*).

With regards to individual difference factors, we administer the 38-point CSI (Cognitive Style Index) [1] to measure cognitive style (*CSI Score*) and the 9-point AMAS (Abbreviated Math Anxiety Scale) [19] to measure math anxiety (*AMAS Level*). We further measure *ability with arithmetic* using a series of custom non-standard exercises in mental arithmetic. We attempted various types and scoring methods for these. The ones that turn out to have some effect, as discussed below, consist of direct multiplication, scored in [0..10] though an exponentially decaying function of the distance between participant response and correct answer, comparisons of two two-number products and comparisons of two linear combinations each containing two terms. Finally the working *approach* that participants followed, between "using their intuition" and "following a specific method" was captured through self-reporting.

Experimental Units. To construct our experimental instruments we first develop a number of goal models. Two (2) sets of models are developed: symbolic and numeric, each containing only the corresponding type of contribution links. All models consist of one (1) OR-decomposition of hard-goals and an hierarchy of soft-goals to act as criteria for choosing the optimal choice within the OR-decomposition. The soft-goal hierarchy has a unique root goal (such as "Overall Scheduling Quality" of Figure 1) and the contribution labels are chosen such that one of the alternatives of the OR-decomposition is optimal compared to the others, with respect to the top goal. The optimal is calculated by evaluating the impact of full satisfaction of each of the children of the OR-decomposition to the satisfaction of the root soft-goal when the satisfaction values of all other decomposition children are set to **N** or zero, and then identifying the child that results to the maximum such satisfaction. The exact mechanics depend on the type of model and the corresponding semantics. Consider, for example, the *Choose Schedule* decomposition of Figure 1. To evaluate the impact of alternative *Manually* in the left model of Figure 1 we assign it satisfaction values {**FS**,**ND**} while

assuming *Automatically* stays {**NS**,**ND**}. Similarly, for the numeric model on the right we set *s(Manually) = 1* and *s(Automatically) = 0*. We then recursively apply the propagation rules of Table 1, or, respectively, Equation 1 for numeric models, in order to evaluate the satisfaction labels of the higher level goals up to the root soft-goal which is the goal of interest.

Model Sampling. We developed the models used for the instrument by picking a goal structure and populating the contribution links with random contribution labels such that the optimal alternative has a fixed distance from the second optimal one, as measured by the satisfaction each induces to the root soft-goal. This is aimed at allowing sufficient difference between the best and second best to allow for some intuitive detection, but not too obviously.

Calculating the distance from best to second best alternative is straightforward in the case of numeric models: the choice of each alternative will result in a number representing the satisfaction value of the root soft-goal for that alternative; we simply ensure that the largest value is about 0.4 higher than the second largest. For the symbolic models, however, the comparison is less straightforward due to the presence of both satisfaction and denial values. Thus, to allow for comparisons, we aggregate the two values into one. To do so we firstly associate qualitative satisfaction labels **N**, **P**, **F** with numeric values 0,1,2, respectively. Let then $sat(g)$ and $den(g)$ be the resulting numeric satisfaction and denial values for goal g. The aggregated satisfaction value is then $sat(g) - den(g)$ which is an integer in [-2,2]. For example, 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**}, $sat(g_2) - den(g_2) = 2 - 0 = 2$. Given this aggregation procedure, we demand that our sample models have a distance of 2 satisfaction levels. For example, a label configuration in which the best alternative makes the root soft-goal {**FS**,**ND**}, hence aggregated value $2-0 = 2$, and the second best makes the root soft-goal {**PS**,**PD**}, hence aggregated value $1-1 = 0$, qualifies as $2-0 = 2$. To see why this distance matches the one chosen for the numeric models for a fair comparison, observe first that the maximum distance between alternatives in the symbolic case in terms of aggregated value is 4 ({**FS**,**ND**} versus {**NS**,**FD**}). The distance we demanded in symbolic models is 2, thus half of that space. Observing now that the corresponding maximum distance in numeric models is 1.0, it follows that half-space-size distance would be 0.5. However we end up with 0.4, slightly biasing against numeric models, as for some of our structures we fail to find label configurations yielding 0.5 distance.

Instrument and Tasks. For the experimental instrument we develop a total of six (6) model structures, representing decision problems within three (3) domains: Choosing an Apartment, Choosing a Course, and Choosing a Means of Transportation. Thus, two (2) structures are dedicated to each domain, a smaller one with two alternatives and a larger one with three alternatives. For each of the six structures two sets of labels (henceforth: labelsets) are sampled in either of the two frameworks (symbolic vs. numeric). In all, two sets of (3 domains)$\times$(2 sizes)$\times$(2 labelsets) = 12 distinct goal models are constructed and placed in two separate instruments, the symbolic and the numeric.

