Noname manuscript No.
(will be inserted by the editor)

On the Intuitive Comprehensibility of Contribution Links in Goal Models: An Experimental Study

Sotirios Liaskos

Received: date / Accepted: date

Abstract Goal models have long been considered to be useful tools for representing and analyzing complex decision problems in various stages of the software development lifecycle. Through compactly representing large numbers of alternative solutions to requirements problems and capturing the impact of each solution to desired high-level qualities, they allow identification of optimal choices with respect to specified quality priorities. To allow expression of how solutions affect qualities of interest, a special diagrammatic modeling construct, contribution links, is utilized. A variety of ways have been introduced both to visualize the construct and to assign to it formal semantics in the form of rules for performing diagrammatic inferences. However, there is little evidence that, during actual use, proposed visualizations evoke a way of performing diagrammatic inferences that is consistent with the corresponding formal semantics. We conduct an experimental study aimed at comparing two visualization choices for contribution links, symbolic versus numeric, with respect to their ability to evoke inferences that are consistent with formal semantics proposed for such visualizations. The experiment also explores if individual psychological differences including trait cognitive style, mathematics anxiety, and mental math ability, affect this evocation. Participants are asked to make a series of diagrammatic inferences over two sets of goal models each adopting one of the two competing visualization formats and semantics, symbolic vs. numeric. We measure accuracy, that is, the level to which participant decisions are consistent with the formal semantics proposed for each visualization, and investigate the effect to accuracy of the relevant factors – visualization choice, individual differences, and reasoning method adopted. Findings include that most participants adopt specific inference rules instead of working intuitively, that such rules are more consistent with the formal semantics in numeric models, that the utilization of negative contributions and notions of goal denial may hinter accuracy, and that the individual differences considered do not play an important role in either accuracy or choice of inference method.

School of Information Technology, York University,
4700 Keele St., Toronto, Canada, M3J 1P3
E-mail: liaskos@yorku.ca

Keywords Requirements Engineering · Conceptual Modeling · Goal Modeling · Multi-criteria Decision Making · Empirical Methods

1 Introduction

Goal models have been known to be effective tools for supporting decisions in various stages of the software engineering life-cycle and particularly during requirements analysis [6,20,25,100]. During that process, analysts need to make decisions with regards to which of the possible system functionalities are consistent with higher-level long term organizational and stakeholder objectives. Goal models can support such decisions through representing several possible sets of functionalities of envisioned systems as alternative solutions of AND/OR goal hierarchies and describing the impact of each such alternative solution to the fulfillment of high-level strategic objectives. In this way, concise (include only what is necessary) and complete (do not omit necessary parts) solutions can be identified among a large set of possible such, and evaluated subject to multiple and often conflicting strategic criteria. This feature of goal models makes them a very promising tool for supporting and documenting decisions not only in early requirements [74] but also in low-level software design, configuration and adaptation [55,58,59].

A goal modeling language construct that is central for allowing such analyses is known as the *contribution link*. Contribution links show how satisfaction of one goal, which may represent an option or alternative, affects the satisfaction of another goal, which may model a high-level decision criterion. Complex decision problems can, thus, be modeled as networks of such links, whereby goals representing low level decisions contribute in various ways to the satisfaction of goals representing high-level criteria. Moreover, contribution links drawn between the latter express mutual satisfaction dependencies among criteria, adding detail to the model.

A variety of visual representations and semantics have been proposed for contribution links. Symbols, such as "$+$" and "$-$" [30,40,100] and words such as *"help"* and *"break"* [19] are often used as contribution link annotations to describe both the quality of the contribution, i.e., if it is positive or negative, and its size, i.e., if it is a strong or weak contribution. Numeric annotations, such as "75" or "-0.3" have also been proposed [5,54,67]. Depending on the representation choice, such annotations can be useful for a visual exploration of the decision space, aimed at identification, by human readers of the diagram, of the set of decision options, how well each such option satisfies qualities of interest, and which option is better with respect one or more such qualities. To allow for such visual reasoning to take place consistently between people and across time and situations, explicit formal semantics are required that exactly describe how inferences about contributions and their effects can be made. Thus, many attempts to define contribution link semantics for different kinds of representations have been made [5,30,57,60] – [39] for a related survey – often geared towards enabling automated reasoning about decisions. Nevertheless, despite this wealth of options, choosing the right visualization for contribution links (e.g., symbols, words, or numbers) to accurately describe their semantics is rarely a primary concern. Of particular interest is whether visualization and semantics align with each other in terms

of whether users of the notation can naturally infer the latter (semantics) from the former (visualizations). Such alignment allows model readers and model developers to make consistent diagrammatic inferences, supporting successful communication between the two. In addition, it allows model readers to perform diagrammatic inferences that are consistent with those of automated reasoners, making the output of the latter more visually explainable.

In this paper, we present an experimental study on the *intuitiveness* of visual representations of contribution links vis-à-vis their semantics. We define intuitiveness of conceptual modeling notation constructs to be the ability of notation users to understand the supposed semantics of construct representations without prior explicit training, and through appeal to established meanings and uses for such representations. For example, the use of symbol "$+$" to represent that a contribution is positive is more intuitive than the symbol "@", in that users know from daily experience and without the need for additional instruction that "$+$", as opposed to "@", is associated with addition (e.g. added influence, added value).

In our study, we firstly compare the intuitiveness of two distinct representations of contribution links, namely symbolic, i.e., ones that use symbols, such as "$+$" and "$-$", versus numeric, i.e., ones that use numbers such as 0.6 and 0.25. To perform the intuitiveness measurement, we construct a number of goal models, each consisting of an OR goal decomposition representing a decision with 2 or 3 options and a small network of high-level decision criteria connected through contribution links of either representation format (symbolic or numeric). The semantics of each representation format, which come in the form of satisfaction propagation rules, prescribes which of the 2 or 3 options is optimal. We then invite experimental participants to simply look at the models and identify the optimal without complete prior training to the semantics of contribution links. The participants are split in two groups: one is exposed to models with symbolic and the other to models with numeric contribution links. We measure the accuracy, i.e., the number of times that participants of either group identify the correct optimal, according to semantics.

In a second follow-up exercise, participants asked to perform a slightly different kind of diagrammatic reasoning. We expose them to a series of diagrams displaying a single contribution link connecting two goals, disclosing to the participants the level of satisfaction of the goal that is origin to the contribution link and asking them to identify the satisfaction level of the destination of the link. The representation style of contribution labels and satisfaction levels is again different in each group (symbolic vs. numeric), and the correct answer is defined by the corresponding semantics. We measure how often participants – who are, again, not made aware of the semantics – guess the answer correctly, and compare the two groups in that measure.

In addition to those two main tasks, participants are also asked to describe the method they adopted for solving the decision exercises, and answer questionnaires that elicit their individual differences in terms of their trait cognitive style [2], mathematics anxiety, [37] and ability with mental arithmetic.

With the experiment we aim at answering four main research questions. The first is asking if the two representations (numeric and symbolic) are different with respect to their ability to lead participants to diagrammatic reasoning that is compliant to the associated semantics. We answer this through comparing the accuracy of re-

sponses between groups. The second question is what process participants are adopting to perform diagrammatic reasoning and how compliant or similar this process is with the authoritative one. We explore this through analyzing participant descriptions. The third question asks if individual differences (cognitive style, math anxiety, mental math ability) affect the accuracy of responses in each group – answered through studying the corresponding correlations, and the fourth research question asks if the measured cognitive style affects the choice of diagrammatic inference method.

A key finding is that participants spontaneously adopt a concrete method for performing inferences, which, further, appears to favor numeric representations and semantics. Nevertheless, despite the fact that participants offer solutions compliant to semantics in such models, the rules adopted for arriving at those compliant solutions may be quite different from (yet partially consistent with) the ones prescribed by the semantics. In addition, models involving negative contributions and negative satisfaction (goal denial) were consistently found to evoke inferences that do not comply with semantics. Finally, individual differences are not found to affect accuracy or inference choices in any significant way. Apart from informing future research and goal modeling language design efforts, our results have some immediate practical implications which we present as examples of concrete modeling guidelines.

Our report combines and extends our earlier conference publications of these studies [62,63] with previously unreported work and details including: (a) inclusion of additional data that have been collected since the publication of the above papers that allow for more useful and confident statistical inferences (particularly on negative results pertaining to individual differences), (b) results from experimental tasks previously not presented including a comparison between numeric and symbolic representations in single-link tasks, and analysis of free-form qualitative data, (c) comprehensive presentation of the theoretical baseline, (d) complete details on experimental design, administration, and acquired data with additional visualizations and statistics, and (e) a discussion on design implications.

The paper is organized as follows. Section 2 offers background on goal models, contribution links, and dominant representation and semantics proposals for such. Then, Section 3 describes the notion of intuitiveness, its measurement, and factors that may influence it in detail. Section 4 describes our experimental design, Sections 5 and 6 present the results, and Section 7 discusses general conclusions and design implications, as well as validity threats and limitations. Then, Section 8 discusses related work and Section 9 offers concluding remarks and future work possibilities.

2 Goal Models and Contribution Links

2.1 Goal Models as Decision Support Tools

Goal modeling languages provide constructs for capturing the structure of the intentions of individual and organizational actors. Our work focuses on a particular family of goal modeling languages that are based on *i** [99,100] and predominately the latest iStar 2.0 standard [19] as well as the Goal-oriented Requirement Language (GRL) which is part of the User Requirements Notation standard (URN) [6]. Two alterna-

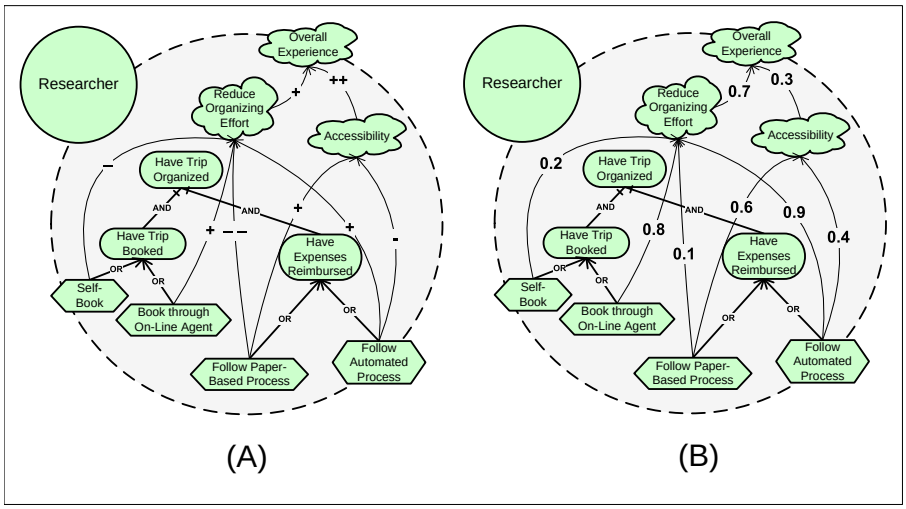

Fig. 1 Goal models featuring the symbolic (A) and numeric (B) approaches to labeling contribution links.

tive graphical representations of a goal model constructed using such languages can be seen in Figures 1(A) and 1(B). These example models present a subset of features of the languages that is interesting for our purposes and are structured in a specific way to support decision exploration.

Focusing on the representation on the left, the model represents the goal structure of actor *Researcher* who wants to have a trip organized for a conference – a case inspired by the running example in the iStar 2.0 guiding document [19]. The oval-shaped elements are *goals* which represent states of the world that *actors* (circular elements) want to achieve, such as for example *Have Trip Organized*. The goals are connected with each other with AND- and OR-decompositions. For an AND-decomposed (resp. OR-decomposed) goal to be considered satisfied, all (resp. one) of its subgoals need(s) to be satisfied. Subgoals can be recursively decomposed to other goals forming and AND/OR tree. At the bottom of such decomposition tree are *tasks* which describe actions that actors need to perform for the fulfillment of parent goals. Some tasks, such as *Follow Automatic Process*, imply the presence of software functions to be executed and as such are indicators of possible software requirements. The root goal of the goal hierarchy can be satisfied by as many subsets of leaf level tasks – henceforth *alternatives* – as the solutions of the AND/OR tree. As such, the goal decomposition implies several possible sets of requirements that can fulfill the main (root) functional goal.

To allow evaluation and comparison of the alternatives, analysts can represent how each of those alternatives supports higher-level strategic objectives. This is represented through *qualities* (also here: *quality goals*) in the diagram – the cloud shaped elements – which are formally defined as attributes for which an actor desires some level of achievement [19], such as, e.g., *Accessibility*.

Qualities do not necessarily have a clear definition, i.e., a precise way to decide when a quality is achieved or not. As such, they are assumed to be satisfied to a certain degree and based on the satisfaction of other goals or qualities for which evidence of satisfaction is more available. This is attained through the use of *contribution links* between goals and qualities and between qualities, which is the focus of this research.

2.2 Contribution Links and their Meaning

We now turn our focus to the notion of contribution links and the various approaches that have been introduced for (a) diagrammatically representing them, and (b) defining their semantics so as to allow consistent reasoning about how satisfaction of one goal affects satisfaction of another. We focus on a two-valued qualitative approach (Section 2.2.2), and a one-valued quantitative approach (Section 2.2.3). This presentation is important for understanding the experimental study we present thereafter, which compares these two approaches.

2.2.1 Contribution links in goal diagrams

Contribution links in goal models represent the idea that satisfaction of one goal or quality has an effect to the satisfaction of some (other) quality. In Figures 1(A) and 1(B) two ways for representing contribution links can be viewed – the diagrams are identical otherwise. In the 1(A) a *symbolic* approach for representing contribution links is presented. Positive symbols, such as "$+$" and "$++$", represent that satisfaction of the origin of the contribution link positively affects satisfaction of the destination of the link. The double sign ("$++$") implies that the effect is somehow of a greater size/impact. The reverse is true for negative symbols such as "$-$" and "$--$", which imply that satisfaction of the origin goal affects negatively the satisfaction of the destination goal in some way. The double sign ("$--$") is, again, used to denote greater impact. Following a *textual* approach (not seen in the figure) we can replace the symbols "$--$","$-$","$+$" and "$++$" with words *"break","hurt","help"* and *"make"*, respectively [19]. The textual labels would have the meaning implied by the words used. The *numeric* approach is to use numbers for labels as in Figure 1(B). In the case depicted, the numbers are from the interval $[0.0, 1.0]$: the higher the number, the higher the contribution.

Irrespective of representation, contribution links, even informally understood as above, can be useful for diagrammatically identifying optimal decisions. For example, in either of the diagrams of Figure 1, if we know that *Reduce Organizing Effort* is an important quality goal, it seems reasonable that the task *Book through On-line Agent* is a better choice for goal *Have Trip Booked* than *Self-Book*. We can assume so through simply intuiting that "+" implies a positive effect and "$-$" a negative one (Figure 1(A)) or that 0.8 implies a larger (positive) effect than 0.2 (Figure 1(B)), based on our prior experience on how such symbols and numbers are interpreted and compared. Subsequently, we make the decision based on which option brings about a comparatively more positive effect to the quality goal of interest.

However, such intuitive inferences may be difficult in larger and more complex models without offering precise semantics both of contribution links and of the notion of goal and quality satisfaction that such links affect. This is particularly true when longer contribution chains need to be traversed, aggregating various contribution links arriving at the same node along the way. For example, it is unclear how one should choose between *Follow Paper-based Process* and *Follow Automated Process* with respect to the top-level goal *Overall Experience*.

2.2.2 A two-valued qualitative framework

To allow more precise and unambiguous reasoning, a variety of definitions for contribution link semantics have been proposed. The original and most expressive semantics for contribution links has been provided by Giorgini et al. [30,31]. According to that framework, each quality goal carries two variables describing its satisfaction status, a *satisfaction* variable and a *denial* variable. Each variable takes a value that describes the level of evidence we possess that the quality is, respectively, satisfied or denied. It is convenient to think about their proposal as offering two options for representing and reasoning about those variables: a *qualitative* and a *quantitative*, represented in their simplest form through symbolic and numeric contribution links as in the diagrams of Figure 1(A) and 1(B), respectively.

The qualitative interpretation assumes that the satisfaction and denial variables take values from the set {**N**,**P**,**F**}, where **F** stands for full evidence, **P** for partial evidence and **N** for no evidence of satisfaction or denial, respectively. The satisfaction/denial status of each quality goal is then described through two such values. For presentation convenience here we appropriately suffix each such value based on whether it represents satisfaction (**S**) or denial (**D**). For example, for a quality we may have full evidence of its satisfaction and no evidence of its denial, hence {**FS**,**ND**} and, for another, partial evidence of satisfaction and full evidence of denial, thus {**PS**,**FD**}. Note that representing conflicting information about the satisfaction status of a quality goal (both satisfied and denied) is perfectly acceptable and one of the features of the framework.

Given this way of representing quality goal satisfaction, contribution links can be seen as mappings from the space of satisfaction and denial values of the origin of the link to the corresponding spaces of the destination of the link. The mapping is defined through a set of *propagation rules*. Different labels decorating the contribution link are associated with different propagation rules. Positive contribution labels $++$, $+$, propagate the labels as they are or with **F** truncated to **P**, respectively. Negative contribution labels $--$, $-$ operate similarly but with the difference that they invert the satisfaction into denial and vice-versa. A list of all possibilities can be seen on Table 1.

The *label propagation algorithm* proposed by the authors [30,31] employs an evidence maximization principle for deciding what the satisfaction and denial value a quality goal should have in the presence of multiple incoming contribution links, as it happens with, e.g., quality *Reduce Organizing Effort* in Figure 1(A). In those cases, the rules are applied for each incoming contribution link, resulting in a set of candidate satisfaction evidence values for each of the satisfaction and denial variables.

Label	Effect	Label	Effect	Label	Effect	Label	Effect
$++$	FS $\rightarrow$ FS PS $\rightarrow$ PS PD $\rightarrow$ PD FD $\rightarrow$ FD	$--$	FS $\rightarrow$ FD PS $\rightarrow$ PD PD $\rightarrow$ PS FD $\rightarrow$ FS	$+$	FS $\rightarrow$ PS PS $\rightarrow$ PS PD $\rightarrow$ PD FD $\rightarrow$ PD	$-$	FS $\rightarrow$ PD PS $\rightarrow$ PD PD $\rightarrow$ PS FD $\rightarrow$ PS

Table 1 Symbolic contribution semantics. The rules in the "Effect" column represent how a value of the origin goal/quality (left hand-side of the arrow) translates into a value of the destination quality (right hand-side of the arrow), when the link is labeled as seen in the "Label" column. Adapted from [31].

Of those, the maximum is selected. For example assume that in Figure 1(A) we are interested in the satisfaction values of *Overall Experience*, when *Reduce Organizing Effort* is {**FS**,**PD**} and *Accessibility* is {**FS**,**ND**}. The candidate satisfaction values are **PS** coming from *Reduce Organizing Effort* and **FS** coming from *Accessibility*. The candidate denial values are, respectively **PD** and **ND**. Hence the values for *Overall Experience* are {**FS**,**PD**}.

Giorgini et al. also present a quantitative version of their label propagation framework [31]. According to this version, both satisfaction and denial values and contribution labels are now numbers as seen in Figure 1(B) – for our purposes we demand them to also be in the interval $[0.0, 1.0]$, though this does not appear to be necessary in the general framework. Instead of an exhaustive list or rules, a generic operator $\otimes$ is used to represent how the origin satisfaction and denial values are combined to produce the corresponding values of the destination. Let g be a quality goal targeted by another quality goal g' using a contribution link with label $w(g', g)$. If $v(g)$ and $v(g')$ are satisfaction or denial values of g and g' respectively, the general form of a propagation rule is $v(g) = v(g') \otimes w(g', g)$. As in the qualitative framework, for label propagation, a maximization of the candidate values is applied in each of the steps.

