Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment

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

Teodora Dundjerovic
Department of Psychology
York University
Toronto, Canada
teah34@my.yorku.ca

Grace Gabriel
Department of Psychology
University of Toronto
Toronto, Canada
race.gabriel@mail.utoronto.ca

ABSTRACT

Decision making is an important part of early requirements analysis. Analysts are faced with the task of describing a large number of solutions to stakeholder problems and assess each of them with respect to high-level objectives. Goal models are regarded to be a useful tool for assisting this decision making activity through explicating relationships among problems, criteria and solutions. To visualize goal models, box-and-line diagrams have traditionally been used. But are diagrams the best way to visualize decision problems within goal models? In this paper, we consider two alternative ways: tree-maps and common pie-charts combined with bar-charts. In an experimental study we evaluate each with respect to whether they can help participants identify optimal alternatives accurately, quickly and confidently. We find that, for at least one of the introduced visualizations, participants tend to respond faster, more confidently and more accurately than when using traditional goal diagrams. The result calls for rethinking the way we visually present goal models and for further experimentation on the matter.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in visualization**; *Visualization*; • **Software and its engineering → Requirements analysis**; • **Applied computing →** *Decision analysis*;

KEYWORDS

requirements engineering; requirements visualization; decision support; cognitive fit; treemaps; goal models

ACM Reference Format:
Sotirios Liaskos, Teodora Dundjerovic, and Grace Gabriel. 2018. Comparing Alternative Goal Model Visualizations for Decision Making: an Exploratory Experiment. In *Proceedings of ACM SAC Conference (SAC'18).* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/https://doi.org/10.1145/3167132.3167269

1 INTRODUCTION

Exploring alternative designs and making decisions is an important activity of requirements engineering. While eliciting high-level objectives of stakeholders, analysts need to devise, describe and compare various design solutions that can potentially meet these objectives. Most often, solutions have conflicting qualities: while one solution might fulfill one objective satisfactorily, it falls short of meeting another. The reverse may be true of an alternative solution. Analysts and their stakeholders are, therefore, in need of methods to represent and reason about such trade-offs and finally make a decision that is weighted by the most important objectives.

Goal models [3, 8, 31] have been proposed as an effective way of exploring decisions in requirements engineering problems [25]. Such models allow concise representation of a great number of alternative ways by which high-level objectives can be met. They also show how each alternative would affect, if chosen, the fulfilment of higher level stakeholder objectives. Various approaches have been proposed for modelling the contribution of alternatives to objectives [2, 8, 18, 31] and several techniques for formalization and automated reasoning of such contribution structures have been introduced (e.g., [2, 8, 12, 18–20]) that allow calculation of optimal solutions given expressed stakeholder priorities.

While an automated reasoner can provide us with lists of good solutions, its output alone might not be sufficient for allowing the participants of a requirements analysis process to comprehend the decision problem itself and its various parameters, so to, for example, better explain why the reasoner generates the particular solutions. Diagrammatic goal model representations have been considered to be potentially suitable for this problem exploration task [12]. Nevertheless, different ways of visualizing decision problems similar to the ones that goal models imply have long been introduced in the literature [23, 32]. At the same time, it is proposed that different ways of representing the same decision problem may lead to different levels of performance depending on the task at hand (e.g., [7, 29]). Thus, whether box-and-line goal diagrams are the best way for comprehending various aspects of the goal structure seems to be an open question.

In this paper we experiment with two alternative ways for visualizing decisions modeled with goal models. These are (i) a combination of a bar-charts and pie-charts and (ii) treemaps. We first discuss the two main quantifiable concepts that are relevant for comprehending decisions within goal models: performance of low-level goals with respect to high-level goals and relative importance of high-level goals. Then, we show how these quantities can be mapped to visual variables of the proposed visualizations, such as size, color and brightness.

To test whether the introduced visualizations merit consideration, we performed an experimental study with participants from

SAC'18, April 9-13, 2018, Pau,France

ACM ISBN 978-1-4503-5191-1/18/04…$15.00
https://doi.org/https://doi.org/10.1145/3167132.3167269

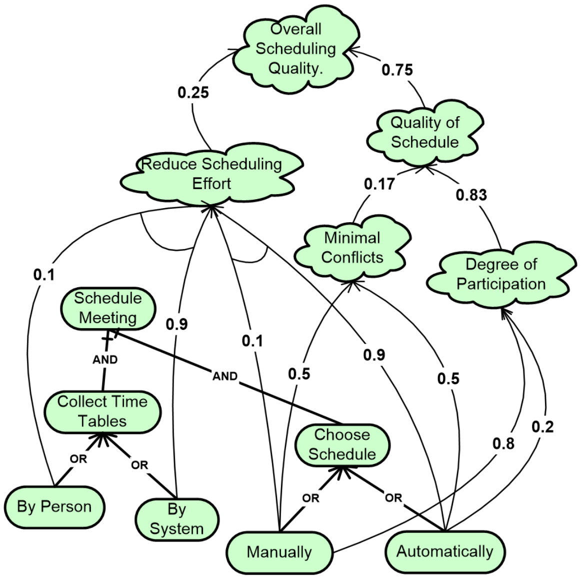

Figure 1: A Goal Model Represented as a Diagram

the University student community. We measured whether participants could identify optimal alternatives by simply looking at the corresponding visualizations. We found that for one of the introduced visualizations, the pie- and bar-charts, participants were in many circumstances more accurate than those who used traditional goal diagrams as well as quicker in making the decision. The results seem to indicate that visualizations of the sort we introduce in this paper can enhance an individual's ability to understand a decision making problem during requirements analysis, and can become a useful add-on for goal modeling tools, thereby increasing their appeal to the practicing analyst.

We organize the paper as follows. In Section 2 we show how goal models can be used for decision making. In Section 3 we describe our visualizations. In Section 4 we describe our experimental design and results. We discuss related work in Section 6 and conclude in Section 7.

2 BACKGROUND

2.1 Goal Models as Decision Structures

A goal model of the variety we consider in this study can be seen in Figure 1 – the example domain is meeting scheduling inspired by Mylopoulos et al. [25]. Alternatives are represented through an AND/OR decomposition tree of hard-goals (ovals). A separate hierarchy of soft-goals – goals that are not precisely defined, represented as clouds – acts as decision criteria. Each alternative of the AND/OR decomposition implies different types and levels of influence to these criteria. This is possible through *contribution links* that originate from the main AND/OR decomposition and target the soft-goal hierarchy.

A few approaches have been proposed with regards to the form and semantics of such contribution links [2, 8, 18, 31]. In the model of Figure 1 a quantitative approach very similar to that of URN [2] is followed. Contribution labels are drawn from the real interval [0, 1], 0 implying no contribution and 1 implying maximal satisfaction contribution of one goal to the other. The labels of the contribution links serve to represent the *share of contribution* each origin goal has to the satisfaction of the destination goal.