Each instrument is then organized as follows. Participants are offered two video presentations introducing them to the concepts of decision alternatives and criteria, as well as goal models and the high-level meaning of either type (depending on instrument) of contribution links. Care is taken so that: (a) the videos are as much as possible identical to each other (e.g. use of same examples and points, about same length, same narrator, same visuals etc.), (b) the videos do not prescribe any exact method for interpreting satisfaction propagation that would allude to specific semantics. Subsequently, participants are sequentially presented with the goal models and are asked to enter which of the two or three alternatives they think is optimal. In the end, they are asked if they used a specific method in making their decision, and what that method is, or whether they used their intuition. The CSI, AMAS questionnaires and math ability test precede the aforementioned tasks. We note that midway in the data collection process, the instrument underwent the following revisions: (a) the math ability test was changed and moved to the end and (b) two questions asking for the participants' confidence in their responses and method followed were added.

Participants. Participation is sought from two sources: (a) undergraduate students of the School of Information Technology, York University, attending a human computer interaction course, and (b) Mechanical Turk (MT) participants with a US college degree. We argue in support of these choices below.

4 Results

Sample. A total of 102 participants are included in the analysis: 27 students (21 males and 6 females) and 75 MT participants (41 males and 34 females). The sample predominantly consists of STEM (Science, Technology, Engineering, Mathematics – 49 total) and Business/Economics (22) students/graduates, but also has a mix of Social Science, Humanities, Arts and other backgrounds (31). Their CSI scores are slightly skewed towards the analytical side – 61 above (analytical) and 41 below (intuitive) population average. Of the AMAS scores, 44 are above (more anxious) and 58 are below (less anxious) population average.

Accuracy Analysis. Accuracy is measured as the raw number (out of 12) of correct (wrt. semantics) choices of optimal alternative. To explore accuracy we first attempt to fit a linear model [38] including representation (numeric vs. symbolic), AMAS Level, CSI Score, and approach as main effects, ignoring interactions for the moment. Most factors seem to offer statistically significant or near-significant results: representation ($F(1,97) = 72.2, p < 0.001$, Cohen $d = 1.51$ – numeric more accurate than symbolic), AMAS Score ($F(1,97) = 5.7, p < 0.05$, $d = 0.33$ – the lower the more the accuracy) and working approach ($F(1,97) = 5.6, p < 0.05$, min robust $d = 0.39$ – methodical approach more accurate than intuitive approach). The representation effect is very large and the rest of the effects are small to medium by Cohen's d. Thus, those with below average AMAS level (less anxious) score 0.96 more correct questions than those above average. Finally, accuracy is the only measure in which a certain type of mathematical ability tests, described earlier, seem to have a marginally

statistically significant effect ($p < 0.025$ tested as a lone factor in a separate model): 2.4 more points (out of the 12) in those arithmetic tests results in 1 more correct response in the decision exercises. A small CSI effect detected presents increased Type I error probability and does not emerge in robust tests; it is, thus, dismissed.

Extending the model with interactions we observe that working approach strongly interacts with representation. Specifically, when participants work methodically (by their declaration), that seems to significantly improve their accuracy (3.4 out of 12 more correct answers) but only in numeric models ($F(1, 91) = 6.7, p < 0.05$, $d = 1.38$). Seeing this through a simple effects analysis, whereby we fix approach to a value and explore the effect of representation to accuracy, the representation effect is only present when participants worked methodically – about 4/5 (symbolic group) and 3/4 (numeric group) of the participants.

Efficiency Analysis. Efficiency, operationalized as the ratio of accuracy over total response time, is considered only for the 27 student sample, where response time can be reliably measured; the 75 MT participants are not invigilated thus their exclusive and uninterrupted focus on the experimental tasks cannot be guaranteed. Representation, CSI level, AMAS level and math ability and their interactions are explored. Approach is not considered due to it being highly unbalanced. Representation appears to have a very strong effect to efficiency (Yuen's $t(9.41) = 3.8, p < 0.01$, min robust $d = 0.93$) with a gain of 3.07 correct answers per minute in numeric models versus symbolic ones. However, no other effect or interaction therewith is observed.