Interestingly for our purposes, the generic operator can be interpreted in different ways. The default is $p_1 \otimes p_2 =_{def} p_1 \cdot p_2$, i.e., the product of the satisfaction value and the contribution label – the authors call this the *multiplicative* interpretation. Under this interpretation, the numbers constitute probabilities: $v(g), v(g')$ are the probabilities of satisfaction (or denial) of the origin and destination goals, and $w(g', g)$ the conditional probability that g' is satisfied given that g is satisfied. However, other interpretations are suggested by the authors as a side note: the *minimum* interpretation $p_1 \otimes p_2 =_{def} min(p_1, p_2)$ (the one applied in the qualitative framework) and the *serial-parallel* interpretation $p_1 \otimes p_2 =_{def} p_1 \cdot p_2/(p_1 + p_2)$. While in our experiments we consider only the qualitative version of the two-valued framework, the alternative ways by which participants combine values $v(g')$ and $w(g', g)$, is, as we will see, relevant to one of our experimental tasks.

Note that the above constitutes a simplified presentation of the framework described by the authors [31]. Specifically, the original framework allows for contribution labels that propagate only satisfaction or denial values, such as, for example, $++_S$, $-_D$ and 0.7_{-D}. The labels and propagation rules as we describe them here represent the co-existence of satisfaction and denial propagation. For example, $++$ is used as a shorthand for two links, $++_S$ and $++_D$, connecting the same goals. This convention, and, generally, the above treatment of contribution links, is in agreement

with the original proposal [31]. However, a simplification that departs from the original, namely the merging of the satisfaction and denial values to allow evaluation of distances between alternatives, will be necessary for the experimental study and is described in the experimental design section. We stress that our intention here is not to evaluate the corresponding frameworks per se but rather use them as starting points for exploring the relationship between meaning and representation of contribution constructs.

2.2.3 A one-valued quantitative approach

The above proposal is only one option for defining the semantics of contribution links – we will henceforth refer to it as the *label propagation* approach. An alternative approach to the above framework has been proposed independently by Maiden et al. [67] and Liaskos et al. [43], which under assumptions we discuss below, is also compliant with the evaluation approach adopted by URN for evaluating GRL models [5]. In this framework, which is quantitative, the satisfaction status of each quality goal is represented using a single value in the real interval $[0, 1]$. Contribution links are also labeled with real values in $[0, 1]$. Rather than propagation of a label, contributions are understood as the *share of satisfaction* of the destination quality due to the satisfaction of the origin goal or quality that connects through the contribution. Assume then that O_g is the set of goals or qualities g' such that there is a contribution link from g' to a quality goal g, and $w(g', g)$ is the numeric weight of that link. Then the satisfaction $s(g)$ of g is calculated from the satisfaction $s(g')$ of each $g' \in O_g$ as follows:

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

Considering again the diagram of Figure 1(B), with respect to the decision under *Have Expenses Reimbursed*, option *Follow Paper-Based Process* wrt. *Overall Experience* has a value of $0.1 * 0.7 + 0.6 * 0.3 = 0.25$ and, respectively, *Follow Automated Process* has a value of $0.9 * 0.7 + 0.4 * 0.3 = 0.75$.

This framework, thus, directly maps goal models to a family of Analytic Hierarchy Process (AHP) [81] decision problems, in which the quality subgraph plays the role of the criteria, and each OR-decomposition is a separate decision process sharing the same criteria and relative importance thereof. Although this approach is much less expressive than the two-valued one and also imposes structural limitations to the goal models (acyclicity), it has the benefit of an established elicitation technique for the numbers (AHP pair-wise comparisons).

The GRL approach to evaluation of contribution links [5] can be seen as a generalization of the above. In GRL both weights and satisfaction values are defined in $[-100, 100]$, rather than $[0, 1]$ and there is no requirement that the multiple incoming weights add up to a maximum (e.g., 100); rather, the outcome of the weighted summation is truncated, when needed, to fit the above interval. Should we restrict values to $[0, 100]$ and demand that incoming weights add up to 100, the two frameworks propose essentially the same aggregation technique, except for presentation style (a decimal versus a percentage-style number). Thus, while we generally follow

the style proposed by Liaskos et al. [54], under these restrictions our findings can be hypothesized to be applicable to GRL as well.

We will henceforth refer to this general representation and inference approach as the *weighted summations* approach to contrast it with the label propagation approach we discussed in Section 2.2.2. In our experimental study the main comparison is between these two approaches.

3 Intuitiveness: Definition, Measurement, and Influencing Factors

We presented above various approaches for representing contribution relationships between quality goals. As we saw, for each representation style, semantics have been proposed, i.e., rules for deciding how satisfaction of the goal or quality that is origin of such a link affects the satisfaction status of the destination quality. The general question we investigate in this paper is whether these semantics, decided by the designers of the language, are consistent with (henceforth also: *align with*) the semantics that users of the notation naturally assign to these visualizations when using them. In the following, we motivate the study of naturally evoked semantics, and discuss intuitiveness as an empirical construct by which we can understand and, respectively, empirically measure such alignment. We then discuss individual psychological traits that may act as factors that affect the emergence or not of alignment. These are also a subject of investigation in our study.

3.1 The Intuitive Comprehensibility Construct and its Measurement

One of the principal properties of successfully designed diagrammatic representations and constituent visual constructs (boxes, arcs and their labels, etc.) is that they are able to communicate their meaning. In conceptual modeling, this quality of a visual construct has been referred to as *semantic transparency* [70] or, more broadly, *comprehensibility* (or understandability [41]).

To empirically measure comprehensibility of a model, we need to unambiguously describe the concept and establish operational definitions (metrics) [80] thereof. For this purpose, is useful to refer to SEQUAL, a semiotic framework for organizing conceptual model qualities [49,50]. In SEQUAL the notion of *(manual) model activation* is proposed to describe the role of models in guiding human behavior. For example, when providing a business process diagram to a participant or observer of the business process represented in the diagram, the participant will organize their work, answer questions, troubleshoot, make decisions etc., in a way that is consistent with the information they believe that the diagram contains. In other words, users of model representations utilize the information they perceive from the representation in order to perform *inferences* which, in turn, inform their own action.

Model activation allows us to think about comprehensibility as the degree of alignment between, on one hand, users' beliefs about the content of the model, manifested through related inferences they perform and observable consequences thereof, and, on the other hand, the corresponding belief held by: (a) the builders of the specific model, (b) the designers of the conceptual modeling language that was used

to build the model. It follows that if users perform inferences with the model that are incompatible with the modeler's and/or the language designer's expectations, the model has arguably not been comprehended. In other words, the evoked (by users) semantics of the constructs does not align with the prescribed semantics defined by the designers of the language (also, henceforth interchangeably: *authoritative*, *normative* semantics); otherwise the inferences would be compatible. In Figure 1(A) for example, we saw that based on the supposed meaning of contribution links and the "+" and "−−" labels that decorate them, we expect that users of the model will infer that one alternative (e.g., *Book through On-line Agent*) is better than another (*Self-book*) with respect to a specific quality (*Reduce Organizing Effort*). If users, however, consistently make the opposite inference, the designers of the labels and their meaning may need to suspect that comprehension has not taken place and there is misalignment between how they want users to understand the labels and how users actually understand them. Thus, observing the frequency or quality of inconsistent inferences appears to be one way to empirically measure comprehensibility.

Incidents of lack of comprehensibility of a specific model representation can be attributed to a variety of factors, such as, the quality of the model, the circumstances of the inference, the person making the inferences and their familiarity with the state of affairs represented, or the modeling language used. Of particular interest here is the modeling language: we are interested to see whether incomprehensibility is the result of sub-optimal language construct design. When the focus is on the language rather than individual models constructed using the language we use the term *comprehensibility appropriateness* of the language [33,97]. In our case, for example, the meaning of a link decorated with a "+" label may not be comprehended as desired due to either "+" being the wrong symbol for representing the concept *"positive contribution"* or the concept itself being unknown, difficult to comprehend or otherwise problematic. This problem concerns not the model in which the link was observed but the language that was used to build the model and proposes the link as one of its constructs.

As a specialization of the above, *intuitive comprehensibility appropriateness* of a language construct, or, henceforth, *intuitiveness*, refers to the comprehensibility appropriateness of the construct by users who have partial and limited prior exposure to the language. The concept, and the need thereof, can be understood through reference to our every day experience with signs [16]. Human computer interface icons, for example, are preferably designed in a way that they easily convey their meaning and function to users, without demanding the latter having to study or otherwise dedicate time for familiarizing themselves with these meanings [79]. In our case, the use of the + label to denote *negative* contribution would not support the intuitiveness of the notation as it would require unnecessary training and probably be the source of errors and inefficiencies in using the construct in the longer term.

Hence, intuitiveness, as defined above, can serve as a concrete empirical construct for our purpose of describing the level of alignment between prescribed and naturally evoked semantics of contribution link visualizations. Note that, with the term "empirical construct" – not to be confused with language construct which refers to constituents of modeling languages – we refer here to an abstract variable that is meant to be used as an explanatory concept and is, as such, operationalized into a con-

crete metric for empirical measurement [80]. The concept of model activation, offers us an idea for operationalizing intuitiveness: we simply observe the inferences users perform with the contribution links (e.g., how they use them to evaluate decision options) and quantitatively and qualitatively compare them with the inferences that the prescribed semantics would allow. We specifically use the term *accuracy* to describe the concrete quantitative measure of the alignment between observed and prescribed inferences that is based on simply counting the number of times, over a number of similar inference tasks, that the two inferences agree. Higher accuracy would then be an indication of more intuitiveness. The precise metric formulations are discussed in the experimental design section.

3.2 Mental Models

The above way to operationalize intuitiveness (measure agreement between observed and normative inferences) relies on a process of semantics evocation, i.e., the adoption of a way of using contribution links based on observing and interpreting their visualization, by possibly utilizing prior knowledge of the meaning of the visualization. It is natural to ask whether there is any theoretical basis for such a phenomenon, to also allow us to obtain a richer and more confident interpretation of some of our results.

A concept that can serve as such a basis are *mental models* [47,75,77,98]. Mental models have been used in the interaction design literature to describe abstractions that users of interactive artifacts form internally for the purpose of predicting and explaining the behavior of said artifacts [75]. For the purpose of diagrammatic reasoning, a visualization of a modeling construct to which a user is exposed for the first time, such as an arc with a label on it, can be understood to evoke an initial theory on how it is to be used – i.e., how the arc is to be combined with other arcs to make a decision. Hence, a visualization that evokes a mental model that is compliant to the actual reasoning mechanism as intended by the designers (such as using "$-$" instead of "$+$" to represent a negative contribution) can be claimed to preferable. As we will see in our results section, the formation and actuation of mental models will help us qualitatively analyze and interpret participant responses.

3.3 Intuitiveness and Individual Differences

In the above, we motivated the notion of intuitiveness and presented the general empirical method we follow in order to measure and compare the intuitiveness of different contribution link visualizations and semantics. In addition to such comparisons, our study is also concerned with exploring if individual psychological characteristics of those who use the models (i.e., their traits and abilities) affect how they interpret and use contribution links, consequently increasing or decreasing alignment with prescribed semantics. In this study, we are specifically interested in three such characteristics: trait cognitive style, mathematics anxiety, and ability with mental arithmetic. We describe and motivate the relevance of each of these in the following.

A first question is whether users adopt and follow any kind of strategy in order to perform a diagrammatic reasoning task with goal models – such as that of identifying optimal solutions in Figure 1. One can, for instance, conjecture that some users make rough, gut-feeling decisions whose rationale and exact procedure that led to them are difficult to articulate. Other users may develop a concrete procedure which they will consistently apply in all decision making instances. An empirical construct that relates to such a distinction is *cognitive style* [2,35]. According to the theory behind this construct, there is a cognitive continuum between analytic and intuitive cognitive work that can be utilized for the solution of a judgment problem. Analytic processing describes conscious, controlled, systematic and detailed-oriented work, while intuitive processing describes quick, approximate, holistic, synthetic and less conscious approach. Hammond et al. supports that a different cognitive style is adopted based on the nature of the task at hand [35].

However, it has been shown that the tendency to adopt a work approach towards one or the other direction of the continuum can be seen as a measurable personality trait. Allinson and Heyes have developed the Cognitive Style Index (CSI) [2] to measure one's propensity to adopt the former or the latter strategy for solving problems. The CSI is measured through a 38-question survey administered to participants including questions such as *"the best way for me to understand a problem is to break it down into its constituent parts"* and *"I am inclined to scan through reports rather than read them in detail"*, to which respondents must answer if they agree or not. A score is then produced characterizing the propensity of the respondent to adopt analytical or intuitive strategies in the given scenarios and situations. In the two above questions, for instance, an analytical person would, respectively, respond *"agree"* and *"disagree"* and an intuitive person the opposite.

The CSI index has been found to correlate to a variety of occupational, learning, or other decision making and information processing preference and performance measures [7,8,26,95]. The applications of the specific or similar indexes have also been observed in the area of conceptual modeling. Türetken et al. [93], for example, found that participants with low CSI (i.e., intuitively-inclined) performed worse in a model comprehension test than their peers with a higher CSI score. A similar index, OSIVQ [12], was found to affect preferences of representation formats (diagrams, structured text, text) for business process models [27].

Such studies motivate the investigation of the role of different cognitive styles in how conceptual models are read and comprehended. In our study, the specific focus is how participants combine various contribution links in order to make a decision using a goal model. We specifically hypothesize that the intuitively-inclined participants will decide based on an abstract impression of which decision option is associated with the most positive contributions, while the analytically-inclined ones will adopt an algorithm to combine different contribution links based on their assumption of the semantics of those links. We further want to explore, for each competing representation, whether either of the strategies leads to more accurate responses, i.e., responses that are more often aligned with the authoritative ones.

As we discussed, our experiment involves asking participants to perform diagrammatic inferences with models of either symbolic or numeric representations of contribution links. When asked to perform inferences with the numeric models specifically,

participants may feel invited to do so via performing some kind of mathematical operations. We may, hence, hypothesize that users with better ability in mental arithmetic could be more effective in, firstly, guessing the normative way to perform such calculations (weighted summations as we saw in Section 2), and, secondly, performing the calculations correctly. At the same time, users with limited such ability and/or a negative attitude towards numbers, might avoid any processing thereof and resort to intuitive or arbitrary choices. It is hence relevant to our research questions to see if attitudes towards numbers and ability in mental arithmetic affects response accuracy.

One construct related to attitude towards math in general is *math anxiety* [10], which describes the presence of feelings of fear, tension and apprehension of mathematics, resulting, as it has been found, in lower performance in math related tasks [11]. As such, math anxiety can be used as a proxy for math ability, and, as we hypothesize in our case, a measure of resistance to engage in mental arithmetic when dealing with a problem presented in the form of numbers. As with cognitive style, an index for measuring math anxiety has been proposed, namely the 9-point Abbreviated Math Anxiety Scale (AMAS) [37].

In addition to attitude towards math, in our experiment we test ability in mental arithmetic. This is tested through a small number of timed questions whereby participants are invited to perform additions, subtractions, multiplications and divisions, and various combinations thereof, without using calculator and as quickly as possible. Our hypothesis is, again, that users that are more capable in mental arithmetic will be able to respond more accurately in numeric models. We discuss how these tests are designed in more detail in the results section.

A summary of the concepts we discussed above, including a description and, where applicable, a sketch of how they are operationalized according to this study is offered in Table 2.

4 Experimental Design

4.1 Research Questions and Design Approach

The study aims at addressing the following main research questions, organized in two groups:

- **Group 1:** The role of representation in intuitive comprehensibility.
  - **RQ1.1:** Do the two ways by which we represent contribution link labels in diagrammatic goal models, numeric and symbolic, differ in terms of their ability to evoke user inferences that are compliant with their semantics?
  - **RQ1.2:** What process do users choose to follow in order to make inferences with the goal models, when concrete guidance for such is absent? Does it align with the normative process under different representations?
- **Group 2:** The role of individual differences in intuitive comprehensibility.
  - **RQ2.1:** Do individual differences, specifically cognitive style, math anxiety and ability with mental arithmetic affect the ability of users to perform inferences that align with the normative semantics?

Term/Construct	Description	Measurement Approach
Comprehensibility/ Understandability (of model)	Level by which understanding of the model by its readers agrees with the understanding of the model by its creator. (see also: [28,41])	Compare inferences made by users with inferences made by model creators.
Comprehensibility Appropriateness (of modeling language)	Level by which understanding of a modeling construct by its readers agrees with the understanding of the modeling construct by the language creator. (see also: [33,49])	As above over several instantiations of the construct in concrete models.
Model Activation	User activity in accordance to a model. [49]	Observe relevant user action following exposure to the model.
Evoked Semantics (of language construct)	The meaning of a modeling language construct as assumed by a user of a construct-instantiating model.	Observe inferences users perform with the construct when using models that contain it.
Authoritative or Normative Semantics	The meaning of a modeling language construct as defined by its creator (the language designer).	Study language guide or manual.
Intuitiveness	Comprehensibility Appropriateness of modeling language (construct) after partial/limited training on it. (see also: [45])	As with comprehensibility appropriateness, with the restriction of partial training.
Mental Model	Abstractions that users of interactive artifacts form internally for the purpose of predicting and explaining the behavior of said artifacts [75]. Interactive artifacts are in our case diagrams to be "interacted" with by viewing. (see also: [47,77,98])	Indirectly through observing and interpreting user action and explanations thereof.
Cognitive Style	Adoption of a decision making or other problem solving strategy from a continuum between analytical and intuitive strategies [35].	(see Hammond et al. [35])
Cognitive Style Index (CSI)	Measures the propensity of adoption of analytic or intuitive strategies for problem solving [2].	A 38-point questionnaire [3].
Abbreviated Math Anxiety Scale (AMAS)	Measures math anxiety, i.e., the level to which respondents have feelings of fear, tension, and apprehension towards mathematics [10].	A 9-point questionnaire [37].
Mental Arithmetic Ability	Ability to quickly and correctly perform common arithmetic operations without the use of calculator and notes.	Success in a series of custom-made mental calculation tasks.

Table 2 Main constructs and measurements assumed in this study.

– **RQ2.2:** Does cognitive style specifically affect the method that users choose to use for performing inferences with the model?

To answer these questions we asked a number of experimental participants to perform two types of tasks. One task is similar to the one we performed in Section 2.2 to demonstrate the intuitiveness of contribution labels – but, this time, with more complex models. Specifically, experimental participants were given a number of decision problems in the form of a goal model with either numeric or symbolic contribution links. According to normative semantics for contribution links offered earlier (Subsections 2.2.2 and 2.2.3), the decision problem has a specific optimal decision with respect to a top level quality goal of interest. We ask participants, who are not revealed the exact semantics of the contribution links, to identify this optimal decision. Participants will then have to intuit and adopt some way of performing inferences using the contribution links in order to decide the optimal decision. Participants are

further asked if they simply followed their intuition to make the decision, or whether they followed a specific method, i.e., worked methodically. In latter case, they are then asked to describe the method they followed.

Utilizing the decision outcomes, we, firstly, calculate accuracy – i.e., the proportion of times that their decision is compliant to what the normative semantics would predict – and investigate the effect of representation (numeric vs. symbolic – RQ1.1), individual differences (RQ2.1) and whether a method was followed (RQ1.2) to accuracy. Then, we also investigate if following a specific method (versus working intuitively) is predicted by trait cognitive style (RQ2.2). If they follow a systematic method which they have described, we qualitatively analyze these descriptions to understand and codify how exactly the participants worked (RQ1.2).