It has been shown [18, 21] that this restricted form is isomorphic to a family of AHP (Analytic Hierarchy Process) decision problems, where the soft-goal structure constitutes a common criteria hierarchy and each OR-decomposition a separate decision problem. As such, the problems of elicitation and aggregation of weights are reduced to the classic pair-wise AHP comparisons and subsequent construction of simple nested linear combinations. We have demonstrated that such systematic elicitation is feasible and scalable [18].

2.2 Computing Optimal Alternatives

Given the contribution labels of Figure 1 each solution of the AND/OR tree is associated with a level of satisfaction of each soft-goal in the soft-goal hierarchy, with respect to each OR-decomposition. Calculation is possible through traversing the soft-goal hierarchy from the bottom-up and linearly combining satisfaction values using the weights from the contribution links. More formally, let G be the set of all goals in the model, $C(g)$ be the set of elements g' such that there is a contribution link from g' to g. Let also $Sat(g) : G \mapsto [0, 1]$ represent the satisfaction value of goal g. The satisfaction value of a goal g is simply calculated as:

$$Sat(g) = \sum_{g' \in C(g)} (Sat(g') \cdot w_{g' \to g}) \qquad (1)$$

where $w_{g' \to g}$ is the weight of the contribution link from g' to g. An alternative, i.e. a solution of the AND/OR tree, comprises of a set of choices in every OR-decomposition of the tree, starting from the top. Considering each decomposition separately, when a choice, i.e., one of the OR-subgoals g_s, is considered as part of the alternative, we assign it maximum satisfaction value $Sat(g_s) = 1$, leaving every other choice in the OR decomposition with minimum satisfaction value 0. As such, a recursive application of formula (1) from the hard-goals of the OR decomposition upwards the soft-goal tree allows for calculation of the satisfaction value of every soft-goal for the choice. If there is a unique root soft-goal – e.g., *Overall Scheduling Quality* in our example – this value constitutes a global fitness score for the OR-decomposition choice. In the more common case in which there are many top level goals – i.e., the decision makers do not want to include an *Overall Scheduling Quality* root goal and thereby avoid commitment to a preference between *Reduce Scheduling Effort* and *Quality of Schedule* – each such calculated satisfaction value constitutes the fitness score for the OR-decomposition choice with respect to the soft-goal in question, to be used for comparison with other top level goals.

In Figure 1, for the alternative *Manually* in the *Choose Schedule* decomposition $Sat(Quality\ of\ Schedule) = 0.17 \cdot [0.5 \cdot Sat(Manually) + 0.5 \cdot Sat(Automatically)] + 0.83 \cdot [0.8 \cdot Sat(Manually) + 0.2 \cdot Sat(Automatically)] = 0.17 \cdot [0.5 \cdot 1.0 + 0.5 \cdot 0.0] + 0.83 \cdot [0.8 \cdot 1.0 + 0.2 \cdot 0.0]$, to be compared with $0.17 \cdot [0.5 \cdot 0.0 + 0.5 \cdot 1.0] + 0.83 \cdot [0.8 \cdot 0.0 + 0.2 \cdot 1.0]$ for alternative *Automatically*. In this way, by identifying such individual optimal choices for every OR decomposition in the diagram, we, in practice, construct the optimal alternative, i.e., solution of the AND/OR tree, for the entire model in a piecemeal fashion.

2.3 Performance and Importance

Looking more closely at the contribution links of the model of Figure 1 allows us to distinguish two kinds of such contributions. The first of these shows how, if at all, each hard-goal contributes to the satisfaction of the soft-goals. Consider a soft-goal g that receives a contribution from hard-goal g'. The weight $w_{g' \to g}$ of the contribution link from g' to g corresponds to what we will call the *performance* $pr(g', g)$ of the hard-goal g' with respect to soft-goal g.

A second kind of information encoded in the diagram is the relative importance of soft-goals. In the figure, satisfaction of soft-goal *Degree of Participation*, for example, is more important than satisfaction of soft-goal *Minimal Conflicts* with respect to the satisfaction of goal *Quality of Schedule*. Consider the path from a soft-goal g' to any of its ancestors g in the soft-goal hierarchy. Let also $w_1, w_2, \dots$ be the labels of each contribution link on that path. We define the *importance* $ir(g', g)$ of g' with respect to g as $ir(g', g) = \prod_i (w_i)$. For example, the importance of *Minimal Conflicts* with respect to *Quality of Schedule* is 0.17. The importance of *Minimal Conflicts* and with respect to *Overall Scheduling Quality* on the other hand is $0.17 \cdot 0.75 = 0.1275$. If the soft-goal with respect to which importance is calculated is the root goal, as in the latter example, then we refer to the *absolute importance* – or simply *importance* – $ir(g)$ of soft-goal g. Note that while different alternatives have different performance with respect to a soft-goal, the importance of the soft-goal remains the same irrespective of alternatives.

2.4 Symbolic and Spatial Representations

A goal diagram such as that of Figure 1 offers us one way for visualizing a goal model. The literature, however, seems to suggest that not all visualizations for such problems are equally effective. The theory of *cognitive fit*, in particular, suggests that humans perform cognitive tasks more quickly and more accurately when the way information is structured matches the structure of the task at hand [16, 29, 30]. According to the theory one can distinguish between two types of tasks. One is *symbolic tasks* in which individual data values are handled (searched for and extracted). In our case a symbolic task is to learn the exact value of a contribution link between two goals. Symbolic tasks are best supported using symbolic representations, tables being the most frequently cited one. Another type of tasks are *spatial tasks* which require high-level assessment of a set of data focusing on identifying relationships, making associations and interpolating values [29]. Spatial tasks are best supported by spatial representations such as graphical representations which visualize relationships between data values, allowing users make holistic assessments about the data.

The decision comprehension and exploration problem that goal models pose to users, is one of evaluating ways to achieve goals (good solutions) subject to higher level objectives. This task can be seen as spatial, in that multiple criteria importance and alternative performance data values need to be aggregated and compared with one another. In traditional goal diagrams, while goals and how they relate to one another are, of course, represented spatially/graphically, the contribution weights are, as we saw, represented symbolically. Users need to locate individual symbolic representations (i.e., contribution labels) and perform mental calculations to combine and compare them. There is, thus, room for a hypothesis that there is a mismatch between the task at hand and the representation, implying that there might be a visualization alternative to goal models (or a way to extend the diagrammatic one) that can allow users to be more effective in this task.

Several techniques for visualising decision problems have been investigated in the literature. Gettinger et al. for example [7] study parallel coordinate plots (PCPs) and heatmaps and empirically compare them to tables. They report mixed results with regards to the effect of representation, although PCPs were found to induce a significantly different decision process in the short term. Miettinen [23] demonstrates uses of bar-charts, spider-web charts and value paths (i.e., parallel coordinate plots) and other visualizations for representing alternatives and criteria. In bar-charts, which are of specific interest here, clusters of bars represent alternatives or criteria and the bars within the cluster represent criteria or alternatives, respectively. In either case, the length of a bar in the cluster represents the performance of the (resp. each) alternative with respect to each (resp. the) criterion. Treemaps [14] are also known to have been used for financial decision support [32] and for representing Analytic Hierarchy Process (AHP) problems [4]. In the latter case, the treemap includes both a hierarchy of the decision criteria and, at the leaf level, the alternatives themselves, all sized according to their relative importance and, in case of alternatives, performance. The representations we experiment with in this paper, namely *charts* and *treemaps*, borrow from these ideas as we describe below.