Confidence Analysis. Response confidence and method confidence measurements were introduced to the instrument for the last 45 MT participants only and thus the analysis is based on that sample. They are measured on a 7-point "Likert"-style scale and treated as ratio as per normal practice [36]. We again attempt to explain differences in both measures subject to CSI, AMAS, representation mode and approach. In the result, highly analytical respondents have slightly lower *response confidence* ($F(1, 40) = 4.8, p < 0.05$, $d = 0.42$) as expected [18]. Representation also appears to have a small ($d = 0.23$) effect to response confidence but with higher Type I error chance ($p < 0.1$). Analysis of *method confidence* does not yield notable effects.

Summary and Explanatory Remarks. The results present substantial evidence that the numeric representation according to the linear model of Equation 1 leads to more compliant decision-making inferences by untrained users and faster than the qualitative one of Table 1. We can attribute this to the familiarity that users have with numbers and proportions, on which the numeric model is based, and the lack thereof for symbolic labels. However, the effect emerges (strongly) only when the participants say they work methodically, which we interpret as them developing a deeper and more explicit mental model. It follows that in the symbolic case either the evoked method/model is in strong disagreement with the authoritative one, or the latter is correctly guessed but poorly executed. At the same time, the general lack of correlation between arithmetic ability and accuracy, assuming that our custom instruments have any reliabil-

ity, may indicate that participants in the numeric group do not perform the exact mental calculations as per Equation 1, which would require to strongly utilize their mental arithmetic skills, but base their success on an evoked heuristic/approximation that works as well. Furthermore, counter to our expectation that AMAS Level would affect only the numeric group it seems to affect both groups, implying the possibility that the requirement for either kind of symbolic inference is akin to a mathematical task, in which, in turn, highly math-anxious individuals tend to perform worse. Finally, we fail to observe any notable effect of cognitive style trait to accuracy, efficiency or even approach taken, indicating that the index might not be useful for studying the phenomena at hand, possibly also indicating exploration of alternative cognitive style constructs [11]. However, the strong effect of self-reported approach taken suggests that cognitive style remains relevant when seen as choice of cognitive strategy inspired by the characteristics of the task at hand [18] rather than a trait.

Validity Threats. We briefly address the most important of construct, internal, external and statistical conclusion validity threats. In terms of *construct validity* our fundamental assumption that intuitiveness can be measured by the alignment between participant-supplied and authoritative inferences can be criticised as avoiding examination of what goes on in participants' minds when confronted with an unknown notation. A possible response is pragmatic: the observed substantial effect on representation accuracy and efficiency is immediately usable even when theoretical clarity is pending: numbers seem to "just" be more intuitive for the particular task. A further criticism can be extended to the ad-hoc development of non-standard math ability tests, which, however, took place in the absence of suitable standard instruments – and are not major effects regardless. Two main threats to *internal validity* revolve around the representation factor. On one hand, the "difficulty" of the symbolic models (distance between first and second optimal) is constructed based on an operation of comparing satisfaction and denial values that may be argued to be arbitrary and off-specification (by [15]). However, in our view, insofar as the two representations can be used for the same purpose (comparing alternatives) they cannot be considered incomparable vis-à-vis that purpose. Thus, one still needs to address the question of what ways, other than the ones adopted here, can be considered for fairly constructing absolute preferability distance between satisfaction levels in a two-valued setting. Furthermore, difference in training quality can be argued to work against one of the conditions. Such bias is difficult to measure and control for. We are hoping that our carefully scripted, video-recorded training videos (versus live lectures commonly adopted in similar studies) offer a first line of defence against this threat. Threats to *external validity* concentrate on the choice of participants and models. We first claim that our participants being non-experts and (some of them) students does not harm generalizability. On one hand, there seems to be an implicit desire in the goal modelling community that non-technical stakeholders (users, owners, clients) should be able to use such models. On the other hand, although we could not find research that describes the typical characteristics of either business and systems analysts or their clients,

we cannot assume that they are exclusively of a technical background. We, thus, find that our participants constitute a good sample of the population that may be a user of goal models. Furthermore, the choice of models that we used for the instruments brings unavoidable structural, size and domain commitments. Larger models, for example, may be less advantageous for numeric representations, when the method followed does not scale in terms of cognitive effort. Likewise, the tasks we tested them against (picking an optimal alternative) were very particular. Thus, until research with different models is conducted, generalizations should be carefully done for models and tasks of similar characteristics. As a final note on *statistical conclusion* validity, while we pre-hypothesized the effect of representation format, the rest of the factors and interactions thereof were the result of some statistical model exploration. This exploratory attitude aimed at identifying candidate future research directions rather than firmly confirming hypotheses. Thus, except for the effect of representation, the remaining effects continue to be tentative and subject for further confirmation.