A second task exposes participants to much simpler models consisting of a contribution link connecting two goals. The participants are given the satisfaction of the origin of the link and are asked to specify what they think the satisfaction level of the destination of the link should be. Aimed at addressing RQ1.1 and RQ1.2, the outcome is again compared with the normative, and the number of responses that are correct is investigated with respect to the kind of representation (numeric, symbolic, strong or weak contribution) and satisfaction status of the origin goal (positive, negative, strong or weak).

The two above types of tasks are organized into two separate sections of a data collection instrument. Moreover the results we report are based on three *rounds* of administration representing three stages in the evolution of the data collection instrument and utilization of three different samples including University students and Mechanical Turk [4, 18] participants. Below we describe our design in more detail starting from the experimental artifacts, i.e., the goal models we developed.

4.2 Experimental Artefacts

The experiment consists of a series of tasks performed sequentially on a computer by individual participants. The tasks that are key to the experimental objectives involve participants being presented with a goal model and asked to perform specific inferences with it. The goal models utilized for these tasks are constructed for the purpose of the experiment. There are two types of models that are developed, corresponding to the two separate sections of the experiment, *Section I* and *Section II*. We describe each type below, followed by a short discussion on the motivation behind devising the specific exercises.

4.2.1 Section I: Decision Models

We develop a set of goal models including an OR-decomposition and a quality goal hierarchy that represents criteria to be considered for the decision. Examples of such models can be seen in Figure 2. The models represent decision problems in three separate decision domains: choosing an apartment, choosing a course within a university program, and choosing a mode of transportation. Through the OR-decomposition, the participants are given apartment/course/mode of transportation choices, and the

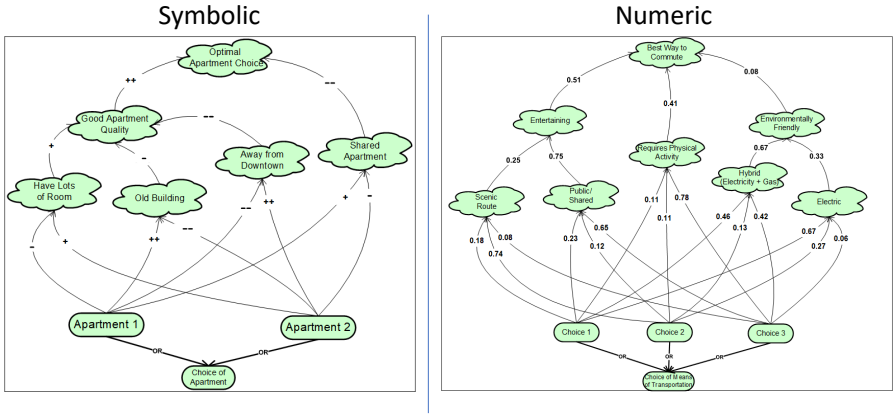

Fig. 2 Examples of goals models utilized in Section I of the instrument.

impact of such choices to high-level qualities such as location, schedule and environmental friendliness, respectively. The decision domains are chosen to be immediately understandable by the participant pool.

The quality goal hierarchy of each model is rooted on a unique quality goal such as *Optimal Apartment Choice* as seen in Figure 2 on the left. The labels are chosen in a way that one of the options is optimal compared to the other options with respect to the degree by which it satisfies the top level goal. Notice first that, depending on the labels of the contribution links, each child of the OR-decomposition implies a different satisfaction value for the root quality goal. To calculate that value of an OR-decomposition child in question we simply assign full satisfaction value to it (1.0 or {**FS**,**ND**} for numeric and symbolic models, respectively) while marking the others with no such evidence (0.0 or {**NS**,**ND**}, respectively). Then we apply the evaluation technique according to the type of contribution representation; for numeric we use weighted summations ([54] - Section 2.2.3), and for symbolic we use label propagation ([31] - Section 2.2.2).

Let us describe the choice of contribution labels in some more detail. In both cases, symbolic and numeric, the labels are chosen randomly, provided that the following condition is met: the satisfaction level of the root quality as it results from the selection of the optimal choice has a fixed distance from the corresponding value of the second best choice. We want this distance to be neither too large, in which case the optimal solution is too easy to spot, nor too small, in which case participants are likely to answer just randomly.

For numeric models, we set this distance to be 0.4 – we justify the choice below. For example, in the numeric model of Figure 2, it can be verified that the three choices have values $0.198, 0.199$ and 0.603, meeting the above requirement. For the model of Figure 1, we saw that the two options under *Have Expenses Reimbursed* have values of 0.25 and 0.75 wrt. *Overall Experience*. The distance is 0.5 hence too large for that specific model slice to meet the requirement.

For symbolic models, the comparison is more complicated due to the adoption of a two-valued qualitative framework [31] in which both satisfaction and denial values may co-exist in a solution, often in conflict. To allow identification of the optimal alternative and control the distance between the two top alternatives we convert the labels into numbers and aggregate them into one. Specifically each of the satisfaction labels **N**,**P** and **F** are associated with numeric values 0,1,2, respectively. Let $sat(g)$ and $den(g)$ be these numeric satisfaction and denial values of quality g in a given evaluation scenario, respectively. We aggregate the two numbers into $eval(g) = sat(g) - den(g)$. Value $eval(g)$ is then an integer in $[-2, 2]$. For example, the aggregated satisfaction value $eval(g)$ of a quality g with {**FS**,**PD**} is $eval(g) = sat(g) - den(g) = 2 - 1 = 1$. If g had a satisfaction status of {**NS**,**FD**}, then $eval(g) = sat(g) - den(g) = 0 - 2 = -2$.

Given this translation from the ordinal two-valued system to the interval one, the distance between the optimal and second-optimal satisfaction values can now be defined. We specifically demand that distance to be exactly 2 satisfaction levels. In the above example, the two satisfaction scenarios for quality goal g, {**FS**,**PD**} and {**NS**,**PD**} meet this requirement as 1 - (-1) = 2. However, neither pair {**NS**,**PD**} and {**NS**,**FD**} (-1 - (-2) = 1, too close) nor pair {**FS**,**ND**} and {**NS**,**PD**} (2 - (-1) = 3, too far apart) meet the distance requirement of 2.

The 2 satisfaction levels distance requirement was chosen based on our intuition of when the distance is becoming too large, revealing the optimal too obviously for meaningful measurement versus when it is becoming too small, when even experts in label propagation cannot guess the optimal without exhaustive calculation. Moreover, the choice of the numeric distance, 0.4, is made to allow comparability. With $eval(g)$ taking values from $[-2, 2]$, the distance of 2 satisfaction levels covers 50% of the available space. In numeric goal models the equivalent distance (50% of the space) would be 0.5. However, for some large model structures it was not possible to identify labels that allow for such large distances. Hence, the level was restricted to 0.4, which is slightly biased in favor of symbolic models given that wider distances are assumed to be easier to spot.

For each of the three domains (apartment finding, course selection, transportation choice), four (4) model structures are developed, two "small" including two choices and a smaller tree of quality goals, and two "large" including three choices and a larger quality goal tree. Two versions of each goal structure are instantiated, one with numeric contribution links and one with symbolic contribution links. Hence, a total of 2 (models) $\times$ 2 (sizes) $\times$ 3 (domains) = 12 models are instantiated for each of the two label representation types (symbolic and numeric). Each participant is exposed to one of the two sets of 12 models, either the symbolic or the numeric, in a *between-subjects* fashion with respect to representation.

Each of the 12 models is used to create a separate task for the participants. Each task includes displaying the model and asking the participant what the optimal alternative is for the displayed model. The tasks are organized into blocks based on the decision domain. Both the blocks between themselves and the models within blocks are randomly sequenced. Three additional warm-up decision problems are presented to participants, one from each of the domains, all small. These problems are otherwise the same as the actual decision tasks, except that responses to these problems

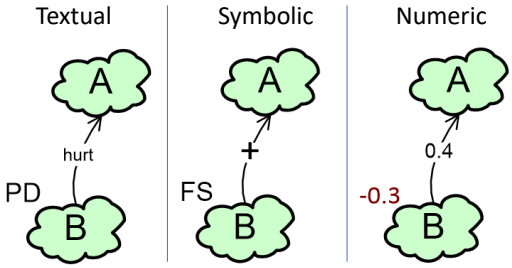

Fig. 3 Examples of goals models for Section II

are not counted towards the final scores. Thus, in all, each participant is exposed to 15 decision problems, the responses to the last 12 of which are the only ones counted. The responses of the 3 warm-up problems are not used for any other purpose such feedback or data filtering/qualification.

Before these task screens are presented, two short video presentations are offered, one describing the domains and another offering an introduction to goal models and contribution links. The latter video discusses the notion of contribution links at a high-level without disclosing any semantics or inference rules. Naturally, that video comes into two different versions, one for the symbolic and one for the numeric representation. The two versions are identical (same narration, structure, visuals, examples) except for the parts where the contribution link annotations need to be presented.

After they make the 12 (plus 3 warm-up) decisions, participants are asked if they followed a specific method, or whether they responded "intuitively". Response to this question constitutes the dichotomous *method* factor in the results. Further, if they answered that they followed a specific method, they were asked to describe in their own words how exactly they worked, using an example diagram as a prop for their explanation. In later rounds (more below) they are further asked how confident they are with the responses and/or the process they followed.

4.2.2 Section II: Individual Links

For the second section of the experiment we focus on a simpler type of model, consisting of two goals connected through a contribution link. We develop three sets of twenty (20) such models each. Each model contains two quality goals **A** and **B**, the former pointing to the latter through a contribution link.

The first set, which we call *symbolic*, all four (4) kinds of symbolic contribution links "++","+","−" and"−−" are considered. For each contribution link, five (5) models are devised corresponding to five different satisfaction levels of the origin goal: **FD**, **PD**, **N**, **PS**, **FS**. The satisfaction level of the origin goal appears as an annotation next to the goal shape. The resulting 4×5 models represent all possible combinations of origin goal satisfaction levels and contribution strengths. The second set, which we will call *textual*, is an exact copy of the first set except that the symbols "++","+","−" and"−−" are replaced with words *help*, *make*, *hurt*, *break*,

respectively – the default iStar 2.0 representation [19]. The third set, which we call *numeric*, is also a copy of the symbolic one with two differences. Firstly, symbols "$++$","$+$","$-$" and"$--$" are replaced with randomly chosen numbers from the intervals $[-1.0, -0.6]$, $[-0.6, -0.2]$, $[0.2, 0.6]$, $[0.6, 1.0]$, using precision of one decimal place. The intervals effectively discretize the interval [-1,1] into five constituent intervals, four representing various levels and qualities of contribution and the one in the middle ($[-0.2, +0.2]$) representing absence of contribution, and, as such, is not utilized. A similar mapping from symbols to numbers takes place at the level of satisfaction of the origin goal, in which the four satisfaction levels **FD**, **PD**, **PS**, **FS** are mapped to a random sample from the aforementioned intervals, respectively, and **N** is mapped to number zero (0). Examples of the three kinds of models can be seen in Figure 3.

Each model is used to create a separate task. Each task asks participants to examine the model and respond with what they think the satisfaction value of the destination goal should be. For *symbolic* and *textual* models an inventory of five satisfaction labels is offered for participants to respond. For the numeric models a text box is offered for the participants to enter a value between -1.0 and 1.0. The screens are given in random order.

4.2.3 AMAS, CSI, and Numeracy Tests

In addition to the core tasks described earlier the treatments include questions for measuring the participants' cognitive style, mathematics anxiety, and their ability with mental arithmetic. As we saw, the 38-point CSI Cognitive Style Index (CSI) [2] as well as the Abbreviated Math Anxiety Scale (AMAS) [37] are utilized for the first two measures. Unable to identify a standardized instrument with mental arithmetic tasks that are close to the ones that we would assume participants of the numeric goal models would perform, we resorted to developing our own. We discuss the exact form of the numeracy tests in the results section.

4.2.4 Section I and II Tasks: Rationale

Let us now discuss the rationale for developing the above artifacts and tasks vis-à-vis our research questions. In the tasks of Section I, goal models are utilized for representing decision problems: alternatives are represented as OR-decompositions and contribution links are used to show how each alternative affects various quality criteria of interest. Assuming contribution links have precise semantics, each model has a clear optimal alternative according to these semantics. If participants, who are unaware of the precise semantics, guess that optimal, this is evidence that the semantics align with how users naturally interpret the labels. This, in turn, supports that the contribution link construct – the package of representation and semantics – is intuitive. If the reverse is observed, i.e., participants cannot guess the optimal, such conclusion is instead discouraged. The tasks check how the two representation approaches compare with regards to intuitiveness (RQ1.1) and if the individual differences summarized above play an additional role (RQ2.1). Solicitation of a free-form description

of the method followed aims at clarifying if success in identifying the optimal can indeed be attributed to correctly guessing the underlying semantics (RQ1.2). We further investigate if following a concrete method at all (vs. working intuitively) is affected by trait cognitive style (RQ2.2).

The tasks of Section II follow the exact same measurement principle at a different level. Rather than intuiting how contribution links are combined, participants are asked to instead combine a satisfaction value with a contribution label to produce the target satisfaction value. Again, whether the response agrees with the normative of each representation is a measure of the intuitiveness of the latter (RQ1.1). Section II tasks are aimed at clarifying and diagnosing the outcome of Section I. For example, if Section I tasks indicate that weighted summations are intuitive, Section II clarifies if participants explicitly multiply weights with satisfaction values, or (as it turns out) follow a different semi-formal procedure that is simply compatible with but not necessarily the same as weighted summations. In addition, Section II models explore the use of negative labels and satisfaction values for numeric models. Likewise, if symbolic models turn out intuitive or unintuitive for making decisions, Section II explains the circumstances that may cause this outcome. Note that the simplicity of the exercise, makes the study of individual differences and chosen method irrelevant. Hence Section II exclusively serves RQ1.1.

4.3 Administration Rounds and Participants

An ordered presentation of Section I and Section II tasks, the CSI, AMAS and Numeracy Tests as well as other questions such as demographics constitute the experimental instrument by which data is collected from participants. PsyToolkit [89,90] is used for administering the tasks. In total, three (3) rounds of data acquisition are performed, each with a slightly different version of the instrument and a different sampling method.

More specifically, round 1 is administered to students of York University, taking a first year undergraduate management course, who are offered bonus grade for their participation. Round 2 is administered to Information Technology students of York University, having just finished a third-year Human Computer Interaction course (they are offered a small gift card for their participation) and, as a follow up, to Mechanical Turk Participants with US college degrees. Round 3 is exclusively administered to Mechanical Turk Participants with US college degrees.

In each round, the instrument undergoes revisions, rearrangements, and improvements. In Table 3 the relevant tasks and the order by which they are offered in different rounds can be viewed. The tasks, listed in the first column, include response to the CSI and AMAS questionnaires (*CSI* and *AMAS*, respectively), response to the *Numeracy Tests*, provision of demographic information (*Demographics*), a video on making decisions under multiple criteria (*Decisions Training*), a video on goal models and making decisions therewith (*Goal Models Training*), the 12 decision exercises including the 3 warm-ups (*Section I Tasks*), the question on how confident the respondent is with their decisions (*Response Confidence*), the question on whether the participants used their intuition or a specific method, followed by a description of – if

Sample:	1st Year Administrative Studies	3rd Year Information Technology & Mechanical Turk	Mechanical Turk (only)
Task:	**Round 1**	**Round 2**	**Round 3**
CSI		1	1 (pre)
AMAS		2	3
Numeracy Tests		3	12
Demographics	1	4	2
Decisions Training		5	4
Goal Models Training	2	6	5
Section I Tasks		7	6
Response Confidence			7
Method Declaration (& Description)		8	8
Method Confidence			9
Contributions Training	3	9	10
Section II Tasks	4	10	11

Table 3 Instrument construction and administration rounds. Each column describes the sequence by which components are presented to participants. Components administered on a separate pre-test are marked with "(pre)".

applicable – the specific method (*Method Declaration (& Description)*), the question on how confident the respondent is with their method (*Method Confidence*), the video describing contribution links in more detail (*Contributions Training*) in preparation to individual links tasks (*Section II Tasks*).

Round 1, specifically, which was devised during early stages of this research, is an initial study solely including the Section II tasks, whereas the remaining rounds include both sections. For the remaining two rounds, the instrument is updated in 3 ways. In round 2, Section I Tasks is added as well as CSI, AMAS and Numeracy Tests. In round 3 the following changes are made: (a) Response Confidence and Method Confidence questions are added (described above), (b) Numeracy Tests are revised based on results from the previous rounds, (c) the order of administration is updated (Numeracy Tests are now at the end). As we discuss below, we consider the differences between rounds 2 and 3 to be minimal enough to allow for pooling of the corresponding data following specific checks.

4.4 Participant Demographics

A total of 196 participants participate in the experiment: 35, 29 and 132, respectively are 1st year business students (round 1), 3rd year IT students (round 2) and Mechanical Turk workers (rounds 2 and 3 – 30 and 102, for each round respectively). Of them, 93 are female and 103 are male. Their fields of (current or former) study are predominately (more than half) Science, Technology and Engineering, Business and Economics. Precise data can be seen in Table 4. Participants of round 1 only provide their sex; though their academic field must be assumed to be in the Business and Economics category.

	1st Year Business		3rd Year IT		MTurk	
	Female	Male	Female	Male	Female	Male
Science, Technology and Engineering	0	0	6	23	18	19
Business and Economics	0	0	0	0	11	27
Health Sciences	0	0	0	0	5	3
Social Sciences	0	0	0	0	9	5
Humanities	0	0	0	0	11	10
Fine Arts	0	0	0	0	3	4
Education	0	0	0	0	4	1
Other	0	0	0	0	0	2
N/A	26	9	0	0	0	0

Table 4 Participant Demographics

For all but round 1 participants, AMAS and CSI indexes are collected. The overall CSI average was 47.47 which is above reported averages in the literature (44.53 according to the CSI manual and Hmieleski and Corbett studying US college students [36]). The overall AMAS average is 20.86 which is just below the reported averages in the literature (21.1 according to D.R. Hopko et al. [37]).

In the two sections that follow we present the results for Section I (decision models) and Section II (single links) respectively. Given the absence of any prior evidence in the literature on the topic – intuitiveness of contribution links for goal models – we consider our analysis to be exploratory [88]. Hence hypotheses are formally constructed for only some of the analysis, where inferential statistics are possible, and by default we hypothesize the presence of an effect for each of the involved the factors. These are supplemented with visualizations and descriptive analyses.

The experimental data as well as complete markdown presentations of the analyses can be found in our data repository [52][1].

5 Analysis and Results: Section I

5.1 Measurements, Factors and Analysis Approach

As we saw, the main measure of intuitiveness (in both sections) is accuracy, i.e., the number of times participant responses agreed with the normative/authoritative ones. Recall that the normative optimal is given by application of symbolic label propagation for symbolic models, and by the weighted summations approach, for numeric models, both discussed in Section 2.

The main explanatory variables are *representation group* (or henceforth interchangeably *representation* or *group*) which refers to whether the models are numeric or symbolic, individual differences measured through *CSI*, *AMAS*, as well as the *method* that participants stated that they followed, i.e. methodically or intuitively.

[1] For review purposes, the data can be accessed using this private URL: `https://borealisdata.ca/privateurl.xhtml?token=ea6aaabc-7ab1-4c77-98e3-a567d1a46184`. Consent agreement allows publication of data as presented there.

Factor Name	Related Task (Table 3)	Factor Description
representation group (or *group* or *representation*)	[random assignment to symbolic or numeric instrument]	Whether participant was exposed to symbolic or numeric models.
CSI	CSI	Whether participants' Cognitive Style Index (CSI) is above or below population average (analytic and intuitive types, respectively).
AMAS	AMAS	Whether participants AMAS score is above or below population average (high and low math anxiety, respectively).
method	Method Declaration (& Description).	Whether participant followed "their intuition" or a "specific method" (according to their own declaration).