3 VISUALIZING ALTERNATIVES

3.1 Charts

In the first alternative visualization we use a combination of two types of popular graphs: pie-charts and bar-charts. For the sake of convenience, we refer to the combined visualizations as *charts*. Pie-charts are used to represent the relative importance of soft-goals. Bar-charts represent the performance of alternatives to the leaf-level soft-goals. An example of this visualization can be seen in Figure 2.

The pie-chart on the upper part of the figure consists of a set of concentric disks. Each disk represents a level of soft-goal decomposition. Hence, the disk at the center represents the top-level decomposition. As usual in pie-charts, the disk is divided in sectors the relative size of each of which is a representation of the relative importance of the soft-goals represented. The ring that surrounds the core pie-chart, offers more detail as to how satisfaction of each soft-goal of the core pie-chart is shared among its children in the decomposition. Again, each soft-goal g' is represented through a ring sector, whose size (in terms of its central angle) relative to the size of the corresponding inner sector represents the importance $ir(g', g)$ of the soft-goal g' with respect to its parent goal g. On the other hand the size of the sector relative to the entire disk is a representation of the absolute importance $ir(g)$ of the corresponding soft-goal g. Thus, as we move outwards from the center to the outer rings, we represent relative and absolute importance of soft-goals that occur lower in the soft-goal hierarchy. Note that each soft-goal is represented through a different color in the pie-chart. Moreover, children of a soft-goal are associated with a higher value of the same hue, i.e., are a lighter version of the color of the parent soft-goal.

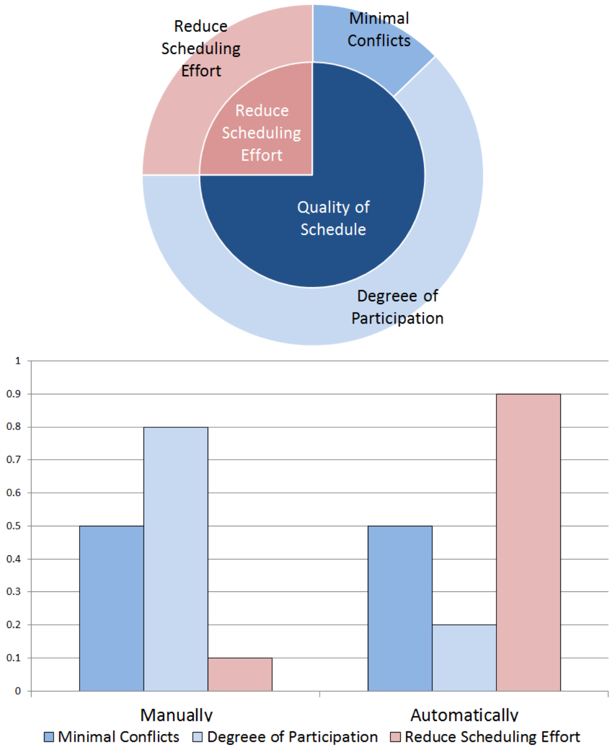

Figure 2: Bar- and Pie-chart

In Figure 2, the top level decomposition of the goal model of Figure 1 is represented through the central disk. As $ir(\textit{Quality of Schedule}) = 0.75$, the sector of the disk corresponding to that goal occupies 75% of the disk. The sector of the outer disk that aligns with the sector of *Quality of Schedule* is divided into two sub-sectors for the two children of *Quality of Schedule*, namely *Minimal Conflicts* and *Degree of Participation*. Given that $ir(\textit{Minimal Conflicts},\textit{Quality of Schedule}) = 0.17$ the corresponding sub-sector occupies 17% of the inner sector. On the other hand, the sub-sector occupies 12.75% of the entire disk, since $ir(\textit{Minimal Conflicts}) = 0.17 \cdot 0.75 = 0.1275$.

The bar-chart represents an OR-decomposition of the goal model and consists of clusters of bars. Each cluster is a representation of an alternative. Each bar in the cluster represents the performance of the alternative with respect to a soft-goal; the value of the performance represented by the length of the bar. The soft-goals that are referred to in each bar are identified by their color and order in which they appear in the cluster as well as using a legend. The color used to represent a soft-goal in the pie-chart matches the color with which the soft-goal is associated in the pie-chart. The bar-chart of Figure 2 represents the decomposition *Choose Schedule*. Focusing on the cluster that represents option *Manually*, the first bar represents the performance of *Manually* with respect to the soft-goal *Minimal Conflicts*, thus, $pr(\textit{Manually}, \textit{Minimal Conflicts}) = 0.5$ in agreement with the goal diagram (Figure 1).

3.2 Treemaps

A second visualization we consider is *treemaps* [14]. Two such treemaps can be seen in Figure 3. A treemap is essentially a rectangle containing other rectangles. Containment of one rectangle in another implies that the former is of a lower level than the latter in some conceptual hierarchy. In our case, each rectangle in the treemap represents a soft-goal in the soft-goal hierarchy and each rectangle that is contained in it represents a child of that soft-goal – which is also a soft-goal. When a soft-goal has no children in the soft-goal hierarchy the corresponding rectangle does not contain other rectangles. A label inside every rectangle is used to convey its association with the corresponding soft-goal.

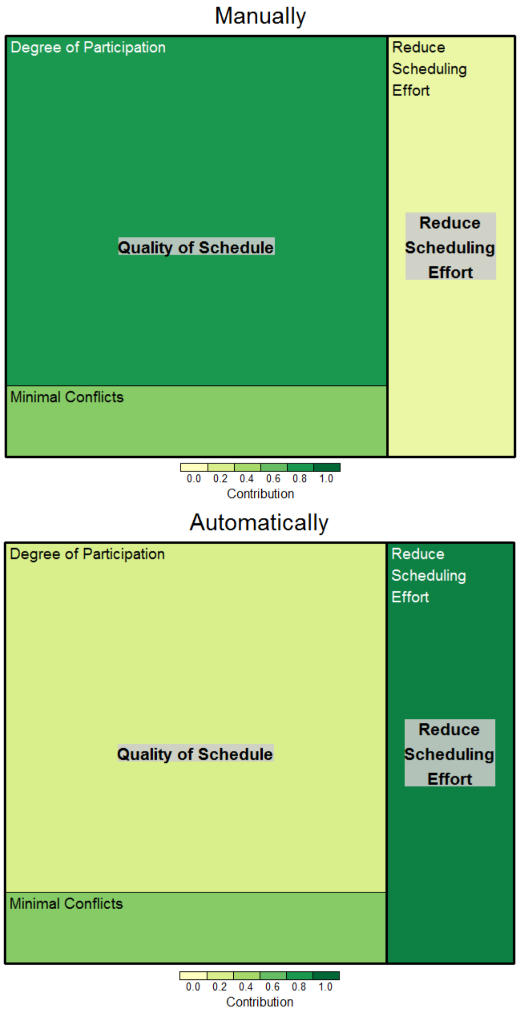

Figure 3: Treemap

The size (in terms of area) of a rectangle with respect to the size of its container rectangle is relevant in treemaps. In our case, the ratio of the size of a rectangle representing goal g' over the size of a containing rectangle representing higher-level soft-goal g is set to be $ir(g', g)$. Thus the relative size of a rectangle representing a goal compared to e.g. the outer rectangle or just its parent, is a representation of the importance of the goal or the relative importance with respect to the parent soft-goal, respectively.