5 Related Work

There are several research efforts dedicated towards exploring the effectiveness of common conceptual modeling notations including UML and ER diagrams [8,35,37,10,14] or process models [5,13,12,31]. Much of the research in the area is based on various *understandability* constructs, though there does not seem to be very strong consensus with regards the definition and exact operationalizations of such constructs [22]. The concept of intuitiveness, as we introduce it here as a dimension of understandability, is less frequently considered explicitly, as in work by Jošt et al., for example, where the *intuitive understandability* of various modeling methods are empirically compared [23].

Work focussing on goal models specifically has also emerged. Notable works are by Horkoff and Yu who devise and evaluate an interactive evaluation technique for goal models [20], by Caire et al. [6] who experimentally assess the success of visualization choices for modelling constructs, by Hadar et al. [17] who compare goal diagrams with use case diagrams on a variety of user tasks and by Carvallo and Franch who studied empirically the development of strategic dependency *i** diagrams by non-technical stakeholders [7].

Compared to these efforts, our research program has been heavily targeted towards a specific construct, i.e., contribution links. In earlier work [28], for example, we attempted an investigation of the qualitative propagation rules of Table 1. Through an experiment of a nature similar to the one described here, we observed, among other things, that positive labels and satisfaction values appear to be more readily understandable than negative labels and denial values. Likewise, we have also compared the various models for quantitative satisfaction propagation including the one used here and three versions of the one proposed by Giorgini et al. [2], to find that there is tendency for participants to follow some models versus others, motivating further research on the subject. Note that in all this work our focus is not the effectiveness of just perceiving information

about contributions, which is what, e.g., Moody et al. [33] attempt to improve, but rather understand how contribution is operationally understood and what reasoning it inspires. In a fashion somewhat more similar to that of Moody, Caire et al. [6], i.e., focussing on perception effectivness, we explored graphical (versus diagrammatic) ways for representing contribution levels and found that simple combinations of pie-graphs and bar-graphs allow for better accuracy [24].

6 Conclusions

We presented an experiment for comparing the intuitiveness of symbolic versus numeric goal models vis-à-vis individual differences and working styles of model users. A number of experimental participants is presented with decision problems formalized in either notation and are asked to identify the optimal alternative, without given much information about the precise meaning of the modelling constructs. Intuitiveness is attained when participant responses accurately match the ones each kind of model prescribes to be correct. We find that numeric models lead participants to more accurate responses when the latter are the result of adopting a specific working method. We further find that mathematics anxiety has a mild negative correlation with performance irrespective of representation. Finally while we fail to observe any notable effect of cognitive style as a trait, we find it to be relevant as a chosen cognitive strategy.

Future work can zero-in on identifying the source of inference errors and inefficiencies through distinguishing between mental model adoption and mental model execution, each being exposed to different sets of biases and influencing factors. For the task, instruments that enhance explanatory analysis need to be devised beyond our black-box technique. Qualitative methods and protocol analysis may prove to be of value. However, rather than just understanding a specialized task within a specific notation, our long-term objective is to develop an empirical perspective and toolset transferable to the study of other important classes of notations, such as business process or entity models.

References

1. Allinson, C.W., Hayes, J.: The Cognitive Style Index: A Measure of Intuition-Analysis For Organizational Research. Journal of Management Studies **33**(1), 119–135 (1996)
2. Alothman, N., Zhian, M., Liaskos, S.: User Perception of Numeric Contribution Semantics for Goal Models: an Exploratory Experiment. In: Proceedings of the 36th International Conference on Conceptual Modeling (ER'17). pp. 451–465 (2017)
3. 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)
4. Amyot, D., Mussbacher, G.: User Requirements Notation: The First Ten Years, The Next Ten Years. Journal of Software (JSW) **6**(5), 747–768 (2011)
5. Birkmeier, D.Q., Klockner, S., Overhage, S.: An Empirical Comparison of the Usability of BPMN and UML Activity Diagrams for Business Users. In: Proceedings of the 18th European Conf. on Information Systems (ECIS'10). pp. 51–62 (2010)