Table 5 The explanatory variables considered in the analysis; all dichotomous.

A summary of these factors is offered in Table 5. The following null hypothesis are tested, corresponding to the research questions posed above (Subsection 4.1):

- $\mathbf{H_0^{I,1}}$: There is no difference in response accuracy between numeric and symbolic groups, i.e. average accuracy measures between the two groups are equal. (RQ1.1).
- $\mathbf{H_0^{I,2}}$: Accuracy does not depend on chosen method, i.e. the mean accuracy scores of those who followed a specific method and those who used their intuition are equal. (RQ1.2).
- $\mathbf{H_0^{I,3}}$: AMAS does not affect accuracy, i.e. those with high AMAS (math anxious) achieve the same accuracy as those with low AMAS (not math anxious) (RQ2.1).
- $\mathbf{H_0^{I,4}}$: CSI does not affect accuracy, i.e. those with high CSI score (analytic) achieve the same accuracy as those with low CSI score (intuitive) (RQ2.1).

Note that, for brevity, the above hypotheses are assumed to also include effects of each factor in the context of interactions.

ANOVA models [68] are developed for exploring the relationships between explanatory and response variables. In particular we test the maximal (in number of factors) model in which all four factors (*group*, *AMAS*, *CSI*, *method*) as well as interactions between each of them and factors *group* and *method* are included. Section I data are available from rounds 2 and 3 only. Data from both rounds are first analyzed together. Given that they have a difference in the sequence of tasks (numeracy tests precede or succeed respectively the tasks in question, and that round 2 has samples from different two sources (students and Mechanical Turk participants) we include two additional factors, *sample* and *phase*. Depending on whether significant effects are found in these two factors or not, we perform a separate analysis for each set (at a discounted α level to limit family-wise error) or continue with analyzing the data together, respectively. We discuss these choices in more detail in the validity section. Further, to simplify modeling and interpretation, CSI and AMAS are discretized into two-value variables based on whether the score exceeds the population average or

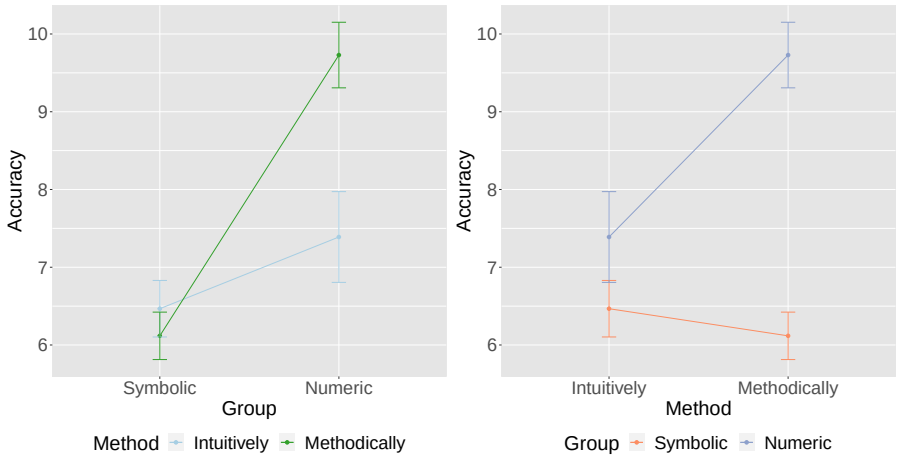

Fig. 4 Interaction plots for Representation Group and Method

not. Finally, separate analyses investigate the relationship between CSI and method chosen, as well as the relationship of numeracy scores with accuracy.

5.2 The role of representation and approach

Fitting an ANOVA model as described earlier produces, among other effects, an interaction between *sample* and *group* ($F(1,145) = 6.59$, $p = 0.011$). As per our methodology, we, hence, proceed with performing separate analyses for the two data sets, i.e., the student sample (n = 29) and the samples from Mechanical Turk (n = 132). The model now is restricted to the factors that appear to be relevant: *group*, *method*, *CSI*, *AMAS* and their in between interactions.

A look at the student data (29 cases, 15 symbolic and 14 numeric) reveals that the sample is too unbalanced for reliable inferences if *method* is included. Thus, for the student data only, we drop this factor and any interaction terms in which it participates. The result with the simplified model indicates a strong (Cohen's $d = -2.43$ *(large)*) main effect on *group*, $F(1,23) = 15.53$, $p < 0.001$, and no other effects or interactions. Hence numeric models evoke more accurate responses than symbolic and by a large margin: group means of accuracy scores are 10.64 vs 5.8 out of a maximum 12, respectively.

The Mechanical Turk sample, which is large enough to allow for the original model (132 cases, 66 symbolic and 66 numeric), yields strong interactions between *group* and *method* ($F(1,116) = 6.55$, $p = 0.012$) and between *group* and *AMAS* ($F(1,116) = 4.82$, $p = 0.03$). While variances appear to be homogeneous across cells, some violations of normality assumptions prompt us to perform also Wilcox's non-parametric equivalents, which identify the same interactions ($p = 0.007$ and $p = 0.022$).

The first interaction is between the method that participants adopted for performing the tasks and the kind of representation that they were assigned to. In Figure

4, the nature of the interaction can be seen more clearly. Referring to the interaction plot on the left, for symbolic models whether or not an intuitive method was followed does not seem to affect accuracy. On the contrary, for the numeric group, following a specific method helped participants achieve better accuracy – Wilcoxon rank sum $W = 211$, p = 0.001, effect size = 0.4 (*moderate*). Measured in terms of difference in the mean correct answers, participants of the numeric group who work methodically perform on average 2.34 more correct tasks (out of 12) than their members in the same group who work intuitively (9.73 vs. 7.39).

Another way to see the same effect, visualized in the interaction plot on the right of Figure 4 is that those who work intuitively do not benefit from working with numeric models more than working with symbolic. Of those who work methodically however, participants working with numeric models answer on average 3.61 more correct questions compared to those working with symbolic models [9.73 vs. 6.12; Wilcoxon rank sum $W = 411$, $p < 0.001$, effect size = 0.58 (*large*).

5.3 Qualitative Descriptions

Recall that after performing the decision tasks, participants are asked it they used their intuition or a specific method to make the decision. (This binary method declaration informs, as we saw, the *method* explanatory variable.) Those who say they used their intuition go to the next task, while those who say they followed a specific method are asked in the next screen to describe that method. We now focus on that data, aimed at understanding the precise method that methodical participants follow that makes them successful with numeric models but not so with symbolic.

For the analysis, we performed a simple iterative labeling task akin to grounded-theoretic open coding [17]. Specifically, by reading the responses we identify labels that describe patterns of work that participants are following to identify optimal decisions. We iterate in order to refine the coding scheme and also identify dimensions along which the participant approaches vary. We identify two such dimensions: the way by which participants compare and/or combine contribution labels, and the direction they follow in order to analyze and compare the alternatives.

Figures 5 and 6 depict the categories we identified for each dimension, the occurrence frequency among those who gave a response (96%) of each, and the accuracy attained by the participants following the corresponding strategy. Although most of the times it was difficult to exactly discern from their descriptions the method the participants used to make the decision (identified as "Unclear" in the graph), for a good part of the descriptions we are able to identify some common themes.

Starting from the Numeric group of Figure 5 on the left, most participants (32) do not offer sufficient detail on how they worked, despite some indications of varying specificity. For example in one participant's words *"I looked at what percentage each choice applied to the optimal choice at the top of the hierarchy, and worked myself down to see which option applied the highest percentage to the top tier choice"* (Excerpt 1), or in another's *"I followed the path with the highest contributions"* (Excerpt 2). These examples seem to indicate some general patterns of work – e.g., the first one may be following the weighted summations approach – but are too ambiguous

to be classified with certainty and/or reproduced. This category includes participants who offer even less detail. For some participants it was clear that they followed a technique which involved some kind of navigation from node to node whereby the contribution link with the strongest label would be followed to the next node of interest. For example, *"I start at the top and whichever number is higher, I go down that route. By following this technique going down, I eventually end up with the optimal choice"* (Excerpt 3). Although the heuristic is not guaranteed to offer the optimal answer vis-à-vis the normative weighted summations procedure, it does however work for the randomly prepared cases of our experiments and participants indeed appear to be successful by following this approach. Other heuristics followed include adding numbers for *"each strand"*, a scheme of node scoring and a swapping scheme. In traversing the links, in responses where it was clear what directions they followed, participants worked predominately top-down (see Excerpts 1 and 3 above), with a few cases declaring bottom-up or a combined approach. Some simply mentioned that they worked along paths (Excerpt 2).

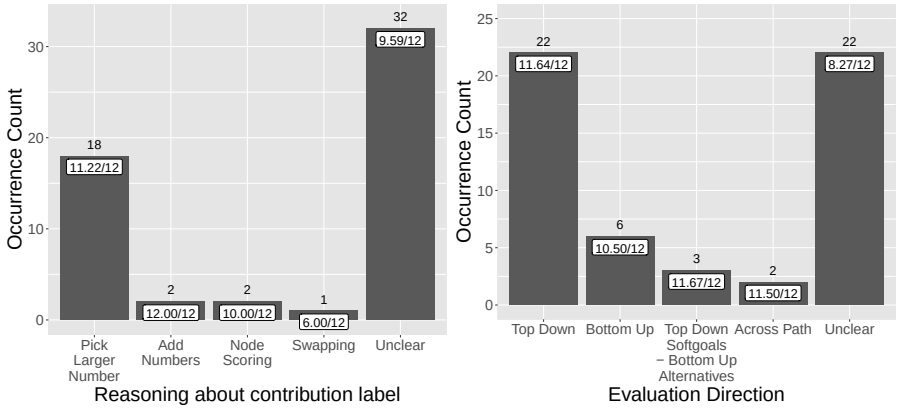

Fig. 5 Self-reported Method Descriptions – Numeric Representation Group. The average accuracy exhibited by the participants in each category is displayed with white background.

Figure 6 offers a view of the descriptions in the Symbolic group. In this group, participants predominately seem to adopt a symbol counting technique, e.g. *"I looked for the option with the most amount of (+) symbols"* and *"Pick the one that has most + signs over − from all routes available to Optimal Choice"*. Those who apparently follow a top-down traversal process similar to the one that was popular in the numeric group are labeled under *"Pick Stronger Symbol"*. For example, *"I started from the main goal 'Optimal apartment choice' and chose the positive link, or the most positive one, and went down the criteria, looking for the most positive route"*. Following such a process would lead participants at a minimum 2 and maximum 9 (mean = 5.66) of the 12 times to the response that is correct according to the authoritative calculation.

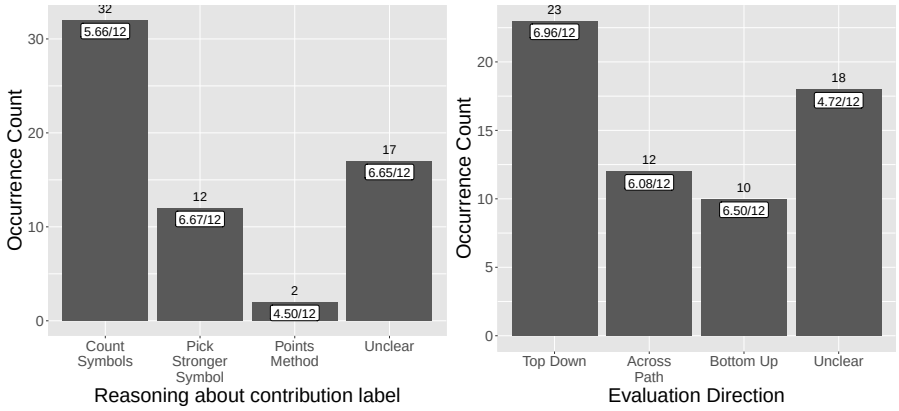

Fig. 6 Self-reported Method Descriptions – Symbolic Representation Group. Average accuracy in white background on top of the category that exhibited it.

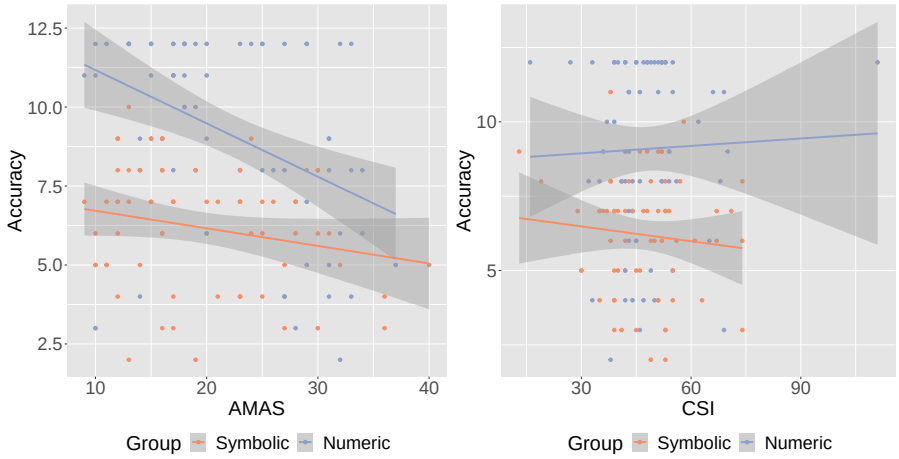

Fig. 7 The effects of AMAS and CSI to accuracy by Group.

5.4 CSI and AMAS

Let us now turn our focus to *CSI* and *AMAS* and their effect on accuracy based also on the representation group, as it emerged in the Mechanical Turk sample. As we saw *AMAS* appears to interact with *group*. However, Figure 7 shows this interaction to not imply a qualitative difference. Increased AMAS indeed implies lower accuracy for both representations, through even more so for numeric models where the difference is statistically significant – Wilcoxon rank sum $W = 318.5$, p = 0.003, effect size = 0.18 (*small*).

CSI scores appear nowhere in the statistically significant results, leading us to the hypothesis that the specific index does not relate with participants' accuracy or

Task	Example	Round 2	Round 3
Addition	$0.76 + 0.19 =?$	4 × (90s)	–
Addition Comparison	$0.92 + 0.16$ vs. $0.25 + 0.89$	4 × (90s, .05)	–
Subtraction Comparison	$0.81 - 0.12$ vs. $0.93 - 0.28$	4 × (90s, .05)	–
Multiplication	$0.28 \times 0.27 =?$	4 × (90s)	4 × (*) + 4 × (20s)
Multiplication Comparison	0.48×0.29 vs. 0.12×0.79	4 × (90s, .05)	4 × (*, .05) + 4 × (15s, .15)
Division	$0.46/0.62 =?$	4 × (90s)	–
Division Comparison	$0.25/0.98$ vs. $0.11/0.55$	4 × (90s, .05)	–
Linear Combinations Comparison	$0.39 \times 0.97 + 0.46 \times 0.58$ vs. $0.90 \times 0.82 + 0.56 \times 0.74$	–	2 × (*, .25) + 4 × (15s, .5)

Table 6 Mental Arithmetic Tests. Round 2 is administered to Students and MTurk participants and round 3 to MT participants only. The number in the cell displays the number of exercises of the specific type times, in parentheses the time participants had to respond comma the distance between responses when there was a comparison. E.g. 4 × (15s,.15) means four questions of the type, 15 seconds each and distance in the comparison is 0.15 and 2 × (*) means two questions of the type, no time limit.

even the method they choose to reason about the diagrams. Hence, analysis of the proportions of high-CSI participants who chose to work intuitively, compared to the low-CSI ones who made the same choice reveals no effect (Fisher's exact test). To see if we can make population inferences from this negative result, equivalence of proportions analysis is then performed assuming equivalence bounds of -0.15 and +0.15. The equivalence test was significant, Z =-2.2, p = 0.015. This means that there is no difference in said proportions that is greater than 0.15 (the actual confidence interval being close to 0.1).

Likewise, we investigate whether accuracy scores from high CSI and low CSI participants (ignoring other factors) are equivalent at a small-to-medium effect level $d = 0.35$. The equivalence test was significant, t(145.8) = 2.109, p = 1.83e-02. That is, the difference between the scores produced by the two CSI types is not greater than small-to-medium in the population (for $\alpha = 0.05$).

5.5 Mental Arithmetic

Recall that one of the tasks that participants performed was a set of tests on mental arithmetic. We devised our own tests that fit the kind of arithmetic that could be used by participants to reason about numeric goal models. The tests consisted of addition, subtraction, multiplication and division exercises with random numbers in the interval $(0, 1)$ with two significant digits. Some exercises ask for the result of an operation whereas others offer two operations with results that have a known fixed distance and ask participants which one is greater. For the former type, a 0-10 scoring is assigned based on an exponentially decaying function of the distance between correct and provided answer. Table 6 offers details on these tests and how they were updated from one round of the experiment to the next.

To measure the effects of these numeracy tests, we calculate Kendall correlations between the test scores and the accuracy scores for each round and representa-

	Numeric			Symbolic		
	Student	MTurk1	MTurk2	Student	MTurk1	MTurk2
Addition	**0.36***	0.17	-	-0.22	0.15	-
Addition Comparison	0.05	-0.14	-	0.08	0.10	-
Division	0.18	-0.21	-	-0.19	-0.04	-
Division Comparison	0.02	0.02	-	0.07	-0.03	-
Linear Comparison	-	-	0.19	-	-	0.22
Multiplication	-0.01	0.10	0.06	-0.11	**-0.30**	0.24
Multiplication Comparison	0.03	-0.10	0.19	-0.15	0.08	0.26
Subtraction Comparison	**0.45**	**0.45**	-	-0.06	-0.24	-

Table 7 Pearson correlation coefficients between numeracy test components and accuracy scores. MTurk 1 are round 2 and MTurk 2 are round 3 participants. None is statistically significant $p < 0.05$, significant with $p < 0.1$ marked with * and correlation values 0.3 and above in **bold**. Note also the presence of (unintuitive) negative correlations.

tion group. The results can be seen in Table 7. Overall, very few strong correlations emerge – only one statistically significant – and in patterns that are not interpretable. For example ability in linear combination comparisons, does not seem to correlate to accuracy in numeric models more than it does for symbolic models, despite the fact that such operations describe the formal procedure for evaluating numeric models.

Given that our tests are not standardized, the construct validity threat is, of course, salient here. Assuming, however, that the tests do successfully measure ability to mentally perform arithmetic operations, the fact that accuracy does not correlate with mental math ability may imply that such mental math operations are never performed by participants. Rather, as evident in the self-reported commentary discussed above, they devise simpler heuristics in which numbers are compared in isolation, rather than through additions or multiplications. This is consistent with our finding in Section II below in which, in their majority, participants do not appear to perform recognizable arithmetic operations even when confronted with a single contribution problem.

5.6 Response and Method Confidence

Recall that a question on how confident participants felt on their responses and the method they followed to make the decisions, was introduced in Round 3 – hence, data on that aspect is collected from 102 Mechanical Turk participants.

The results show that participants are overwhelmingly confident in both their responses and the method they used: 71 of the 102 participants agree or strongly agree that they are confident with the method they followed and 73 agree or strongly agree that they are confident with their responses. Individual correlation tests do not reveal notable differences in confidence between representation *group*, *method* chosen or *AMAS* score.