In the treemaps of Figure 3, the rectangle *Quality of Schedule* consists of two smaller rectangles labelled *Degree of Participation* and *Minimal Conflicts*. The first occupies $ir($ *Quality of Schedule*$)$

of the treemap, i.e., 75% of the total area. The sizes of the latter contained rectangles with respect to the parent rectangle are respectively $ir(\textit{Degree of Participation}, \textit{Quality of Schedule}) = 0.83$ and $ir(\textit{Minimal Conflicts}, \textit{Quality of Schedule}) = 0.17$.

To represent a decision in an OR-decomposition we associate each choice of the OR-decomposition to a treemap and we draw and present all corresponding treemaps together. Color is used to represent the performance of each choice to each of the soft-goals of the treemap. Thus, the value (light or dark), hue, or both, of a leaf-level rectangle are associated with the performance pr of the choice the treemap represents with respect to the soft-goal that is represented by the rectangle. Hence, every OR-decomposition is represented by a group of as many treemaps as the choices of the decomposition, each with the same sizes of rectangles, since the soft-goal hierarchy is the same, but with different colors, as the performances of the choices with respect to soft-goals normally differ.

In Figure 3, an OR-decomposition of 2 alternatives is presented in form of treemaps. As such, there are two treemaps with same rectangle sizes and different colors. In the treemap representing choice *Manually*, rectangle representing *Degree of Participation* is significantly dark, implying that the alternative performs well with respect to the soft-goal. A legend shows the association of the value/hue level with pr. The corresponding rectangle in the treemap for choice *Automatically*, though, is light green, which according to the legend corresponds to a smaller pr value, as indeed is the case for that choice with respect to *Degree of Participation*.

For generating the treemaps we use M. Tennekes's implementation in R ([28]). For color, we associate a light yellow with $pr(\cdot,\cdot) = 0$ and dark green $pr(\cdot,\cdot) = 1.0$; values between those extremes are mixes between the two colors.

4 EXPERIMENTAL STUDY

4.1 Goals

To understand whether the two visualizations we introduce have merit with respect to perceiving optimal solutions compared to diagrams, we perform an experimental study with human participants. The goals of our study are three:

Accuracy. Assess which visualization leads to correct identification of the good quality alternatives, through simple exposure of participants to the visualization.

Efficiency. Understand the effect of visualization to the time it takes to perform the above task.

Confidence. Understand whether the degree of confidence of the participants with respect to their selection of good alternatives is affected by the presented visualization.

4.2 Experimental Design

4.2.1 Design Overview. To fulfill the above objectives, we first devised a number of models to be tested using different visualizations. The models come from three different domains: choosing an apartment to live, choosing a course and choosing a means of transportation. We chose these domains to ensure that our participants (University students) work on problems with which they have some familiarity, allowing the experimental tasks to be appreciated as realistic decision problems.

For each domain we devised three different model structures. The structures are of the form we discussed above, i.e, they all imply a ranking of alternatives from the best to the worst based on a top level root soft-goal. They all include one OR-decomposition. The three structures have different sizes: a small (6 soft-goals, 2 alternatives), a medium (7 soft-goals, 3 alternatives) and a large (9 soft-goals, 4 alternatives). For each structure we devised three randomly generated sets of weights, henceforth *weightsets*. During sampling of the weightsets we ensure that the difference between the total score of the optimal and the second optimal choice are in the interval $(0.09, 0.11)$ in small and medium size models and in $(0.19, 0.21)$ in larger models. We did this to avoid models whose optimal solution is very obvious, while for complex models specifically we keep solutions of the exercise attainable. Finally, the samples are such that they could have resulted from consistent (Consistency Ratio, CR≤0.1) AHP pair-wise elicitation processes. To acquire random weightsets fulfilling these criteria, specially designed search scripts were developed. For each combination of structure and weightset we developed each of the three visualizations. Thus, in total (3 domains)×(3 structures)×(3 weightsets) = 27 separate models were constructed, each represented in 3 different ways, goal diagrams, charts and treemaps, resulting in 81 separate visualizations.

We expose the models to the experimental participants in a *between-subjects* fashion with respect to the type of visualization. Thus, the participants are randomly split into three groups: the first works exclusively with goal diagrams, the second on charts and the third on treemaps.

The tasks the participants perform for each model are as follows. First, training videos about decision making (criteria and alternatives), the visualization they are assigned, and the domains are presented to them. Then they are shown the visualizations and for each they are asked: (a) to rank the alternatives in the model from the best to the worst and (b) to select how confident they feel about their response through a 4-level Likert-type scale (*Very Confident, Confident, Unconfident, Very Unconfident*). The time they spend on each model is also measured; no time limit is set. They repeat this task for each of the 9 models of each complexity level (small, medium, large). In other words, they repeat the task for all 27 available models in a given visualization.

Note that while domains and weightsets are provided in random order, complexity levels are given in fixed increasing complexity order. This obviously introduces an ordering effect, which is here deliberate: rather than making statements about the effect of model complexity in isolation, we want to examine whether the participants become quickly familiar with the visualization as they use it. Our question is whether increased familiarity with each visualization "beats" the increased level of complexity or the other way around, and whether there are differences in that process between visualization choices.

4.2.2 Participants and Groups. Participants are students of York University. Most of them were attending two 3rd year Information Technology classes from which they were recruited through announcements in class and mailing list. Additional participation was solicited from students outside the classes. A small monetary award ($35 CAD) was awarded to those with the most correct answers in the least amount of time. The instrument was administered in a classroom in afternoon or evening sessions with the authors supervising the process to ensure that participants focused on the task and were devoid of distractions.

5 RESULTS

5.1 Sample Characteristics

A total of 129 participants were recruited for this experiment. Of these, 13 participants did not understand the task or had difficulty using the task's drag-and-drop feature. This was assessed by a questionnaire following the instructions video. The remaining 116 cases are used for the analysis. They are 80 males and 36 females, their ages are mainly 21-29 and their field of study primarily Science, Technology and Engineering. Fields such as Humanities, Health Sciences and Social Sciences are also represented in the sample. Of all participants, 40, 38 and 38 are assigned to goal diagrams, charts and treemaps, respectively.

5.2 Analysis Approach

We analyze the data from our experiment using a standard 3 × 3 mixed-factorial design: one between-subjects (visualization choice) and one within-subjects factor (complexity level), each with three levels. First an overall effect is sought in each of the factors and their interaction. Wherever interactions are discovered, simple effects are sought by fixing the within factor (complexity) to a specific level and looking for differences between groups (visualization).