6. 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 (jul 2013)
7. Carvallo, J.P., Franch, X.: An empirical study on the use of i* by non-technical stakeholders: the case of strategic dependency diagrams. Requirements Engineering **24**(1), 1–27 (2018)
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 Engineering **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. De Lucia, A., Gravino, C., Oliveto, R., Tortora, G.: Data model comprehension an empirical comparison of ER and UML class diagrams. In: Proceedigns of the 16th IEEE International Conference on Program Comprehension (ICPC'08). pp. 93–102. Amsterdam, The Netherlands (2008)
11. Epstein, S., Pacini, R., Denes-Raj, V., Heier, H.: Individual differences in intuitive–experiential and analytical–rational thinking styles. Journal of Personality and Social Psychology **71**, 390–405 (08 1996)
12. Figl, K., Laue, R.: Cognitive Complexity in Business Process Modeling. In: Proceedings of the 23rd International Conference on Advanced Information Systems Engineering (CAiSE 2011). pp. 452–466. London,UK (2011)
13. Figl, K., Recker, J., Mendling, J.: A study on the effects of routing symbol design on process model comprehension. Decision Support Systems **54**(2), 1104–1118 (2013)
14. Genero, M., Poels, G., Piattini, M.: Defining and validating metrics for assessing the understandability of entity-relationship diagrams. Data and Knowledge Engineering **64**(3), 534–557 (2008)
15. Giorgini, P., Mylopoulos, J., Nicchiarelli, E., Sebastiani, R.: Reasoning with Goal Models. In: Proceedings of the 21st International Conference on Conceptual Modeling (ER'02). pp. 167–181. London, UK (2002)
16. Guizzardi, R.S., Franch, X., Guizzardi, G., Wieringa, R.: Ontological distinctions between means-end and contribution links in the i* framework. In: Proceedings of the 32nd International Conference on Conceptual Modeling (ER 2013). pp. 463–470. Hong-Kong, China (2013)
17. 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)
18. Hammond, K.R., Hamm, R.M., Grassia, J., Pearson, T.: Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. IEEE Transactions on Systems, Man, and Cybernetics **17**(5), 753–770 (1987)
19. Hopko, D.R., Mahadevan, R., Bare, R.L., Hunt, M.K.: The Abbreviated Math Anxiety Scale (AMAS): Construction, Validity, and Reliability. Assessment **10**(2), 178–182 (2003)
20. Horkoff, J., Yu, E.S.K.: Interactive goal model analysis for early requirements engineering. Requirements Engineering **21**(1), 29–61 (2016)
21. Horkoff, J., Yu, E.S.: Comparison and evaluation of goal-oriented satisfaction analysis techniques. Requirements Engineering (REJ) **18**(3), 1–24 (2011)
22. 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 (2012)

23. Jošt, G., Huber, J., Heričko, M., Polančič, G.: An empirical investigation of intuitive understandability of process diagrams. Computer Standards and Interfaces **48**, 90–111 (2016)
24. Liaskos, S., Dundjerovic, T., Gabriel, G.: Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18). pp. 1272–1281. Pau, France (2018)
25. 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)
26. 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)
27. Liaskos, S., McIlraith, S., Sohrabi, S., Mylopoulos, J.: Representing and reasoning about preferences in requirements engineering. Requirements Engineering Journal (REJ) **16**, 227–249 (2011)
28. Liaskos, S., Ronse, A., Zhian, M.: Assessing the Intuitiveness of Qualitative Contribution Relationships in Goal Models: an Exploratory Experiment. In: Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'17). pp. 466–471. Toronto, Canada (2017)
29. Liaskos, S., Tambosi, W.: Comparing the comprehensibility of numeric versus symbolic contribution labels in goal models: an experimental design. In: Proceedings of the MODELS 2018 Workshop on Human Factors in Modeling (HuFaMo'18). pp. 738–745. Copenhagen, Denmark (2018)
30. Maiden, N.A.M., 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)
31. 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)
32. Moody, D.L.: The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Transactions on Software Engineering **35**(6), 756–779 (nov 2009)
33. Moody, D.L., Heymans, P., Matulevičius, R.: Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation. Requirements Engineering **15**(2), 141–175 (2010)
34. Mylopoulos, J., Chung, L., Liao, S., Wang, H., Yu, E.: Exploring Alternatives During Requirements Analysis. IEEE Software **18**(1), 92–96 (2001)
35. 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)
36. Rosnow, R.L., Rosenthal, R.: Beginning Behavioral Research: A Conceptual Primer. Pearson Prentice Hall, NJ, USA, 6 edn. (2008)
37. Shoval, P., Frumermann, I.: OO and EER Conceptual Schemas: A Comparison of User Comprehension. Journal of Database Management (JDM) **5**(4), 28–38 (1994)
38. Tabachnick, B.G., Fidell, L.S.: Using Multivariate Statistics. Pearson, 6 edn. (2012)
39. 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)