Some relationship of *CSI* and response and method confidence can also be observed. According to Hammond et al. [35] intuition implies high-confidence in answer but low confidence in method, while analysis is associated with the opposite. As seen in Figure 8, a slightly higher response confidence can indeed be observed among the intuitive respondents (those with *CSI* below population average) com-

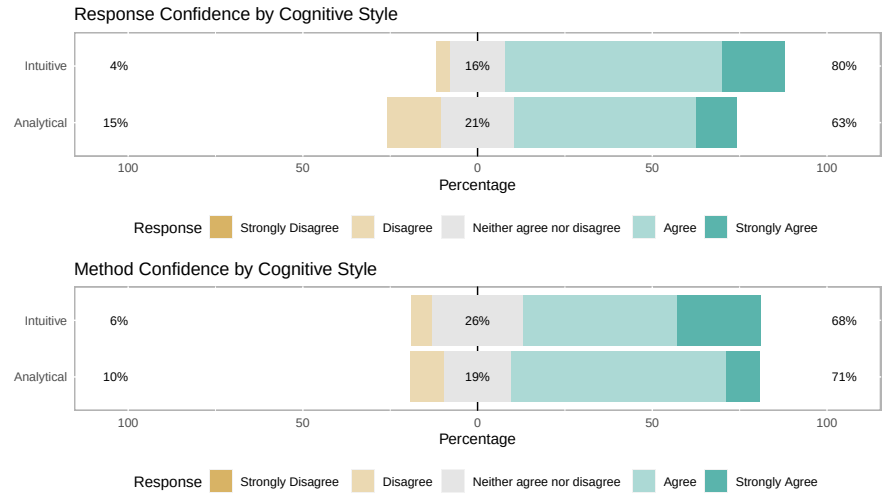

Fig. 8 Response and method confidence with respect to cognitive style (intuitive is CSI score below population average 45.1 and analytical is above that average). Questions: *"I am confident of the answers I gave in the optimal decision exercises I just completed"* and *"I am confident of the method I used to find the optimal alternative in the decision exercises."*

pared to their analytical peers. Less can be inferred about method confidence from the graph. Accordingly, although correlation between *CSI* and response confidence agrees with theory and the graph (Kendall's $\tau = -0.19, p = 0.017$) the correlation between *CSI* and *method* is too weak ($r_s = -0.12$) and statistically insignificant for conclusions.

5.7 Section I: Summary of findings

To summarize the findings of Section I, let us, first, examine the status of the null hypotheses put forth earlier. Hypothesis $\mathbf{H_0^{I,1}}$ (group effect) is rejected in the student sample as a main effect and in the Mechanical Turk sample in the context of interactions: the effect occurs for methodical participants. $\mathbf{H_0^{I,2}}$ (effect of method chosen) is also rejected in the Mechanical Turk, again, in the context of interaction with group (working methodically or not matters only for the numeric group) but is not tested in the student sample, due to highly unbalanced data. $\mathbf{H_0^{I,3}}$ (AMAS effect) is also rejected through in the Mechanical Turk data, again for the numeric group only but with low effect size; it is not rejected in the student data. We fail to reject $\mathbf{H_0^{I,4}}$ (CSI effect) in any of the two samples.

Given the above, combined with the qualitative and descriptive analyses, some general observations can be made with regards to the outcomes of Section I. Firstly, the majority of participants appear to adopt a specific method for reasoning about the models, instead of working intuitively – i.e., abstractly or even randomly. This shows that the visualization itself and the abstract introduction to it may evoke some kind

of a mental model (see Section 3) of what the conceptual model means and how it "works".

The representation group effect that was observed among those that claimed to have followed a specific method, combined with the qualitative data supports this method adoption hypothesis. Specifically, participants exposed to numeric contributions were successful by seemingly adopting a heuristic that led them to the authoritative optimal with high likelihood. The corresponding heuristics adopted by the symbolic group led to solutions that did not coincide with the authoritative ones. An explanation of the high accuracy of the numeric group is the familiarity of participants with numbers, on one hand, and the naturalness of viewing the numbers as proportions as per the normative weighted summations approach, on the other. One can go on and specifically hypothesize that numbers evoke more accurate responses due to their affording familiar mental arithmetic that unfamiliar symbols do not. However, according to the method descriptions offered by the participants, rather than complex arithmetic calculations, they seem to work along paths simply making comparisons along the way. It is, hence, the compatibility of the normative approach with the participants' ad-hoc approach that seems to bring about the accuracy effect. As we discuss towards the end, this has useful design implications.

Further, we could not find evidence that the cognitive style index (CSI) appears to play a role in attaining accuracy for either group, that it interacts with the group factor or that it is even a strong predictor of the method that participants choose. We instead find that effects that are small-to-medium or larger are not likely to exist in the population. Future studies may attempt alternative assessments of the construct – e.g. by Epstein et al. [24].

Finally, despite the inconsistencies in accuracy, participants are confident of both the response and the method they followed, and their confidence does not appear to be affected by representation or other factor. Consistent to expectations there is an effect of CSI to response confidence, albeit a weak one.

6 Analysis and Results: Section II

We now turn to the tasks of Section II of the experiment. Recall that for Section II, participants assign satisfaction values (e.g., **FD**, **PD**, 0.4, 0.8, etc.) to the destination of a contribution link displayed to them given the satisfaction value annotating the origin of the contribution link (Figure 3). The resulting values are analyzed with respect to the agreement within participants (Section 6.2) and accuracy vis-à-vis the authoritative values (Section 6.3); both measures defined below. Further, focusing on numeric models we look at whether and what kind of arithmetic operation participants are likely to perform (Section 6.4). Finally, we explore the data for models with zero satisfaction origin in a separate analysis (Section 6.5).

6.1 Measurement and Analysis Approach

To calculate either agreement or accuracy in a way that numeric and symbolic models can be compared, we first need to map satisfaction values **FD**, **PD**, **N**, **PS**, **FS**

(or intervals $[-1.0, -0.6]$, $[-0.6, -0.2]$, $[-0.2, 0.2]$, $[0.2, 0.6]$, $[0.6, 1.0]$ for numeric models) into integers in the interval [1,5]. Depending on the answer each participant offers, the corresponding code is used for the analysis. For example **N** is coded as 3 in the textual or symbolic groups and 0.5 is coded as 4 in the numeric group.

We calculate the *agreement* within participants with respect to their responses, via measuring the average distance between each pair of participant responses. Let $r_i(l) \in [1, 5]$ be the code of the response of a participant i in exercise l. To calculate average pair-wise distance, for each exercise l we identify all pairs of participant responses $r_i(l)$ and $r_j(l)$, $i, j = 1 \ldots N, i \neq j$, N being the number of respondents for the exercise. For each pair we then calculate the normalized distance $|r_i(l) - r_j(l)|/4$, and average over all $N(N-1)/2$ pairs. Hence, average pairwise distance $apd(l)$ for each exercise l is given by:

$$apd(l) = \frac{|r_i(l) - r_j(l)|/4}{N(N-1)/2} \tag{1}$$

The lower the apd is for an exercise l the higher the agreement among participants.

Considering the authoritative response according to the theories detailed in Section 2, we calculate *accuracy* through computing the distance between participant response $r_i(l)$ in exercise l and the authoritative response $a(l)$, both coded as above:

$$dist_i(l) = r_i(l) - a(l)$$

Again, the lower the distance the higher the accuracy. Further, when $dist_i(l) > 0$ we say that the participant i *overestimates* the satisfaction of the destination goal, assigning to it values higher than the normative. Likewise, when $dist_i(l) < 0$, i *underestimates* the satisfaction of the destination goal.

For both agreement and accuracy analyses we consider three main relevant factors. One is *contribution quality* with levels *positive* and *negative*, representing the corresponding effect of the contribution link in each exercise. Hence, links $-$ and $--$ and their corresponding textual and numeric versions are negative, while $+$ and $++$ and their corresponding versions are positive. A second factor is *origin (satisfaction) quality* with level *denied* if the origin goal in the exercise is denied with **FD**, **PD** or an equivalent numeric, level *satisfied* if the goal is satisfied with **FS**, **PS** or an equivalent numeric, and level *none* if the goal is marked as **N** or with 0 satisfaction. We will further refer to combinations of contribution and satisfaction qualities as *configurations*, e.g. the *Denial-Positive* configuration. Thirdly, the factor *group* represents the contribution link representation approach with levels *symbolic*, *textual* and *numeric*.

Wherever inferential procedures are possible, which is in accuracy analysis, the following null hypotheses are tested, all relating to the research question RQ1.1 of Section 4.1:

- $\mathbf{H_0^{II,1}}$: There is no difference in response accuracy between symbolic, textual and numeric groups.
- $\mathbf{H_0^{II,2}}$: Accuracy does not depend on the configuration of contribution quality and satisfaction level.

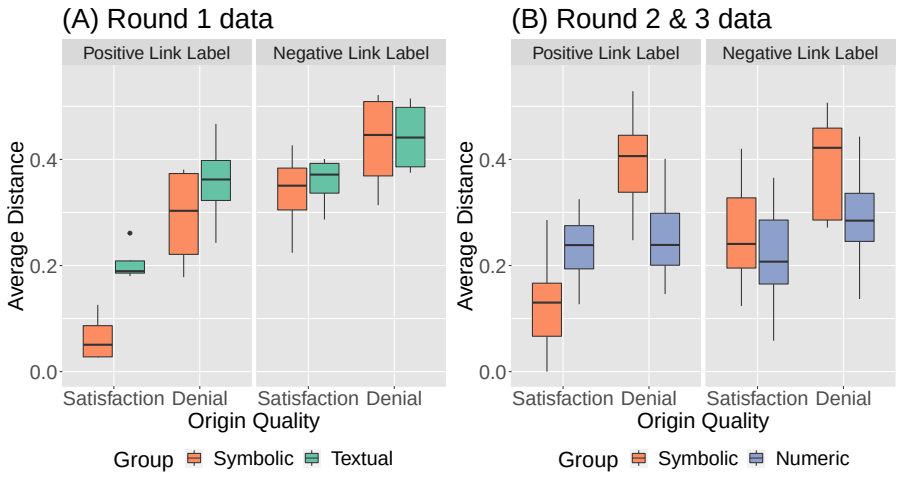

Fig. 9 (A) Agreement for round 1 data [Students, Symbolic vs. Textual](B) Agreement for round 2 and 3 data [Students, MTurk, Numeric vs. Symbolic]

6.2 Agreement Analysis

We compare descriptively the role of satisfaction and link quality to the overall agreement among participants, measured as above. Recall, that for round 1 (Table 3) the comparison is between symbolic and textual representation while for rounds 2 and 3 it is between symbolic and numeric. This analysis excludes the cases in which the satisfaction is “none", which are dealt with separately (Section 6.5).

The data from round 1 can be seen in figure Figure 9(A), noting that the lower the number the higher the agreement. It is clearly the case that satisfaction and positive links lead to better agreement, which decreases with the presence of a denied origin or a negative link, and becomes even lower when denied origin and negative link are combined.

For rounds 2 and 3, where numeric labels are compared against symbolic the result is seen in Figure 9(B). While agreement in symbolic models decreases with the presence of denial in the origin goal and negative contributions, agreement in numeric models remains largely unaffected. It is not clear, however, if any of the groups evokes higher agreement overall.

6.3 Accuracy Analysis

6.3.1 Round 1: Symbolic vs. Textual

We first visualize accuracy with respect to, again, origin satisfaction quality, link quality as well as link representation (group). The first comparison concerns the round 1 data in which symbolic and textual representations are compared. A visualization can

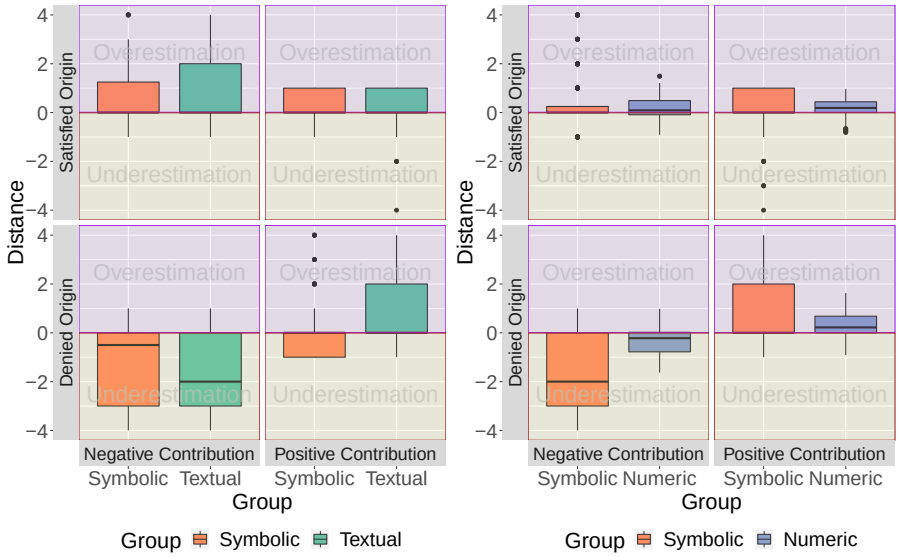

Fig. 10 (A) Accuracy for round 1 data [Students, Symbolic vs. Textual] | (B) Accuracy for round 2 and 3 data [Students, MTurk, Numeric vs. Symbolic].

be seen in Figure 10(A). Recall that when the distance from the normative is positive, the participant has overestimated the satisfaction of the destination goal, and vice-versa when it is negative. We observe that while in the cases of a satisfied origin and a positive link participants generally overestimate satisfaction of the destination, the opposite is strongly the case when origin denial and negative contribution are combined.

To explore this effect better we compare the distributions of responses of the two extreme cases in Figure 11. Graph 11(A) presents the response count for each combination of partial or full origin denial with weak (*hurt* / $-$) and strong (*break* / $--$) negative contribution labels, while the second graph (B) represents the corresponding counts of partial or full origin satisfaction with weak (*help* / $+$) and strong (*make* / $++$) positive contribution labels. In both graphs the bars representing responses that are compliant to the normative are marked with a thicker outline as "Correct".

Focusing on graph (A) of Figure 11, we observe that responses are often symmetric around **N** with the respondents ambivalent between a positive and a negative satisfaction value. In the *break* / $--$ and **FD** combination (top right histogram) of graph (A)) there are almost as many **FS** (correct) as there are **FD** (wrong). A similar pattern can be seen in all combinations. In other words participants fail to recognize the satisfaction reversal effect that a negative contribution has, in which, according to the designed semantics, a denied goal and a negative contribution becomes satisfaction evidence for the destination goal.

Moving to graph 11(B), on the other hand, disagreement is between the strength of satisfaction rather than its quality. In three of the cases the majority of participants offer a compliant response, except for the case of full satisfaction and a weak

contribution, where participants believe should still cause full satisfaction of the destination, instead of partial. Regardless of this, the absence of satisfaction reversal allows for more compliant responses.

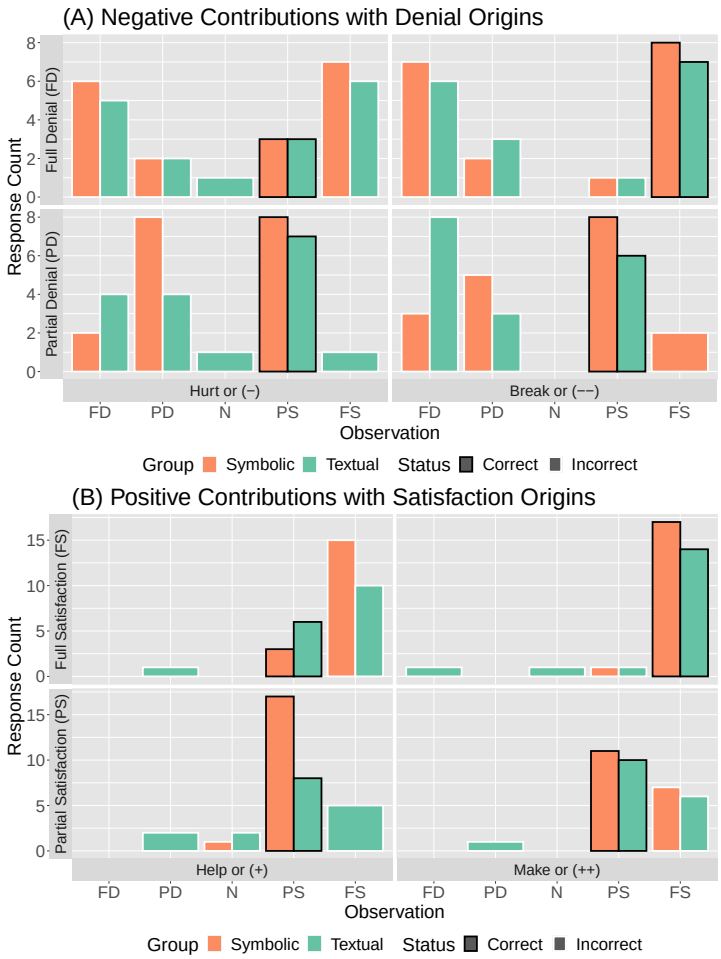

Fig. 11 Response counts for round 1 data per contribution label and origin satisfaction value. (A) Denial Origin with Negative Contributions | (B) Satisfaction Origin with Positive Contributions.

We can attempt an inferential analysis through a 2×4 ANOVA in which the first factor is the representation group and the second is the configuration, i.e., each of the four possible combinations of origin and link quality – the latter is also treated as a repeated measures factor. The test offers no effect for representation and no effect for interaction thereof with model configuration. The effect of the configuration itself however was found to be statistically significant (Pillai $F(3, 31) = 11.8, p < 0.001$). A view of the cell means can be seen in Figure 12. If we perform Bonferroni ad-

justed pairwise paired t-tests we find differences as per Table 8: the *Denial-Negative* configuration is distant from all other configurations ($p < 0.01$).

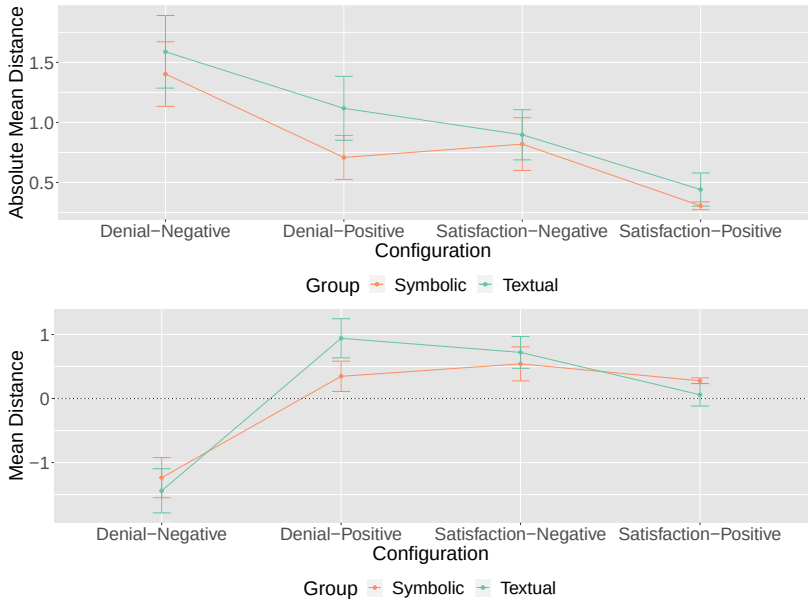

Fig. 12 Mean distance and absolute mean distance interaction plots for round 1.

	Denial-Negative	Denial-Positive	Satisfaction-Negative
Denial-Positive	0.000		
Satisfaction-Negative	0.000	1.000	
Satisfaction-Positive	0.000	0.313	0.293

Table 8 p-values of the pairwise comparison between configurations (both groups). Coding format [Origin Quality]-[Link Quality]

The results suggest that origin denial, combined with negative contribution links, leads quite certainly to less accuracy than all the other categories. However, the representation style (*symbolic* vs. *textual*) does not seem to matter. In the following experimental rounds we, hence, switched focus to the symbolic representation style, as featured in the original *i** publications, and moved on to perform a similar comparison with numeric representations.