A summary of the acquired data can be seen in Figure 4. The F statistic is calculated in search for statistically significant ($p < 0.05$) effects. In the particular case of between group effects (overall comparisons of visualizations), due to the presence of heteroscedacity in the data, we apply Kruskal-Wallis' non-parametric test as well as Welch's W, which is a parametric modification of the F-test that is tolerant to heteroscedacity [22]. For the within-subjects factor (complexity level), we follow a multi-variate approach which does not require homogeneity of variance or sphericity [22]. In the case of simple effects, i.e. effects per complexity measure, heteroscedacity is dealt with transformations: accuracy measures are represented by their square root, time by its log, and the confidence measure by its third power. For simple effects, we generate confidence intervals for post-hoc comparisons. R is used for the analysis.

5.3 Accuracy

Recall that the experimental models are crafted in a way that the alternatives can be ranked from the best to the worst based on the evaluation technique we described in Section 2. Our first response variable reflects the degree by which experimental subjects identify this ranking correctly. We consider two different measures for accuracy. In the first one, which we call ranking identification, we calculate the Kendal Tau [17] correlation between the authoritative ranking of alternatives –i.e., the one calculated based on the evaluation formulae of Section 2 – and the ranking supplied by the participants. We calculate this for every response and add them up for each participant and complexity level – so (3 domains)×(3 weightsets) = 9 different numbers added. In the second one, we consider only the top alternative selection – i.e. we compare the authoritative best with the best identified by the participant ignoring the other alternatives.

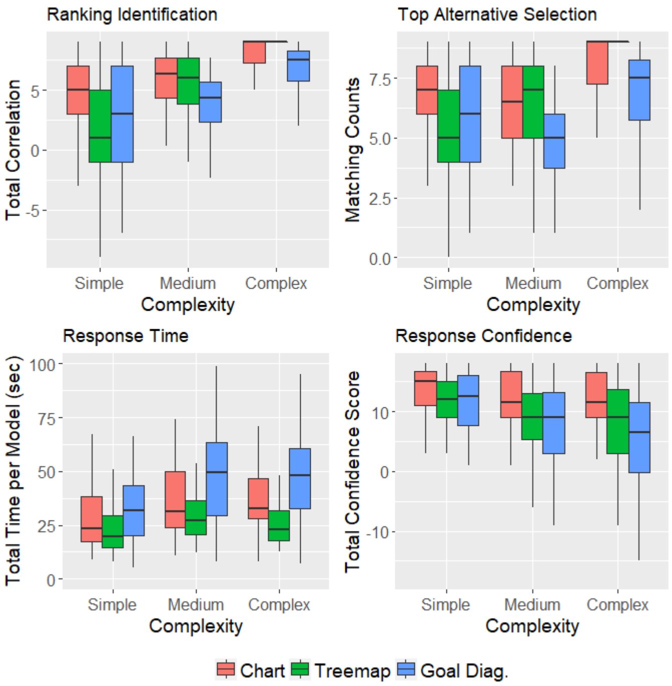

Figure 4: Box Plots of Main Measures

5.3.1 Ranking Identification. In terms of ranking identification, there are significant ($p < 0.05$) main effects of the chosen visualization – Kruskal-Wallis $H(2) = 13.36, p < 0.01$ (Welch's $W(2, 72.7) = 5.71, p < 0.01$), meaning that the choice of visualization affects the ability of participants to specify the rankings of optimal solutions. The level of complexity also has a very significant effect $F(2, 112) = 145, p < 0.01$ on ranking identification. Interestingly, visualization and complexity level seem to have a statistically significant interaction $F(4, 226) = 3.75, p < 0.01$. This means that the level of complexity seems to affect success rate in different ways for different visualizations.

To further examine these interactions, we follow the simple effects approach we described above in which we fix different levels of the complexity factor and study effects of the visualization factor. Considering simple models only, none of the spatial visualizations (treemap or chart) lead to better performance than the goal diagrams with significance – charts do so with marginal significance $p = 0.07$ (Dunnett post-hoc). Moving on to medium-size goal models however, we observe that charts become significantly $p = 0.04$ more effective than diagrams (Dunnett post-hoc). In complex models charts also appear to perform better than diagrams, albeit with $p = 0.08$ in the Dunnett tests, above our 0.05 threshold. The 95% family-wise confidence intervals of the left side of Figure 5 shed more light on these effects. *Chart, Tree* and *Diag.* represent charts, treemaps and goal diagrams, respectively. We observe that we can be reasonably confident that charts are consistently better than diagrams, while we remain inconclusive for treemaps.

The question of whether complexity level affects negatively success in rank identification has a negative answer. As Figure 6 demonstrates, this success increases with complexity for all visualizations,

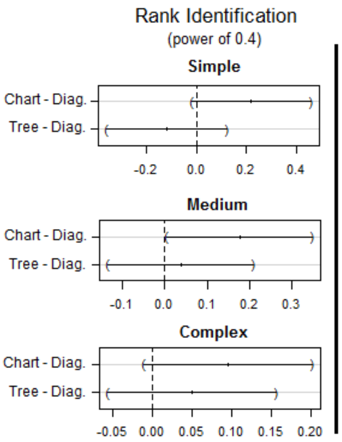

Figure 5: Confidence Intervals for Ranking and Top Identification

meaning that as participants get more and more familiar with the visualization, model size does not deter them from finding the correct answers.

5.3.2 Top Alternative Selection. If we restrict our focus to comparing how many times the participants' top choice matches that of the evaluation algorithm, we see similar results. There are significant main effects both due to the visualization – Kruskal-Wallis $H(2) = 19.27, p < 0.01$ (Welch's $W(2, 73.04) = 6.32, p < 0.01$) – and due to the complexity level $F(2, 112) = 46.57, p < 0.01$ as well as a significant interaction $F(4, 226) = 5.54, p < 0.01$.

Moving on to simple effects, in Figure 5, right side, the confidence intervals comparing visualizations for each complexity level can be seen. As with rank identification, for simple and complex models charts appear to be more effective than diagrams to a near-significant level. For medium complexity the effect is statistically significant.