6.3.2 Rounds 2 and 3: Numeric vs. Symbolic

In rounds 2 and 3 we repeat the same exercise with the second student group and the two Mechanical Turk groups. The modes of representation under comparison are

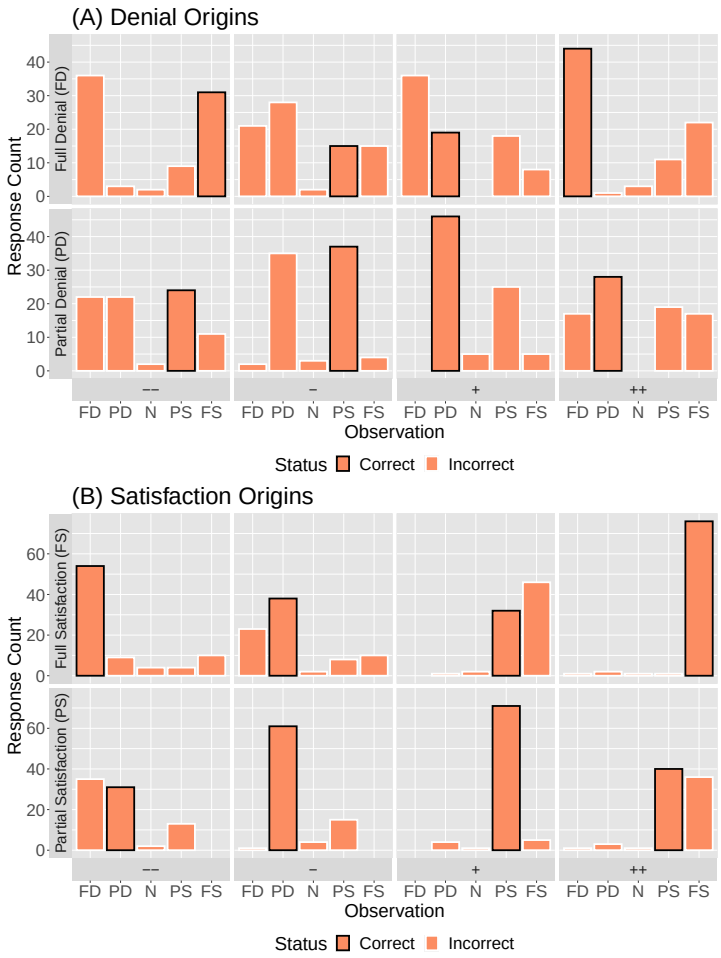

Fig. 13 Response counts for round 2 and 3 per contribution label and origin satisfaction value (symbolic models). (A) Denial Origins | (B) Satisfaction Origins.

now the symbolic against the numeric. A visualization of the data can be seen in Fig. 10(B). Symbolic representation follows the same pattern observed in round 1: accuracy substantially decreases when denial and/or negative contribution are featured in the diagram. The same is less true with numeric models. Qualitatively, this lack of accuracy is overestimation in all cases except, again, in the case where origin denial and negative contribution are combined as seen in Figure 10(B). A look at the corresponding distributions of responses for symbolic data can be seen in Figure 13, where exactly the pattern of non-detection of satisfaction reversal is observed when the origin goal is partially or fully denied (upper graph (A) – responses cover the entire range) but not when the origin goal is partially or fully satisfied (lower graph (B) – responses are concentrated to one side of the graph).

We again perform a 2x4 ANOVA as before: between factor is group (representation style) and within factor are the four configurations – i.e., combinations of link and origin qualities. We find a significant main effect on configurations – Pillai $F(3, 157) = 28.2, p < 0.001$ as well as an interaction – Pillai $F(3, 157) = 3.22, p = 0.024$. The result can be seen in Figure 14 where mean distance is measured in both as-is and as absolute value. It can specifically be seen that for both representations, configurations including denied origin are the least accurate, particularly when contribution is negative. The interaction is further studied through simple effects analysis [68] of the group factor, after fixing configuration levels. Out of the four simple effects tests, configurations *Denial-Negative* (Wilcoxon $W = 2488, p = 0.011$) and *Satisfaction-Positive* ($W = 2275.5, p < 0.001$), are the ones achieving statistical significance ($\alpha = 0.05/4 = 0.0125$) observed also in Figure 14. However, effects are small (0.2 and 0.26, respectively) and do not lend themselves to any useful interpretation. On the other hand, Bonferroni-corrected post-hoc pairwise paired t-tests over the within-subjects factor (*configuration*) can be seen in Tables 9 and 10, for fixing group level to symbolic and numeric respectively. The difference between the *Denial-Negative* and all other configurations is salient indicating, again, the denial inversion problem.

	Denial-Negative	Denial-Positive	Satisfaction-Negative
Denial-Positive	0.00		
Satisfaction-Negative	0.00	0.02	
Satisfaction-Positive	0.00	0.00	0.43

Table 9 p-values of the pairwise comparison between configurations for the *symbolic* group. Coding format [Origin Quality]-[Link Quality]

	Denial-Negative	Denial-Positive	Satisfaction-Negative
Denial-Positive	0.00		
Satisfaction-Negative	0.00	0.00	
Satisfaction-Positive	0.00	0.06	0.32

Table 10 p-values of the pairwise comparison between configurations for the *numeric* group. Coding format [Origin Quality]-[Link Quality]

6.4 Quantitative Theory Adoption

We now explore what method respondents of the numeric group follow to arrive to the satisfaction value they report. The goal is to understand the mental operation (if any) that participants perform with the origin satisfaction value and the contribution label – e.g., in Figure 3 how the number -0.3, the origin goal satisfaction, and the number 0.4, the contribution label, are combined by the participant to calculate the satisfaction level of the destination. We are particularly interested to see if participants perform

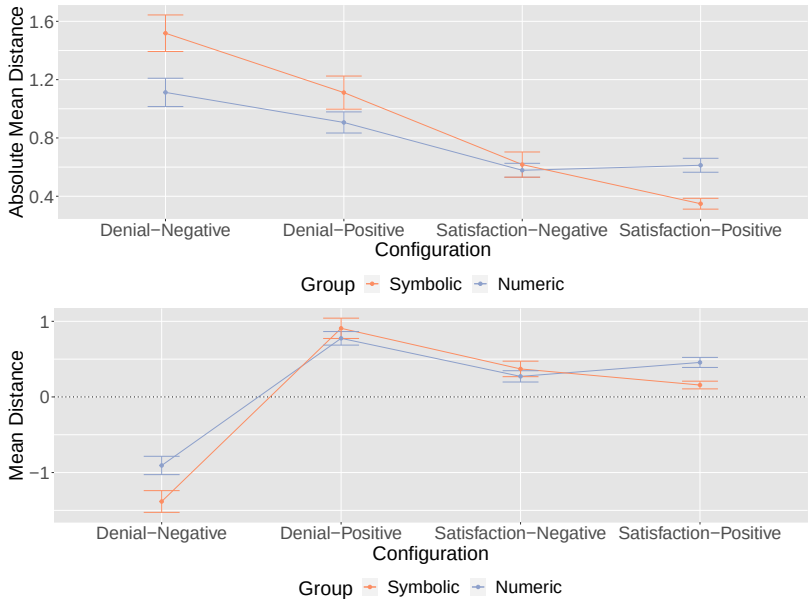

Fig. 14 Mean distance and absolute mean distance interaction plots for rounds 2 and 3.

any of the candidate calculation approaches described in Section 2, namely addition (so, in the example above $-0.3+0.4=0.1$ or $0.3+0.4=0.7$), multiplication (0.12 or -0.12), minimum (-0.3 or 0.3), maximum (0.4). We thus allow for participants to ignore or misuse negative signs, as long as the operation they perform on the absolute values matches the hypothesized one. We decide that the participant has used one of the operations if their response is 0.02 or less away from the corresponding normative value.

Based on the above design, in most cases, we cannot strongly associate the response to a specific operation, assuming instead that, participants predominantly offer an intuitive value or choose some other operation not covered here. Recall, for comparison, that in the decision problems of Section I there is no evidence that calculations are taking place. Table 11 offers the distribution of number of participants who consistently (at least three out of the four times in each configuration), followed an identifiable calculation on the absolute values. Thus, several of the participants followed an addition or subtraction approach, followed by some adoption of multiplication.

	Origin Qual.	Link Qual.	Add or Subt.	Mult.	Min	Max	Other
1	Satisfaction	Positive	23%	14%	5%	0%	58%
2	Satisfaction	Negative	31%	15%	0%	5%	49%
3	Denial	Positive	32%	10%	0%	5%	53%
4	Denial	Negative	21%	10%	0%	0%	69%

Table 11 Calculation method per origin satisfaction and contribution type.

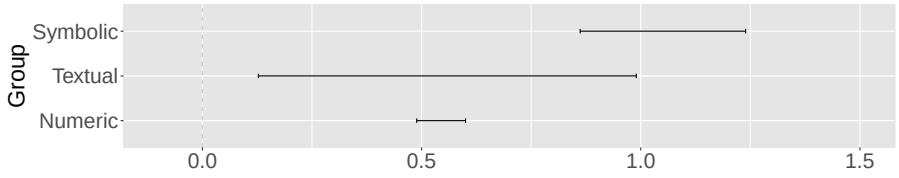

Fig. 15 t-test confidence (Bonferroni adjusted) intervals on the existence of a difference between positive and negative contribution links in the average reported destination satisfaction when origin satisfaction level is **N** or **0**.

6.5 Zero Satisfaction Analysis

We finally turn our focus to the cases in which the origin goal has no satisfaction or denial; in other words, it is marked as **N** (symbolic, textual) or **0** (numeric). In such cases, any satisfaction propagation framework would assume that the destination goal should be marked with zero satisfaction. However, that does not appear to be the case in the data. In Tables 12 and 13 we see the average assessed satisfaction value of the destination observed for rounds 1 and 2 & 3, when the origin satisfaction level is zero. In all cases and independent of representation mode, when the link is positive it appears to somehow imply satisfaction of the destination goal while when the link is negative it implies denial. Participants therefore tend to, to some extent, see contribution links are generators of satisfaction or denial rather than mere propagators. The observation is statistically significant – Figure 15 presents t-test confidence intervals.

Group	Break/−−	Hurt/−	Help/+	Make/++
Symbolic	-0.44	-0.22	0.22	0.44
Textual	-0.35	-0.24	0.12	0.41

Table 12 Observed satisfaction level for destination goal when origin goal is N or 0 (round 1).

Group	Large Negative/−−	Large Negative/−	Small Positive/+	Large Positive/++
Numeric	-0.22	-0.16	0.16	0.56
Symbolic	-0.57	-0.32	0.48	0.90

Table 13 Observed satisfaction level for destination goal when origin goal is N or 0 (rounds 2 & 3).

6.6 Section II: Summary of findings

Let us summarize the findings of Section II of the experiment. In terms of hypotheses we fail to reject $\mathbf{H_0^{II,1}}$ for the comparison between textual and symbolic models. We reject it for the comparison between symbolic and numeric, but the effect is small and

not suitable for useful generalizations: the two representations appear to each be more suitable compared to the other for *Satisfaction-Positive* and *Denial-Negative* configurations, respectively. More importantly, the *Denial-Negative* configuration appears to offer a much lower accuracy score due to what we identified as the satisfaction/denial inversion problem, i.e., failure to assume conversion of a satisfaction (resp. denial) value of the origin goal of the link to a denial (resp. satisfaction) value for the destination of the link, due to a negative contribution label. Hence $\mathbf{H_0^{II,2}}$ is rejected. The finding emerges descriptively in agreement data as well. Through further analysis, we find that the majority of respondents of the numeric group do not follow an easily identifiable numeric calculation approach, though some seem to have adopted some version of addition, subtraction or multiplication. Finally we find that participants assign satisfaction or denial to the destination despite the absence of satisfaction or denial in the origin goal, due to simply the presence of a positive or, respectively, negative contribution link.

7 Design Implications, Validity Threats, and Limitations

7.1 Summary of Findings and Language Design Implications

We summarize the key observations from our various analyses in Table 16. We summarize the results in relation to our original research questions as follows.

Firstly, considering RQ1.2, users appear to adopt specific (though hard to precisely elicit and describe) methods for exploring the decision structure of the goal model, and such methods may have some common characteristics (e.g., top-down navigation, simple local comparisons). The methods, however, may not be compatible with the normative semantics designed by researchers for the purpose of e.g. automatic reasoning. In other words, methods for visually navigating a presentation of a decision problem may need to be designed distinctly from and in addition to methods for automated generation of optimal solutions not meant to be used by humans.

The use of numbers (RQ1.1) appears to allow for more consistent reasoning compared to symbols. However, this may be because it so happens that the ad-hoc methods adopted by participants are compliant in the particular examples with the authoritative method, without however the two methods necessarily being the same. As a design implication, it may, thus, be ideal that such compliance is by design rather than by coincidence. An interesting future exploration, for instance, would be to devise decision problem visualizations whereby the natural way of exploring them leads to results that are guaranteed to be consistent with, e.g., the label-propagation theories we took up in this study [30,31].

Secondly, the role of individual differences (RQ2.1, RQ2.2) turned out much less important than we originally conjectured. Cognitive style, specifically, measured by CSI, does not appear to be relevant to the phenomena in question, including, to our surprise, the choice to work methodically or intuitively. It follows that, either the choice of index is sub-optimal – alternative measures, such as the Rational-Experiential Inventory (REI) [24] have been proposed – or that the cognitive work needed to perform the tasks in question is not within the scope of the cognitive style

Key Findings

Section I: Decision Problems

- The majority of participants adopt a specific method for performing decision making inferences on the diagram.
- Participants appear to predominantly follow a top-down approach exploring paths from top level criteria to decisions, making path choices and ultimately decisions based on simple local comparisons (numeric and symbolic models) or cumulative weight (e.g. symbol counts) accumulated along the path (symbolic models).
- Compared to using symbolic weights, using numerical weights evokes a diagrammatic reasoning method that is more compatible with the corresponding normative one (weighted summations), without evidence however that the latter is precisely recognized and followed.
- Cognitive Style, measured by the CSI, does not have a large predictive power over the phenomena in question.
- Math anxiety, as measured by the AMAS index, has a small effect in successfully performing the diagrammatic reasoning exercises and is more pronounced in numeric models.
- No evidence emerges that complex mental arithmetic is taking place in reasoning with the diagrams.
- Participants are generally confident of their responses despite the inconsistency of accuracy scores.

Section II: Individual Contributions

- There does not appear to be substantial difference between using symbols, such as $++$, $-$, etc., words, such as *makes*, *hurts*, etc. or numbers, such as 0.5, -0.7 etc., for labeling single contribution links, in terms of identifying a satisfaction value for the destination of a link, given the satisfaction level of its origin.
- Whenever negative contribution is combined with denial of the origin participants fail to agree on the satisfaction quality of the destination.
- Participants misunderstand contribution as a generator of satisfaction (or denial) rather than a mere propagator thereof.
- Most participants of the numeric group do not appear to perform any kind of arithmetic. When they do this tends to be some form of addition, subtraction or, less likely, multiplication - not minimum, or maximum.

Fig. 16 Key Findings from Section I and II analysis

construct – e.g., they are too low-level. AMAS's small effect shows that the particular construct may affect diagrammatic reasoning in general, especially if the latter includes numbers. The effect, however, may be too low to be significant part of a design process or future investigation.

7.2 Implications to Modeling using Current Languages

The results of the study may help improve diagrammatic practices even when utilizing the current goal visualization languages. To see how, we focus on symbolic and textual contribution annotations such as “+", “–" or “helps" and “breaks". These are highly desirable in many cases in which a rough idea of the contribution structure needs to be conveyed and/or when systematic measurements (e.g. application of AHP comparisons) are not available or practical. As we saw in Section 2, their abil-

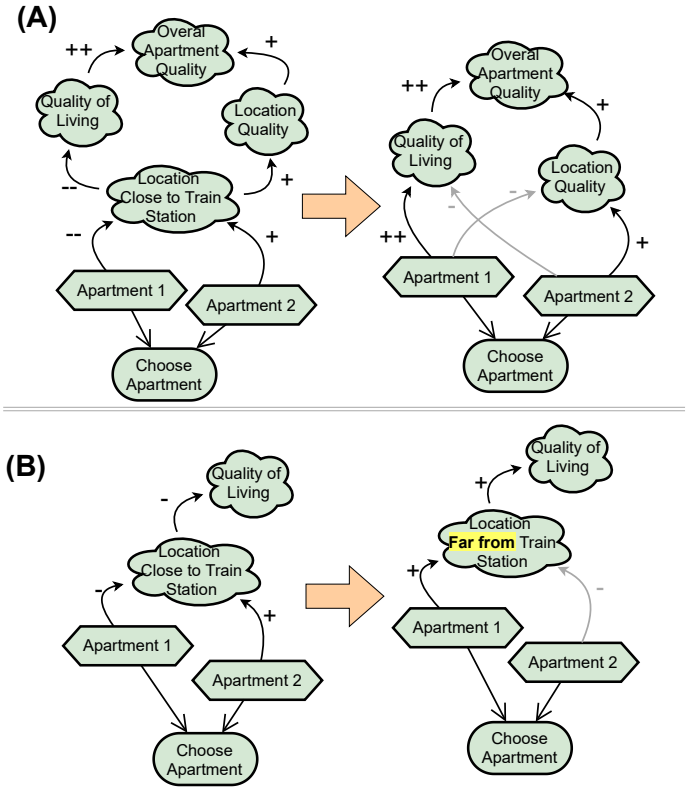

Fig. 17 Chain shortening (A) and semantics reversal (B) examples

ity to allow for intuitive diagrammatic reasoning in simple models such as those of Figure 1 is highly compelling. However, our study suggests that the presence of negative contribution links and the emergence of satisfaction denial can be detrimental to compliant reasoning with more complex models. Thus, if we follow a diagramming approach that avoids these elements while preserving meaning, diagrams can become more amenable to accurate diagrammatic reasoning. As examples, subject to future evaluation and formalization, we sketch four possible guidelines that may help achieve that:

– **Chain Shortening.** Our results revealed participants' difficulty in interpreting negative contribution links and goal denial values that these links result in. Such problems will tend to emerge when negative links appear in chains of contribution links, i.e., series of quality goals each contributing to the next. In the Figure 17(A), left side, an analysis of the problem of choosing between two apartments is presented. In the specific example, one option is far from the train station and the other one is close. Closeness to the train station may have conflicting qualities: it may be a noisy, lowering *Quality of Living*, but it allows quick access to transportation, supporting *Location Quality*. Considering *Apartment 1*, two chains are

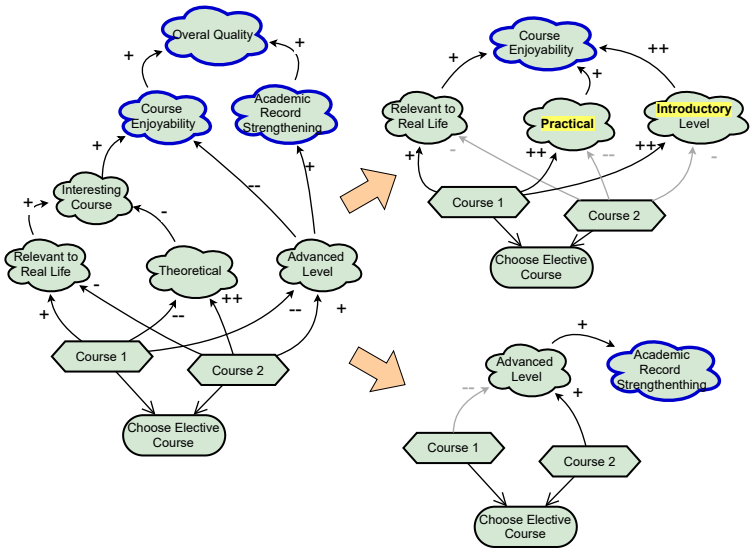

Fig. 18 Slicing example.

formed, one consisting of two negative contribution links and a positive one and one consisting of one negative and two positive ones. Our evidence suggests that it is likely that users of the diagram will be confused with regards to how they should combine the negative links in the chain. Recall for example that some participants work along paths and count the number of symbols. Following such technique they may for example infer that *Apartment 1* hurts *Quality of Living*, due to the number of negative symbols along the corresponding path, when, according to semantics, it actually helps it.

To increase the chance of accurate reasoning, the modelers may prepare a simplified version of the diagram, such as that of the right side of Figure 17(A). The quality *Location Close to Train Station* has now been removed, and the first two contributions have been replaced by one that aggregates them according to semantics. The implications of each decision are now intuitively clearer – at the cost of removing the intermediating goal and its explanatory function. It is, hence, likely that symbol counting participants will more readily select *Apartment 2* on account of the stronger contribution of *Location Quality* to the root goal.

– **Reverse goal semantics.** It is often the case that to eliminate a negative contribution it suffices to reverse the semantics of a goal. In Figure 17(B) left side, the negative contribution between the two quality goals is the source of two problematic chains. However, if we reverse *Location Close to Train Station* with its dual *Location Far from Train Station*, and to preserve model semantics, reverse all incoming and outgoing contribution links, we arrive at a model (Figure 17(B) right side) in which only one of the chains contains a negative link, making the optimal easier to spot.

- **Slicing.** In Figure 18, left-side, a problem with two criteria is presented. The context here is to choose an elective university course. The student has two choices with different qualities, ultimately contributing to top level goals *Course Enjoyability* and *Academic Record Strengthening* (how good the course looks on the student's record).
  Our results suggest that the representation is not really amenable to accurate diagrammatic reasoning, unless it is somehow simplified into one or more models that avoid negative contributions in addition to being smaller. Analysts may first observe that the top-level contributions do not offer much to the decision problem; they merely suggest that the two sub-qualities are equally important. Hence, the analysts may decide to remove the top goal and split the model into two separate ones. In each of the latter, the above guidelines can be used to further simplify them. By looking at the models of Figure 18 right-side it is quicker to understand the impact of each course to each of the two important qualities.
- **Avoid negative contributions**. A final technique to make models more comprehensible is to avoid the emergence of negative contributions, which we saw invite inaccuracy. We can achieve that via assuming by default and when possible, that the worst possible contribution toward a goal is no contribution. Let us go back to the model of Figure 17(B), right side. In that model, *Apartment 2* is next to the train station and *Apartment 1* far from it. From an optimal decision viewpoint it makes no difference to say that *Apartment 2* denies the goal *Location Far from Train Station* – i.e., causes it to have a negative satisfaction value, versus saying that *Apartment 2* has no contribution whatsoever to the same goal. In either case, we make an assessment of how close a distance needs to be from the station for the goal to be deemed not only not satisfied, but even worse (denied). In most contexts where symbolic goal models are used, such as, for example, sketching decision problems during early requirements, such assessment is rarely based on concrete information or method. Meanwhile, as far as the decision problem is concerned, *Apartment 1* is still the preferred choice. We may, thus, choose to avoid the negative contribution link. In all Figures 17(A), 17(B), and 18, negative contributions that can be eliminated without harm to the corresponding decision problem representation are grayed out. Note that the avoidance of unnecessary negative contribution links and, subsequently goal denial, does not contravene Giorgini et al. semantics.

Guidelines such as the above are rather informal and in need for further specification and formalization into concrete rules that allow systematic transformations that also make equivalence guarantees between the original and transformed representations. They show, however, the kind of follow up work that the evidence from our study inspires, aimed at making goal models more useful visual instruments.

7.3 Validity Threats and Limitations

We now turn to validity threats of our study, focusing specifically on construct, internal, external, and statistical conclusions validity.

With regards to *construct validity* a central question is the validity and usefulness of our main quality concept, intuitive comprehensibility appropriateness, both itself as a theoretical construct and with respect to the ways we operationalize it. We define the theoretical construct on the basis of the traditional understanding of comprehensibility – in our case defined as leading to model activation that is consistent with language designer and modeler expectations – specialized to further demand that the participants have no prior training to the modeling notation at hand. At the theoretical level, the assumption is that there are representations that make better use of users' prior experiences and knowledge than others. For example, we hypothesized that users are more comfortable with reading and manipulating numbers than idiosyncratically defined symbols, as they are familiar with the former from their daily lives, but have never seen the latter before.

At the operationalization level, measuring intuitiveness through observing intuitive reactions of untrained participants – instead of educated choices after complete training – naturally follows the theoretical definition. Training participants to the exact method would not allow us to detect any prior participant expectations and inclinations, as participants would simply execute the third-party method they learned; i.e. the training itself would become a strong confounding factor. One can, however, hypothesize that even in the full training scenario, error frequencies and response time discrepancies may offer indications of the sought intuitiveness: representations and (imposed) methods in which participants take longer or make many mistakes may indirectly indicate unintuitive choices. Future studies may attempt this strategy while carefully: (a) defining the construct being measured which is now more akin to a form of learnability, (b) dealing with training and its quality as a nuisance variable.

Further, the use of accuracy for measuring comprehensibility appropriateness directly follows from the definition of the latter as the level of agreement between the user and designer vis-à-vis the meaning of language constructs, via comparing observable inferences the two parties make. A caveat is that, as we saw, such agreement may be coincidental, that is, although inferences agree, the underlying meaning and thought process, which are unobservable, may be different. This difference may or may not reveal itself in different sets of examples.

Two comments can be made with respect to this last concern. Firstly, if we are restricted to observation of inferences, there appears to indeed be no guarantee that we can ever completely learn if participants and designers follow the exact same mental model; our confidence only increases as we consider more and more varied examples. Secondly, there is a pragmatic benefit in simply measuring observable model activation: even if the mental models of participants are very different than those of the designers, it is still useful to know that they are such that the majority of inferences will coincide. This was observed in our results: participants are unlikely to have precisely followed a weighted summation approach to make decisions in the numerical models. Whatever method they used, however, seems to have properties that make it lead to the same answers as the aforementioned method. The analogy with mental models is salient here: users may form and employ only an incomplete or surrogate [75,98] model of the actual reasoning technique, which is nevertheless compliant with the latter. This may be acceptable in practice.

The above discussion is crucial also from an *internal validity* standpoint, which is concerned with the claims of causal relationships between variables. Thus, while the numeric models appear to lead to better accuracy, as we saw, there might be other factors at play than the representation format per se, including the specific normative reasoning approach attached with the representation. It is, hence, the combination of representation and authoritative reasoning approach that is understood to bring about the result, specifically in the decision problems. Future work can investigate different such combinations, such as for example, the application of AHP-style decision making where numbers are discretized as symbols. Given the wealth of such options, however, the space of possible experiments is large.

An additional internal validity concern is that of training. Participants in our experiments do attend some training videos in which they are presented to the concept of goal models and contribution links, so that they can perform the exercises. They are told, for example, that + is a positive contribution, or that a larger number implies a stronger contribution. As we saw, however, this training does not discuss any specific method for making the complex decisions or combining origin satisfaction and contribution label to decide destination satisfaction level. Nevertheless, despite the care that we took to keep that information hidden, the way by which we abstractly described contributions could affect participant behavior. Furthermore, effort is made for the training material between the two groups to be as similar as possible: the same narration, voice, models, visuals, video length etc., with necessary differences only when the contribution annotations are different. We find that detecting biases in a training process, even when it is highly controlled (e.g. use of videos rather than live lectures), is a non-trivial matter, addressed primarily through replications with different training approaches.

The same difficulty emerges when we perform transformations in order to make the two representation approaches, symbolic and numeric, comparable. This primarily affects model generation for Section I, where we needed to arrange so that the symbolic and numeric version of each of the 12 decision models allow for fair comparison, via keeping the distance between best and second best alternative consistent across the models. In all cases, we needed to use our judgment with regards to the appropriateness of the coding and transformation procedures employed to make the two representation approaches comparable without favoring one of the two. In Section II, for example, accuracy and agreement distances for numeric models are preceded by discretization, so as to control for the advantage that numeric models may have due to their expressiveness and allow for a fair comparison with symbolic ones. As with training, however, replications with alternative coding procedures may be needed to explore the sensitivity of such procedures to bias.

Further, some obvious *external validity* concerns can be raised with regards to sampling of both participants and models. Firstly, to appreciate the rationale for participant sampling, (students and Mechanical Turk participants), one needs to think of the population of supposed users of goal model visualizations. While goal models have been designed to be used primarily by requirements analysts [100], the decisions that they can represent are really ones of arbitrary stakeholders. Hence, rather than being a tool for exclusive use by analysts, goal models are much more attractive for adoption in the requirements analysis practice when the stakeholders themselves can

use the visualizations to explore and understand the decision problems themselves. It is, thus, reasonable to expect that goal models aspire to offer visualizations that make them usable to a wide range of decision-making professionals that can be involved as stakeholders in a requirements analysis process in a variety of domains. While there are no statistics on the exact profile of such a participant, we can assume that this population is ultimately bound to primarily include people who have finished high-school, and most likely attended a few years of University. Hence, samples from the student population or the on-line participant pool with university degree qualifications appear appropriate for this investigation.

A more pertinent external validity threat is the sampling of goal models. While we have created 24 of them in Section I, we imposed certain structural constraints (e.g. one decision only, distance between the best and second best is fixed, specific layouts, colors, shapes, fonts, etc.) that may be limiting their representativeness. As we saw, measuring comprehensibility of a model does not amount to a measurement of the comprehensibility appropriateness of the language that was used to construct it [61]. Rather, diverse samples of models need to be tested prior to making statements about the language. As such, replications with different models will be needed to address the inherent pragmatic limitations of a single experiment.

An additional threat is also the size of goal models. In practical applications, goal models are meant to be used for organizing large numbers of goals and their in-between interactions (tens or often hundreds, see [38]), which raises the question whether our small experimental models generalize to such realistic models. A first comment is that if a certain kind of representation is ineffective for small models, it is not problematic to also assume that such ineffectiveness also emerges in larger models. For example, phenomena such as difficulty in combining denial of the origin goal with a negative contribution link, or erroneously ascribing non-zero satisfaction to a goal that is targeted only by goals with zero satisfaction, are not expected to correct themselves if we increase model size. They are rather pointing to foundational design/visualization choices that need to be attended to prior to exploring larger models. Secondly, even large goal models are likely to contain a number of decisions, in the form of OR-decompositions, that can be dealt with separately as smaller problems. Each such decision problem typically includes not all but a subset of relevant quality goals. For example, even in the small goal models of Figure 1 it can be observed that *Have Trip Booked* is a separate decision from *Have Expenses Reimbursed* and the first decision is concerned with only two of the three quality goals. Hence, even in large goal models, the need to visually reason with small or medium size slices thereof is usually pertinent. Finally, to experiment with larger models is to investigate an activity – unguided visual reasoning against large and complex models – that stakeholders will unlikely engage in in the first place. When models are large and cannot be compartmentalized as above, rather than unguided visual reasoning, it is more appealing to use – and hence study the effectiveness of – alternative visualization techniques – e.g. [53], guided evaluation – e.g., [40], or automated reasoning – e.g., [56,60]. Hence, while generalization of our findings to larger models can indeed only be hypothesized given our results, such generalizations might be of limited practical use.

Furthermore, utilizing our observations to make general statements about the goal modeling and analysis frameworks utilized in this study (Giorgini et al. [30], GRL [5], Liaskos et al. [54]) is not supported by our methodology. As we saw, simplifying assumptions needed to be made for comparisons to be possible and only subsets of corresponding modeling languages were utilized. For example, in the decision problems, numeric contribution links do not feature arbitrary weights in the [-100,100] range, as proposed by GRL [5], and symbolic contribution labels do not distinguish between propagation of satisfaction or denial as in the original framework [30]; e.g., $+$ vs. $+_S$ and/or $+_D$. Rather than evaluating these frameworks, our study focuses on the effect of specific design decisions (choice of label representation and meaning) for specific tasks (visual reasoning) over small and medium size models, in order to guide future investigation and notation design efforts.

One final comment on external validity concerns possible generalizations beyond goal models to cover conceptual models in general. Although our study was not designed for such, its results may offer useful indications of investigative directions that are or are not worth pursuing. One is the question whether CSI is a predictor of effectiveness or style for diagrammatic reasoning (in any diagram). Our results discourage hypotheses that this may be the case, without however excluding a role for CSI or other cognitive style index in, e.g., developing models, or choosing one representation or model development approach over another. A second is the method adoption construct, in which some participants operate intuitively and others adopt a specific method. This may occur in any kind of model when participants are given freedom as to how they should work with the model. In our results, the majority of participants did adopt a concrete method. This seems to suggest that mental models is a possible theoretical basis on which we can talk about diagrammatic reasoning in general, especially when intuitiveness is the main subject – i.e., the evocation of a way of working with the model.

Finally, in terms of *statistical conclusion validity*, a point of discussion may seem to be the way we approach different samples and administration rounds in the analysis. One my be consider that each of these rounds constitutes a member of a family of experiments [82], and analyze each separately followed by meta-analysis. However, in our case, the changes to the instrument are minimal and restricted to reordering the mental math exercises. The models and the task remain exactly the same and so is the response variable. The sample origin (students vs. Mechanical Turk) may be argued to be a candidate for some effect. As we saw, we chose to originally include those variables (round and sample) as additional factors and proceeded with separate treatment only if those factors turned out to be relevant. That happened once in Section I, where students were treated separately from Mechanical Turk participants.

8 Related Work

The role of problem representation in decision making has long been known to be important in the literature, as representations both help decision makers understand the problem at hand [78] and may actually influence the corresponding decision [44, 46, 65]. Several approaches to visualizing multi-criteria decision problems specifically

have been proposed with a focus on representing alternatives and their impact on criteria: tables, treemaps [9], value paths, parallel coordinate plots [29,69] as well as a variety of more interactive and specialized approaches such as WeightLifter [76], Grower Plots and Decision Balls [66] among others. Efforts for empirical evaluation of decision support visualizations have also been reported, such as, for example, Stone and Schade who compare numeric versus textual attribute values for the evaluation of alternatives [91], or Dimara et al. [22] who study parallel coordinate graphs, scatterplot matrices, and tabular visualizations. At the same time, a wealth of individual studies on comprehensibility of conceptual models exist in the literature – see Houy et al. [41] for an earlier survey and a presentation of the comprehensibility construct problematic – while, more recently, the general problem of systematizing evidence-based notation design in conceptual models has attracted increasing attention from researchers. Bork and Roelens, for example, offer a technique based on iterative evaluation and improvement of notations [13].

Research has also focused on the relevance of cognitive fit theory [96] in predicting which visualizations will work best for a task at hand, e.g. [42,53,64,87,94]. The role of individual differences has also been studied. For example in a study by Engin and Vetschera [23], CSI is reported to be a predictor of suitability of graphical versus tabular representations, while Luo et al. [64] use the verbalizer-visualizer questionnaire [48] to obtain a similar result.

As we saw, goal models have long been considered to be tools for effectively guiding decision problem understanding and exploration [74] via a variety of formal, semi-formal or visual analysis approaches. Gonzàlez Baixuali et al. for example [32] propose a tool for visualizing qualities of goal model alternatives through a variety of techniques including pie-charts, bar-charts and tree-views. Horkoff and Yu propose a way to semi-automatically evaluate satisfaction propagation, whereby model users intervene to resolve conflicts [40]. Many other ways to reason about goal satisfaction propagation and thereby resolving goal alternative selection have been proposed in the literature, e.g. [5,51,55,57,60] – Horkoff and Yu offer a survey [39].

Despite the wealth for proposals for reasoning with goal models, efforts for empirical exploration of such proposals are limited in number. Horkoff and Yu, for example, perform an evaluation of their own proposal [40] while Hadar et al. [34] report on a family of studies in which goal diagrams and use case diagrams are compared on a variety of user tasks, such as reading and modification. In a similar vein, Abrahão et al. [1] present an empirical comparison of *i** with a specialization of GRL [99] called value@GRL, and, through similar experimental practices, Morales et al. [72, 73] compare *i**, KAOS (a goal modeling language [20]) and TRiStar (an extension to *i** for teleo-reactive systems). Elsewhere, Teruel et al. [92] compare again *i** with an extension thereof for collaborative systems requirements. The role of representation becomes the subject of a study by Caire et al. [15] where, using Moody's "physics of notations" [71] as motivating theory, symbols used to represent goal modeling constructs are the result of participant selection. In a similar vain, aimed at improving the semantic transparency of *i**, Santos et al. compare the standard visualization with an alternative one [85]. Tasks included answering comprehension question after studying a model and identifying issues in defective models and metrics included accuracy, speed, and ease, the latter assessed with the assistance of eye tracking. Sim-

ilar work has been done by the same group on KAOS goal models [20, 84]. Despite these efforts, however, to our knowledge, no work reports empirical effort focusing exclusively on the comprehensibility of contribution links for decision making.

9 Conclusions and Future Work

The ability of goal models to represent and support decisions [74] is arguably one of their most appealing properties that makes them potentially valuable tools for every stage of the IT planning and development lifecycle where decisions and tracking of their rationale is involved. Hence, we consider evidence-based optimization of their utility as visual aids to be a worthwhile research program. The study we presented is meant to be used as a starting point for further empirical investigation aimed at, firstly, informing the design of goal model based notations and decision support visualizations and techniques and, secondly and more generally, developing new or utilizing and advancing existing empirical constructs (e.g., intuitiveness) and theoretical approaches (e.g., mental models) to allow systematic study of modeling notation design, beyond goal models.

With regards to goal model-specific research, we are interested in exploring novel visual representations that are consistent with the more expressive semantics that have been proposed for contribution links, so that formal reasoning is more explainable and transparent. In earlier work [53] we showed, for example, that simple bar-charts and pie-charts are, under specific circumstances, better tools for helping users identify the correct – according to weighted summation semantics – optimal alternatives compared to diagrams. It is, thus, possible that there is a visualization that is optimal for conveying the semantics of label propagation, which, as we saw, are not always served well by the current diagrammatic notation. Further, investigation can go beyond static visualizations and into more interactive decision space exploration experiences. This may also allow for measuring intuitiveness of the specific steps of formal procedures. For example, a step-by-step interactive execution approach, such as that proposed by Horkoff and Yu [40] where users intervene to resolve conflicts resulting from the application of formal rules, can also be implemented as a stepwise evaluation of the rules of formal reasoning themselves. Thus, instead of training users to a given predefined reasoning mechanism, the latter is specially designed to fit intuitive expectations of the former.

Furthermore, we plan to continue to study methodological aspects and particularly the interaction between comprehensibility appropriateness, training, and learnability, both within and outside the context of goal models. As we discussed above, the process of measuring the former is confounded by adequate application of the latter: with sufficient training, any notation can become comprehensible, one may claim. Intuitiveness as discussed here, becomes then a function of the amount of training needed to reach a fixed level of comprehensibility or, reversely and as implemented here, a measure of comprehensibility that is reached after a fixed amount of training. Sound ways to measure training “amounts” will be, hence, needed. Moreover, at the measurement and data collection level, our experience in this study underlines the importance of free-form verbalization as a way of contextualizing the observa-

tional data. We plan to integrate such components in future studies focusing not only on written retrospective comments but also oral ones offered during performance of the activity [86]. Finally, the introduction of questionnaire-style measures of comprehensibility, analogous to widespread standardized instruments utilized in interaction design such as SUS [14] or TAM [21], can allow for more reliable assessment and potentially for a more refined theoretical model of comprehensibility.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available in the York University's Dataverse repository: `https://borealisdata.ca/privateurl.xhtml?token=ea6aaabc-7ab1-4c77-98e3-a567d1a46184`. (private draft link for review purposes – embargo to be lifted upon publication of this report).