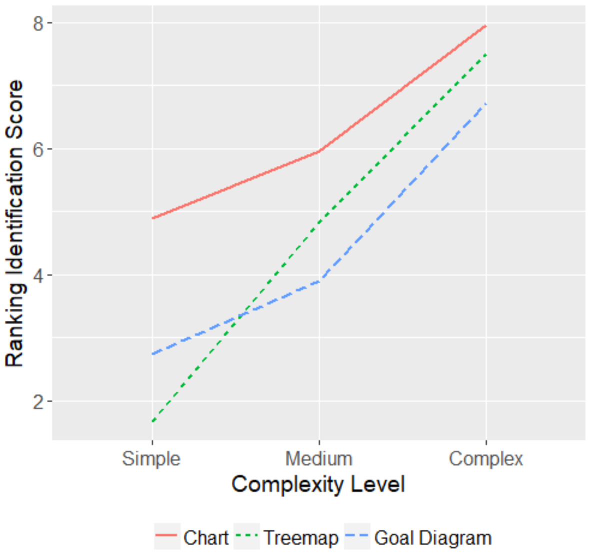

Figure 6: Ranking Identification: Visualization vs. Complexity

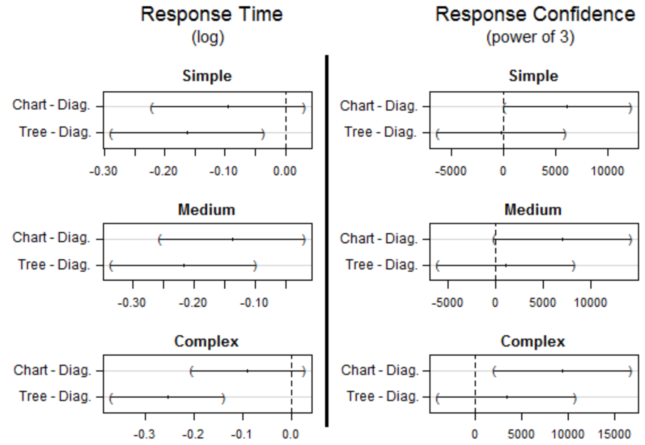

Figure 7: Confidence Intervals for Response Time and Response Confidence

5.4 Response Time

Response time is measured as the time difference between loading of the screen with the visualization and the question, and the time that the participant clicks to proceed to the next page. We add up the response times of the nine (9) tasks associated with each combination of visualization and complexity level, and perform our analysis using these totals.

Analyzing differences in response time across visualizations we also observe significant main effects due to the chosen visualization $H(2) = 44.14, p < 0.01$ (Welch's $W(2, 70.3) = 15.36, p < 0.01$) and due to complexity level $F(2, 122) = 11.71, p < 0.01$ as well as some interaction between the two factors $F(4, 226) = 2.39, p = 0.052$. Confidence intervals per complexity level can be seen in Figure 7. Participants generally respond with treemaps and charts quicker than with goal diagrams, and this can be claimed with statistical significance for treemaps. Figure 8 shows the effect difference in minutes, averaged for individual tasks. In non-simple cases Treemaps take nearly half as much time as goal diagrams and charts take from 2/3 to 3/4 as much time.

With regards to the effect of complexity, if we normalize response time for model size, measured using average number of contribution links in the model, the time spend per such link is 2.3, 1.93 and 1.54 sec/link for simple, medium and complex models, respectively. It, thus, reduces as model size increases. As seen in Figure 9, the effect is observed in all visualizations. If participants had followed precise mathematical or other systematic procedures for ranking the alternatives, we would have likely observed the opposite effect: time spent per link would at least stay constant, if not increase, as complexity of the overall task at hand increased. Instead, the decreasing time we actually observed allows us to hypothesize that participants seem to use cognitive processes of a lower level

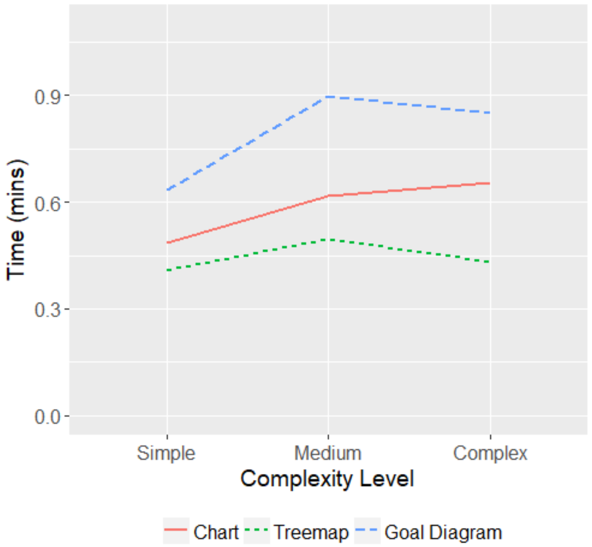

Figure 8: Response Times per Visualization and Complexity

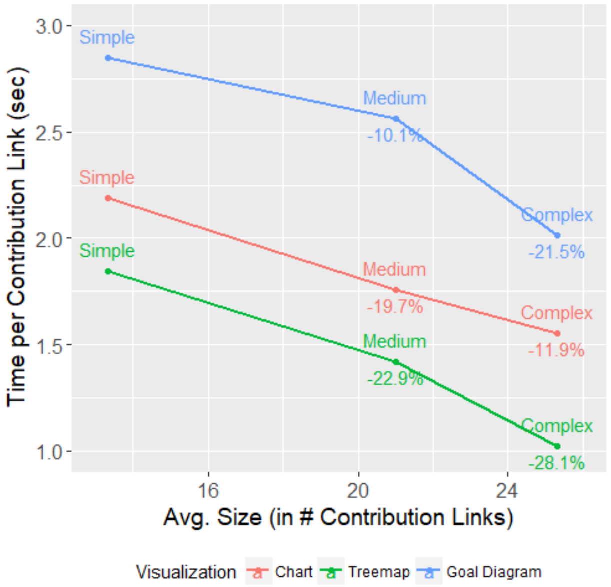

Figure 9: Response Times per Contribution Link versus Model Size (in # of Contribution Links)

when confronted with larger models. Moreover, as we saw, this strategy does not seem to compromise accuracy in any of the three visualizations.

5.5 Response Confidence

Participants confidence on their answer is acquired through a four-value rating scale. The responses are mapped to the values {-3,-1,+1,+3} which are in turn treated as an interval scale, as widely done in practice [27]. As in the case of accuracy and time, we sum up the nine (9) values provided for each combination of visualization

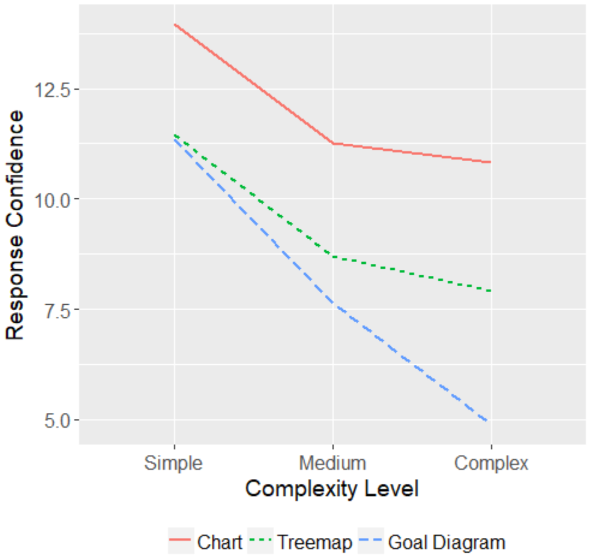

Figure 10: Response Confidence per Visualization and Complexity

and complexity level, and perform the analysis with the resulting total.

As with the previous measures, confidence also presents us with significant effects both due to visualization $H(2) = 19.92, p < 0.01$ (Welch's $W(2, 74.74) = 4.91, p < 0.01$) and due to complexity level $F(2, 112) = 30.5, p < 0.01$. There is no statistically significant interaction.