Compliance with Ethical Standards

Data collection has received ethics review and approval by the Human Participants Review Sub-Committee, York University's Ethics Review Board (certificate # e2018 - 024 and subsequent renewals) and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines.

References

1. Abrahão, S., Insfran, E., de Guevara, F.G.L., Fernández-Diego, M., Cano-Genoves, C., Pereira de Oliveira, R.: Assessing the effectiveness of goal-oriented modeling languages: A family of experiments. Information and Software Technology **116**, 106171 (2019). DOI https://doi.org/10.1016/j.infsof.2019.08.003. URL https://www.sciencedirect.com/science/article/pii/S0950584919301673
2. 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). DOI 10.1111/j.1467-6486.1996.tb00801.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6486.1996.tb00801.x
3. Allinson, C.W., Hayes, J.: The Cognitive Style Index: Technical Manual and User Guide. Tech. rep. (2012). URL https://www.cognitivestyleindex.com/_files/ugd/832dc3_0da009728f13415dbc4e08daa0943f8f.pdf
4. Amazon Mechanical Turk (2022). URL https://www.mturk.com/
5. 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)
6. Amyot, D., Mussbacher, G.: User Requirements Notation: The First Ten Years, The Next Ten Years (Invited Paper). Journal of Software (JSW) **6**(5), 747–768 (2011)
7. Armstrong, S.J.: The Influence of Individual Cognitive Style on Performance in Management Education. Educational Psychology **20**(3), 323–339 (2000). DOI 10.1080/014434100750018020. URL https://doi.org/10.1080/014434100750018020
8. Armstrong, S.J., Qi, M.: The Influence of Leader-Follower Cognitive Style Similarity on Followers' Organizational Citizenship Behaviors. Frontiers in Psychology **11** (2020). DOI 10.3389/fpsyg.2020.01265. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01265
9. Asahi, T., Turo, D., Shneiderman, B.: Using Treemaps to Visualize the Analytic Hierarchy Process. Information Systems Research **6**(4), 357–375 (1995). DOI 10.1287/isre.6.4.357. URL http://dx.doi.org/10.1287/isre.6.4.357

10. Ashcraft, M.H.: Math anxiety: Personal, educational, and cognitive consequences. Current Directions in Psychological Science **130** (2002). DOI 10.1111/1467-8721.00196
11. Ashcraft, M.H., Kirk, E.P.: The relationships among working memory, math anxiety, and performance. Journal of Experimental Psychology: General (2001). DOI 10.1037//0096-3445.130.2.224
12. Blazhenkova, O., Kozhevnikov, M.: The new object-spatial-verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology **23**, 638–663 (2009)
13. Bork, D., Roelens, B.: A technique for evaluating and improving the semantic transparency of modeling language notations. Software and Systems Modeling **20**(4), 939–963 (2021). DOI 10.1007/s10270-021-00895-w. URL https://doi.org/10.1007/s10270-021-00895-w
14. Brooke, J.: SUS: A quick and dirty usability scale. Usability Evaluation In Industry **189** (1995)
15. 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). DOI 10.1109/RE.2013.6636711
16. Chandler, D.: Semiotics: The Basics, 2nd edn. Routledge (2007)
17. Corbin, J., Strauss, A.: Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory. Sage Publications (2012). DOI 10.4135/9781452230153
18. 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). DOI 10.1371/journal.pone.0057410. URL http://dx.doi.org/10.1371%2Fjournal.pone.0057410
19. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. The Computing Research Repository (CoRR) **abs/1605.0** (2016). URL http://arxiv.org/abs/1605.07767
20. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-Directed Requirements Acquisition. Science of Computer Programming **20**, 3–50 (1993)
21. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly **13**(3), 319–340 (1989). URL http://www.jstor.org/stable/249008
22. Dimara, E., Bezerianos, A., Dragicevic, P.: Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support. IEEE Transactions on Visualization and Computer Graphics **24**(1), 749–759 (2018). DOI 10.1109/TVCG.2017.2745138
23. Engin, A., Vetschera, R.: Information representation in decision making: The impact of cognitive style and depletion effects. Decision Support Systems **103**, 94–103 (2017). DOI https://doi.org/10.1016/j.dss.2017.09.007. URL https://www.sciencedirect.com/science/article/pii/S0167923617301744
24. 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 (1996). DOI 10.1037/0022-3514.71.2.390
25. Eric Yu Paolo Giorgini, N.M., Mylopoulos, J.: Social Modeling for Requirements Engineering. MIT Press (2010)
26. Evans, C., Harkins, M.J., Young, J.D.: Exploring teaching styles and cognitive styles: evidence from school teachers in Canada. North American Journal of Psychology **10**, 567 (2008)
27. Figl, K., Recker, J.: Exploring cognitive style and task-specific preferences for process representations. Requirements Engineering **21**, 63–85 (2016)
28. 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). DOI 10.1016/j.datak.2007.09.011
29. Gettingera, J., Kieslingb, E., Stummerc, C., Vetscherad, R.: A comparison of representations for discrete multi-criteria decision problems. Decision Support Systems **54**(2), 976–985 (2013)
30. 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)
31. Giorgini, P., Mylopoulos, J., Nicchiarelli, E., Sebastiani, R.: Formal Reasoning Techniques for Goal Models. In: S. Spaccapietra, S. March, K. Aberer (eds.) Journal on Data Semantics I, pp. 1–20. Springer Berlin Heidelberg, Berlin, Heidelberg (2003). DOI 10.1007/978-3-540-39733-5_1. URL https://doi.org/10.1007/978-3-540-39733-5_1
32. Gonzales-Baixauli, B., Leite, J.C.S.P., Mylopoulos, J.: Visual variability analysis for goal models. In: Proceedings of the 12th IEEE International Requirements Engineering Conference (RE'04), pp. 198–207 (2004). DOI 10.1109/ICRE.2004.1335677
33. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. Ph.D. thesis, University of Twente (2005)

34. 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)
35. 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). DOI 10.1109/TSMC.1987.6499282
36. Hmieleski, K.M., Corbett, A.C.: Proclivity for improvisation as a predictor of entrepreneurial intentions. Journal of Small Business Management (2006). DOI 10.1111/j.1540-627X.2006.00153.x
37. 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). DOI 10.1177/1073191103010002008. URL https://doi.org/10.1177/1073191103010002008
38. Horkoff, J.: Using i* Models for Evaluation. Master's thesis, University of Toronoto (2006)
39. Horkoff, J., Yu, E.: Comparison and evaluation of goal-oriented satisfaction analysis techniques. Requirements Engineering (REJ) pp. 1–24 (2011)
40. Horkoff, J., Yu, E.S.K.: Interactive goal model analysis for early requirements engineering. Requirements Engineering **21**(1), 29–61 (2016). DOI 10.1007/s00766-014-0209-8. URL http://dx.doi.org/10.1007/s00766-014-0209-8
41. 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), vol. (LNCS 7532), pp. 64–77 (2012)
42. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems (2011). DOI 10.1016/j.dss.2010.12.003
43. Jalman, R., Aranda, J., Liaskos, S.: On Eliciting Preference and Influence Measure in Goal Models. Tech. rep., School of IT, York University, http://www.yorku.ca/liaskos/AHPGoals.pdf (2012)
44. Jones, D.R., Schkade, D.A.: Choosing and translating between problem representations. Organizational Behavior and Human Decision Processes **61**(2) (1995). DOI 10.1006/obhd.1995.1017
45. 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). DOI 10.1016/j.csi.2016.04.006
46. Kelton, A.S., Pennington, R.R., Tuttle, B.M.: The effects of information presentation format on judgment and decision making: A review of the information systems research. Journal of Information Systems (2010). DOI 10.2308/jis.2010.24.2.79
47. Kieras, D.E., Bovair, S.: The role of a mental model in learning to operate a device. Cognitive Science **8**(3), 255–273 (1984). DOI https://doi.org/10.1016/S0364-0213(84)80003-8. URL https://www.sciencedirect.com/science/article/pii/S0364021384800038
48. Kirby, J.R., Moore, P.J., Schofield, N.J.: Verbal and visual learning styles. Contemporary Educational Psychology (1988). DOI 10.1016/0361-476X(88)90017-3
49. Krogstie, J.: Model-Based Development and Evolution of Information Systems. Springer (2012)
50. Krogstie, J., Sindre, G., Jørgensen, H.: Process models representing knowledge for action: a revised quality framework. European Journal of Information Systems **15**(1), 91–102 (2006). DOI 10.1057/palgrave.ejis.3000598. URL https://doi.org/10.1057/palgrave.ejis.3000598
51. 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. ACM Press, Newport Beach, CA (2004). URL http://www2.info.ucl.ac.be/people/eletier/
52. Liaskos, S.: Replication Data for: On the Intuitive Comprehensibility of Contribution Links in Goal Models: An experimental study (2023). DOI 10.5683/SP3/T38E48. URL https://doi.org/10.5683/SP3/T38E48
53. 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). URL http://www.yorku.ca/liaskos/Papers/SAC2018/Visualizations/SAC2018.pdf
54. 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)
55. Liaskos, S., Khan, S.M., Litoiu, M., Jungblut, M.D., Rogozhkin, V., Mylopoulos, J.: Behavioral adaptation of information systems through goal models. Informations Systems (IS) **37**(8), 767–783 (2012)

56. Liaskos, S., Khan, S.M., Mylopoulos, J.: Modeling and reasoning about uncertainty in goal models: a decision-theoretic approach. Software & Systems Modeling **21**, 1–24 (2022). DOI https://doi.org/10.1007/s10270-021-00968-w
57. 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)
58. Liaskos, S., Lapouchnian, A., Wang, Y., Yu, Y., Easterbrook, S.: Configuring Common Personal Software: a Requirements-Driven Approach. In: Proceedings of the 13th IEEE International Requirements Engineering Conference (RE'05). IEEE Computer Society, Paris, France (2005)
59. Liaskos, S., Litoiu, M., Jungblut, M.D., Mylopoulos, J.: Goal-based behavioral customization of information systems. In: H. Mouratidis, C. Rolland (eds.) Proceedings of the 23rd International Conference on Advanced Information Systems Engineering (CAiSE'11), pp. 77–92. Springer Berlin Heidelberg, London, UK (2011)
60. Liaskos, S., McIlraith, S., Sohrabi, S., Mylopoulos, J.: Representing and reasoning about preferences in requirements engineering. Requirements Engineering Journal (REJ) **16**, 227–249 (2011)
61. Liaskos, S., Mylopoulos, J., Khan, S.M.: Empirically Evaluating the Semantic Qualities of Language Vocabularies. In: A.K. Ghose, J. Horkoff, V.E.S. Souza, J. Parsons, J. Evermann (eds.) 40th International Conference on Conceptual Modeling (ER 2021), *Lecture Notes in Computer Science*, vol. 13011, pp. 330–344. Springer (2021). DOI 10.1007/978-3-030-89022-3_26. URL https://doi.org/10.1007/978-3-030-89022-3%5C_26
62. 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 (2017). URL http://www.yorku.ca/liaskos/Docs/ESEM17.pdf
63. Liaskos, S., Tambosi, W.: Factors Affecting Comprehension of Contribution Links in Goal Models: An Experiment. In: A.H.F. Laender, B. Pernici, E.P. Lim, J.P.M. de Oliveira (eds.) Proceedings of the 38th International Conference on Conceptual Modeling (ER'19), pp. 525–539. Springer International Publishing, Cham (2019)
64. Luo, W.: User choice of interactive data visualization format: The effects of cognitive style and spatial ability. Decision Support Systems (2019). DOI 10.1016/j.dss.2019.05.001
65. Lurie, N.H., Mason, C.H.: Visual representation: Implications for decision making. Journal of Marketing (2007). DOI 10.1509/jmkg.71.1.160
66. Ma, L.C., Li, H.L.: Using Gower Plots and Decision Balls to rank alternatives involving inconsistent preferences. Decision Support Systems (2011). DOI 10.1016/j.dss.2011.04.004
67. 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)
68. Maxwell, S.E., Delaney, H.D.: Designing Experiments and Analyzing Data, 2 edn. Taylor and Francis Group, LLC, New York, MA, USA (2004)
69. Miettinen, K.: Survey of methods to visualize alternatives in multiple criteria decision making problems. OR Spectrum **36**(1), 3–37 (2014). DOI 10.1007/s00291-012-0297-0. URL http://dx.doi.org/10.1007/s00291-012-0297-0
70. Moody, D.L.: The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Transactions on Software Engineering **35**(6), 756–779 (2009). DOI 10.1109/TSE.2009.67
71. Moody, D.L.: The "Physics" of Notations: A Scientific Approach to Designing Visual Notations in Software Engineering. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 2, ICSE '10, pp. 485–486. ACM, New York, NY, USA (2010). DOI 10.1145/1810295.1810442. URL http://doi.acm.org/10.1145/1810295.1810442
72. Morales, J.M., Navarro, E., Sánchez, P., Alonso, D.: A controlled experiment to evaluate the understandability of kaos and i* for modeling teleo-reactive systems. Journal of Systems and Software **100**, 1–14 (2015). DOI https://doi.org/10.1016/j.jss.2014.10.010. URL https://www.sciencedirect.com/science/article/pii/S0164121214002143
73. Morales, J.M., Navarro, E., Sánchez, P., Alonso, D.: A family of experiments to evaluate the understandability of tristar and i* for modeling teleo-reactive systems. Journal of Systems and Software **114**, 82–100 (2016). DOI https://doi.org/10.1016/j.jss.2015.12.056. URL https://www.sciencedirect.com/science/article/pii/S0164121216000042

74. Mylopoulos, J., Chung, L., Liao, S., Wang, H., Yu, E.: Exploring Alternatives During Requirements Analysis. IEEE Software **18**(1), 92–96 (2001). DOI http://dx.doi.org/10.1109/52.903174
75. Norman, D.A.: Some Observations on Mental Models. In: Mental Models, pp. 7–14. Psychology Press (1983)
76. Pajer, S., Streit, M., Torsney-Weir, T., Spechtenhauser, F., Möller, T., Piringer, H.: WeightLifter: Visual Weight Space Exploration for Multi-Criteria Decision Making. IEEE Transactions on Visualization and Computer Graphics (2017). DOI 10.1109/TVCG.2016.2598589
77. Payne, S.J.: A descriptive study of mental models. Behaviour & Information Technology **10**(1), 3–21 (1991). DOI 10.1080/01449299108924268. URL http://dx.doi.org/10.1080/01449299108924268
78. Pracht, W.E.: Model visualization: Graphical support for DSS problem structuring and knowledge organization. Decision Support Systems (1990). DOI 10.1016/0167-9236(90)90011-F
79. Preece, Y.R., Sharp, H., Jennifer: Interaction Design: beyond human-computer interaction. Wiley (2011)
80. Rosnow, R.L., Rosenthal, R.: Beginning Behavioral Research: A Conceptual Primer, 6 edn. Pearson Prentice Hall, NJ, USA (2008)
81. Saaty, T.L.: Decision making with the analytic hierarchy process. International Journal of Services Sciences (IJSSCI) **1**(1), 83–98 (2008)
82. Santos, A., Gómez, O., Juristo, N.: Analyzing families of experiments in se: A systematic mapping study. IEEE Transactions on Software Engineering **46**(5), 566–583 (2020). DOI 10.1109/TSE.2018.2864633
83. Santos, M., Gralha, C., Goulão, M., Araújo, J., Moreira, A., Cambeiro, J.a.: What is the impact of bad layout in the understandability of social goal models? In: 2016 IEEE 24th International Requirements Engineering Conference (RE), pp. 206–215 (2016). DOI 10.1109/RE.2016.51
84. Santos, M., Gralha, C., Goulão, M., Araújo, J.: Increasing the Semantic Transparency of the KAOS Goal Model Concrete Syntax. In: J.C. Trujillo, K.C. Davis, X. Du, Z. Li, T.W. Ling, G. Li, M.L. Lee (eds.) Proceedings of the 37th International Conference on Conceptual Modeling (ER'18), pp. 424–439. Springer International Publishing, Cham (2018)
85. Santos, M., Gralha, C., Goulão, M., Araújo, J., Moreira, A.: On the Impact of Semantic Transparency on Understanding and Reviewing Social Goal Models. In: 2018 IEEE 26th International Requirements Engineering Conference (RE), pp. 228–239 (2018). DOI 10.1109/RE.2018.00031
86. Schweiger, D.M.: Is the simultaneous verbal protocol a viable method for studying managerial problem solving and decision making? The Academy of Management Journal **26**(1), 185–192 (1983). URL http://www.jstor.org/stable/256146
87. Speier, C.: The influence of information presentation formats on complex task decision-making performance. International Journal of Human Computer Studies (2006). DOI 10.1016/j.ijhcs.2006.06.007
88. Steinle, F.: Entering new fields: Exploratory uses of experimentation. Philosophy of Science **64**, S65–S74 (1997). URL http://www.jstor.org/stable/188390
89. Stoet, G.: PsyToolkit: A software package for programming psychological experiments using Linux. Behavior Research Methods **42**(4), 1096–1104 (2010). DOI 10.3758/BRM.42.4.1096. URL https://doi.org/10.3758/BRM.42.4.1096
90. Stoet, G.: PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. Teaching of Psychology **44**(1), 24–31 (2017). DOI 10.1177/0098628316677643. URL https://doi.org/10.1177/0098628316677643
91. Stone, D.N., Schkade, D.A.: Numeric and linguistic information representation in multiattribute choice. Organizational Behavior and Human Decision Processes (1991). DOI 10.1016/0749-5978(91)90041-Q
92. Teruel, M.A., Navarro, E., López-Jaquero, V., Montero, F., Jaen, J., González, P.: Analyzing the understandability of requirements engineering languages for CSCW systems: A family of experiments. Information and Software Technology **54**(11), 1215–1228 (2012). DOI https://doi.org/10.1016/j.infsof.2012.06.001. URL https://www.sciencedirect.com/science/article/pii/S0950584912001036
93. Türetken, O., Vanderfeesten, I., Claes, J.: Cognitive Style and Business Process Model Understanding. In: A. Metzger, A. Persson (eds.) Advanced Information Systems Engineering Workshops, pp. 72–84. Springer International Publishing, Cham (2017)
94. Umanath, N.S., Vessey, I.: Multiattribute Data Presentation and Human Judgment: A Cognitive Fit Perspective. Decision Sciences **25**(5-6), 795–824 (1994). DOI 10.1111/j.1540-5915.1994.tb01870.x. URL http://dx.doi.org/10.1111/j.1540-5915.1994.tb01870.x

95. Vance, C.M., Groves, K.S., Paik, Y., Kindler, H.: Understanding and measuring linear-nonlinear thinking style for enhanced management education and professional practice. Academy of Management Learning & Education **6**, 167–185 (2007). DOI 10.5465/AMLE.2007.25223457
96. Vessey, I.: Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. Decision Sciences (1991). DOI 10.1111/j.1540-5915.1991.tb00344.x
97. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal **3**(4), 217–237 (1993). DOI 10.1111/j.1365-2575.1993.tb00127.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.1993.tb00127.x
98. Young, R.M.: Surrogates and Mappings: Two Kinds of Conceptual Models for Interactive Devices. In: Mental Models, pp. 35–52. Psychology Press (1983)
99. Yu, E.S.: GRL - Goal-oriented Requirement Language. URL http://www.cs.toronto.edu/km/GRL/
100. 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)