The 95% family-wise confidence intervals are seen in Figure 7. We observe that participants are in all cases more confident in their responses with the charts than with the goal diagrams. That cannot be said about treemaps. Finally, as complexity increases, despite the participants getting more familiar with the visualization, the confidence ratings drop (Figure 10). The differences between visualizations seem to amplify as complexity increases, with goal diagrams performing the worse.

5.6 Summary and Interpretation

We can summarize the above results as follows. As expected in theory, in almost none of the cases or measures does diagrammatic representation outperform the other two visualizations. Of the latter, it is charts that seem to lead to significantly more accurate responses, that are moreover arrived at in less time and with more confidence compared to goal models. Treemaps do not prove to be more accurate or confidence-inspiring than diagrams but elicit a response quicker. The superior performance of charts over treemaps can probably be hypothesized to be due to the familiarity of subjects to the former. Further, as model size increases, confidence in the response decreases irrespective of visualization. More surprisingly, accuracy increases with model size for all visualizations, which might be due to various factors including increased familiarization, or, more intriguingly, decreasing reliance on mental math or other high-level cognitive processes; the latter interpretation also not contradicting with the fact that response time normalized for model size declines as model size increases.

5.7 Validity Threats

Like any experimental study, ours is also subject to various validity threats, which we take up in this section.

5.7.1 Statistical Conclusion Validity. Our statistical analysis approach is based on Multivariate ANOVA (MANOVA) followed by univariate tests for specific effects; these tests assume normality, independence of individual errors and homogeneity of variance. The tests are known to be robust to violations of normality [22], especially when sample sizes are relatively large, as in our case [15]. We could further find no reason to believe errors might be dependent. Homogeneity of variance, nevertheless was an issue in some of our measures which we approached through transformations (checked through Levene's tests) or applications of Welch's W, which is tolerant to the phenomenon.

5.7.2 Internal and Construct Validity. Internal validity concerns our ability to establish causality between dependent and independent variables. As we saw, our choice to progressively increase complexity level instead of completely randomizing it within groups was conscious to allow us to test learnability of the visualization against complexity. The effect of fatigue, however, as an additional factor detrimental to performance vis-a-vis complexity level cannot be quantified and must be seen as a subject for future investigation. Nevertheless, a fatigue effect cannot affect our results with respect to visualization choice which is a between factor.

A factor with stronger potential to influence the between-subjects effect is the training videos. The video training method (versus live classroom sessions) was chosen to exactly control for this factor. The script of the narrator in the video is exactly the same across the three visualizations, with necessary differences to describe the different visualizations. The narrator's voice is also the same across groups. All videos use the exact same running example. Importantly, directions on how to make decisions using the given visualization were presented as high-level requirements rather than specific methods. Thus, the participants needed to find out by themselves how to make decisions using the visualizations. Despite these efforts to control for the effect, its absence cannot be guaranteed without a follow-up.

An additional point is the subjects' possible unequal familiarity with the visualizations. It can specifically be argued that participants were likely more familiar with charts than with goal diagrams at the time of the experiment – as the former are more commonly used in day-to-day life – which may be why the results tended to favor charts. We, however, see this as a plausible explanatory viewpoint, mostly threatening the cognitive fit explanation rather than the visualization's attractiveness: if charts bring with them the advantage of wide familiarity among decision makers, so be it, as long as they perform better in practical use.

In terms of construct validity, one should be conservative as to what our accuracy test measures. For instance, we cannot draw general conclusions about comprehensibility which is a more complex concept. The operationalization also does not measure decision quality, in that the participants do not *make* a decision but rather detect an optimal decision within a model constructed based on e.g., somebody else's preferences and performance evaluations.

5.7.3 External Validity. Generalization potential also deserves some discussion. We assume that our sample, mostly students of IT, is a good enough proxy of a population that may use goal models for making decisions. Although this population is commonly assumed to be requirements analysts, there is a reasonable aspiration that goal models are to also be used by arbitrary stakeholders; i.e., the sources of requirements and the final decision makers. We intuitively assume that the population of the latter is not much different compared to the one we draw from. We note that we could not find data on the average background of either practicing requirements analysts or real stakeholders/decision makers, on which to base a comprehensive assessment with regards to this validity aspect.

A more challenging generalization is on the types of goal models we consider, which have distinct features: they are hierarchical (acyclic), contain one OR-decomposition and one top-level goal. We, firstly, support that acyclic soft-goal hierarchies, which are an assumption of the evaluation theory we adopt, are capable of capturing useful and realistic soft-goal contribution structures. Adoption of a different satisfaction propagation and aggregation theory, such as that of Giorgini et al. [8], can indeed produce different results and need to be investigated in a future study. Furthermore, the presence of one root soft-goal – which aggregates a number of top-level goals hence reducing multiple criteria to one – and one OR-decomposition are simplifications necessary for maintaining experimental feasibility. Nevertheless, even for larger models with more OR-decompositions and many top-level goals, the essence of the decision making tasks is similar as far as visualization quality is concerned: analysis takes place one OR-decomposition at a time, with respect to, in turn, one top level goal at the time. Thus, the nature and difficulty of the decision-making tasks that the experimental models afford does not seem to be much distant from the ones that are afforded by larger practical models.

6 RELATED WORK

Visualization has been argued to be an important aspect of requirements engineering [1, 10]. The question regarding whether a diagramatic approach is effective in supporting comprehension, problem solving or other tasks has attracted considerable attention in the research community. For example, several studies investigate the comprehensibility of diagrammatic notations such as UML state diagrams or ER diagrams [6, 26].

In goal modeling specifically, there has been effort toward understand how goal diagrams can support comprehending the requirements problem and, by implication, making relevant decisions. González-Baixauli et al. introduced a tool for visualizing qualities and characteristics of alternatives within goal models [9]; the tool employs a variety of visualization techniques including pie-charts, bar-charts and tree views. Horkoff and Yu propose and evaluate an interactive evaluation of goal models, in which satisfaction of goals is propagated through the graph, requiring from the evaluator to intervene in cases of conflicting or partial satisfaction evidence [12, 13]. The authors also demonstrate possibilities for visually highlighting starting points of analysis and conflicts. The visual properties of goal modeling languages such as i^* have been the subject of investigation as well. Moody et al. [24] offer an analytical evaluation of the i^* visual syntax based on established rules for designing effective such (termed the "Physics of Notations"). An empirical analysis was followed by Caire et al. [5] in which experimental participants' comprehension of visual syntax is used for adapting the latter. Elsewhere, Hadar et al. [11] compare goal diagrams with use case diagrams on a variety of tasks, including model-reading and modifying. While the above efforts show interest in understanding the value of goal diagrams as visual artifacts, to our knowledge, no empirical investigation that compares diagrams with other ways to visualize goal models with a focus on decisions has been attempted to date.

7 CONCLUSIONS AND FUTURE WORK

We presented and evaluated two alternatives for visualizing decision problems represented as goal models. The introduced visualizations are based on the assumption that the task of mentally evaluating alternatives in goal models must match the way the relevant parameters are presented. We hypothesized that spatial visualizations – visualizations that make use of visual variables – will be more effective than symbolic ones – numbers – in accordance to cognitive fit theory. In the experiment to evaluate this claim we ask participants to identify optimal alternatives in different visualizations and measure the accuracy, time and confidence of their response. At almost no occasion do we find goal models to be more effective in either of the three measures. Instead, in several situations, spatial visualizations, particularly charts, seem to be more accurate, quicker to work with and inspiring more confidence on the decision made.

Given these results, tool developers may want to explore ways to enrich their diagrammatic notations in a way that symbolic representations are replaced by or augmented with representations amendable to low-level visual inferences whenever needed. Our future research agenda includes exploration of interactive experiences in which diagrammatic goal models are complemented on-demand by spatial visualizations such as charts. Such investigation focuses both on applicable forms of such visualizations (e.g. are charts scalable for larger models?) and on the underlying automated reasoning mechanism that supports the interaction.

REFERENCES

[1] Z. S. H. Abad, M. Noaeen, and G. Ruhe. Requirements engineering visualization: A systematic literature review. In *Proc. of the 24th International Requirements Engineering Conference (RE'16)*, pages 6–15, Sept 2016.

[2] D. Amyot, S. Ghanavati, J. Horkoff, G. Mussbacher, L. Peyton, and E. S. K. Yu. Evaluating goal models within the goal-oriented requirement language. *International Journal of Intelligent Systems*, 25(8):841–877, 2010.

[3] D. Amyot and G. Mussbacher. User requirements notation: The first ten years, the next ten years (invited paper). *Journal of Software (JSW)*, 6(5):747–768, 2011.

[4] T. Asahi, D. Turo, and B. Shneiderman. Using treemaps to visualize the analytic hierarchy process. *Information Systems Research*, 6(4):357–375, 1995.

[5] P. Caire, N. Genon, P. Heymans, and D. L. Moody. Visual notation design 2.0: Towards user comprehensible requirements engineering notations. In *Proc. of the 21st IEEE International Requirements Engineering Conference (RE'13)*, pages 115–124, 2013.

[6] J. A. Cruz-Lemus, M. Genero, M. E. Manso, S. Morasca, and M. Piattini. Assessing the understandability of UML statechart diagrams with composite states—a family of empirical studies. *Empirical Software Engineering*, 14(6):685–719, 2009.

[7] J. Gettinger, E. Kiesling, C. Stummer, and R. Vetschera. A comparison of representations for discrete multi-criteria decision problems. *Decision Support Systems*, 54(2):976–985, 2013.

[8] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani. Formal Reasoning Techniques for Goal Models. *Journal on Data Semantics*, LNCS 2800, pages 1–20, 2003.

[9] B. Gonzales-Baixauli, J. C. S. P. Leite, and J. Mylopoulos. Visual variability analysis for goal models. In *Proc. of the 12th IEEE International Requirements Engineering Conference RE'04*, pages 198–207, Sept 2004.

[10] O. C. Gotel, S. J. Morris, and F. T. Marchese. On requirements visualization. In *Proc. of the 1st International Workshop on Requirements Engineering Visualization*, New Delhi, India, 2007.

[11] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi. Comparing the comprehensibility of requirements models expressed in use case and Tropos: Results from a family of experiments. *Information and Software Technology*, 55(10):1823 – 1843, 2013.

[12] J. Horkoff and E. Yu. Finding solutions in goal models: an interactive backward reasoning approach. In *Proc. of the 29th International Conference on Conceptual modeling (ER'10)*, pages 59–75, Vancouver, Canada, 2010.

[13] J. Horkoff and E. S. K. Yu. Interactive goal model analysis for early requirements engineering. *Requirements Engineering*, 21(1):29–61, 2016.

[14] B. Johnson and B. Shneiderman. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proc. of the 2nd Conference on Visualization (VIS'91)*, pages 284–291, Los Alamitos, CA, USA, 1991.

[15] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, NJ, USA, 6 edition, 2007.

[16] A. S. Kelton, R. R. Pennington, and B. M. Tuttle. The effects of information presentation format on judgment and decision making: A review of the information systems research. *Journal of Information Systems*, 24(2):79–105, 2010.

[17] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[18] S. Liaskos, R. Jalman, and J. Aranda. On eliciting preference and contribution measures in goal models. In *Proc. of the 20th International Requirements Engineering Conference (RE'12)*, pages 221–230, Chicago, IL, 2012.

[19] S. Liaskos, S. M. Khan, M. Soutchanski, and J. Mylopoulos. Modeling and reasoning with decision-theoretic goals. In *Proc. of the 32th International Conference on Conceptual Modeling, (ER'13)*, pages 19–32, Hong-Kong, China, 2013.

[20] S. Liaskos, S. McIlraith, S. Sohrabi, and J. Mylopoulos. Representing and reasoning about preferences in requirements engineering. *Requirements Engineering Journal (REJ)*, 16:227–249, 2011.

[21] N. Maiden, P. Pavan, A. Gizikis, O. Clause, H. Kim, and X. Zhu. Making decisions with requirements: Integrating i* goal modelling and the AHP. In *Proc. of the 8th International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ'02)*, Essen, Germany, 2002.

[22] S. E. Maxwell and H. D. Delaney. *Designing Experiments and Analyzing Data*. Taylor and Francis Group, LLC, New York, MA, USA, 2 edition, 2004.

[23] K. Miettinen. Survey of methods to visualize alternatives in multiple criteria decision making problems. *OR Spectrum*, 36(1):3–37, 2014.

[24] D. L. Moody, P. Heymans, and R. Matulevičius. Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation. *Requirements Engineering*, 15(2):141–175, 2010.

[25] J. Mylopoulos, L. Chung, S. Liao, H. Wang, and E. Yu. Exploring alternatives during requirements analysis. *IEEE Software*, 18(1):92–96, 2001.

[26] H. C. Purchase, R. Welland, M. McGill, and L. Colpoys. Comprehension of diagram syntax: an empirical study of entity relationship notations. *International Journal of Human-Computer Studies*, 61(2):187 – 203, 2004.

[27] R. L. Rosnow and R. Rosenthal. *Beginning Behavioral Research: A Conceptual Primer.* Pearson Prentice Hall, NJ, USA, 6 edition, 2008.

[28] M. Tennekes. Treemap visualization. R package 'treemap', CRAN repository., 2015.

[29] N. S. Umanath and I. Vessey. Multiattribute data presentation and human judgment: A cognitive fit perspective. *Decision Sciences*, 25(5-6):795–824, 1994.

[30] I. Vessey and D. Galletta. Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, 2(1):63–84, 1991.

[31] E. S. K. Yu. Towards modelling and reasoning support for early-phase requirements engineering. In *Proc. of the 3rd IEEE International Symposium on Requirements Engineering (RE'97)*, pages 226–235, Annapolis, MD, 1997.

[32] B. Zhu and H. Chen. Information visualization for decision support. In *Handbook on Decision Support Systems 2: Variations*, pages 699–722. Springer, 2008.