A Splash of Color: A Dual Dive into the Effects of EVO on Decision-Making with Goal Models

Yesugen Baatartogtokh · Irene Foster · Alicia M. Grubb

Received: 7 December 2023 / Accepted: 4 June 2024

Abstract Recent approaches have investigated assisting users in making early trade-off decisions when the future evolution of project elements is uncertain. These approaches have demonstrated promise in their analytical capabilities; yet, stakeholders have expressed concerns about the readability of the models and resulting analysis, which builds upon Tropos. Tropos is based on formal semantics enabling automated analysis; however, this creates a problem of interpreting evidence pairs. The aim of our broader research project is to improve the process of model comprehension and decision-making by improving how analysts interpret and make decisions. We extend and evaluate a prior approach, called EVO, which uses color to visualize evidence pairs. In this article, we explore the effectiveness of EVO with and without the impacts of tooling through a two-phased empirical study. All subjects in both phases were untrained modelers, given training at study time. First, we conduct an experiment to measure any effect of using colors to represent evidence pairs. Second, we explore how subjects engage in decision-making activities (with or without color) through a user study. We find that the EVO color visualization significantly improves the speed of model comprehension and is perceived as helpful by study subjects.

1 Introduction

Goal-oriented requirements engineering (GORE) aims to assist individuals to make decisions about their projects. To do so, analysts create models consisting of actors and intentions (e.g., goals, tasks), as well as connections between them. These models can then be evaluated for a given scenario by placing a label on each intention of interest to the user. In the domain of qualitative evaluations of goal models, there are multiple methods for evaluating intentions. For example, iStar and GRL use visual labels (e.g., checkmarks and Xs), while Tropos uses evidence pairs (e.g., (F, P)). In comparing these approaches, the visual labels in iStar are more understandable to end-users but lack formal semantics, while the evidence pairs in Tropos allow for automation but are hard for users to understand.

This tension between model comprehension and automated analysis is further exacerbated by evaluating models over time [3,20] and with families of models [1], where users evaluate collections of models. Given the potential for automating analysis of goal models [30] and connecting them with downstream activities [25], the broader aim of this research program is to improve the *cognitive effectiveness* [33] of Tropos evidence pairs, making them more accessible to end-users.

The comprehensibility of Tropos models has already been investigated in the literature. Hadar et al. compared Tropos and Use Case models and found that Tropos models seem to be more comprehensible with respect to some requirements analysis tasks, although Tropos models were found to be more time consuming [23]. In a replication of Hadar et al.'s work, Siqueira found no difference in model comprehensibility and effort between Tropos and Use Case models, when those models have equivalent complexity [42]. While an important foundation, this work is tangential to our investigation because we are interested in improving the comprehensibility of Tropos relative to itself, rather than comparing it to other approaches.

Yesugen Baatartogtokh · Irene Foster · Alicia M. Grubb
Department of Computer Science
Smith College, Northampton, MA, USA
E-mail: amgrubb@smith.edu

In prior work, Grubb and Chechik developed automated analysis techniques for Tropos models with evolutionary information [22]. Building on this framework and the BloomingLeaf tool, Varnum et al. proposed using colors to assist users in interpreting evidence pairs in Tropos, which they called EVO (Evaluation Visualization Overlay) [45]. Varnum et al. completed a preliminary evaluation with an example but did not validate this approach with users [45]. Ben Ayed et al. extended EVO by adding alternative color palettes based on regional and individual interpretations of color and enabling users to create their own custom palettes [6]. Prior work suggests that color can help individuals interpret certain graph types faster [29], but should be used as a secondary encoding [33].

Contributions. The high-level objective of this work is to investigate to what extent, if any, using EVO affects how individuals understand and make decisions about goal models with timing information, using Tropos evidence pairs. Additionally, we continue our ongoing investigation of the utility of goal models [9,18] and continue to observe how individuals interact with evolving goal models and our tool, BloomingLeaf [21].

In this article, we report the results of a two-phased empirical study conducted in a laboratory setting with *undergraduate students* at Smith College. In the first phase, we completed an experiment with 32 subjects to directly compare subjects' ability to answer goal modeling questions with and without the use of EVO (called the "Experiment" study in this article). In the second phase, we conducted an *experimental simulation* [43] and *user experience evaluation* [47] with 11 subjects (called the "User" study), in which we simulated the process of stakeholders reviewing and performing analysis on a goal model with the assistance of a trained modeler. We observed subjects directly using BloomingLeaf, on a model created with and for them.

In the Experiment study, we aimed to answer three research questions:

RQ1 To what extent are subjects able to learn EVO, and then use EVO to answer goal modeling questions?

RQ2 How does EVO compare with the control in terms of time and subjects' perceptions?

RQ3 How do subjects rate the study experience and instrument?

In the User study, we ask three additional research questions:

RQ4 To what extent did subjects engage in goal modeling and decision making activities using BloomingLeaf?

RQ5 How do subjects perceive and use EVO during an in-person goal modeling session?

RQ6 How do subjects assess the in-person session and BloomingLeaf?

In the Experiment, we found that with minimal prior training in goal modeling, subjects were able to learn and use the EVO extension to make decisions. We found no evidence that EVO altered the quality of understanding or decision making, either positively or negatively. However, we found that EVO significantly decreased the time required to make decisions. Finally, the subjects responded positively to EVO and the study protocol.

In the User study, we found that subjects were able to make real-life decisions using goal models in BloomingLeaf with minimal training. Subjects understood and engaged with the base model, altering it to reflect their needs, and analyzed the simulation to make decisions or comment on the believability of the results. Eight out of eleven subjects used EVO in their exploration, expressing that they found it helpful or enjoyable. In the study debrief, all subjects had a positive reaction to EVO, even ones who had not used it.

Organization. The remainder of this article is organized as follows. Sect. 2 reviews goal modeling, the BloomingLeaf tool, the EVO approach, and its extensions. Sect. 3 describes our methodology for the Experiment and User studies. We report on the results of our studies in Sect. 4 and Sect. 5, and discuss lessons learned and validity in Sect. 6. Finally, we review related work in Sect. 7 and conclude in Sect. 8.

2 Background

In this section, we review the goal modeling notation, goal modeling tool, visualization overlay and visualization extensions used in this study.

2.1 Goal Model Notation

We use the Employee model shown in Fig. 1 to illustrate our notation. A goal model consists of actors, intentions, and links. Intentions describe the intentionality of each actor and consist of four types: goals, soft goals, tasks, and resources. For example, Fig. 1 contains one actor, named Employee, and nine intentions that describe the Employee's motivations.

Intentions can be decomposed or contribute to the fulfillment of one another via links, forming one or more graphs of nodes in the model. Decomposition links (i.e., and, or) decompose an intention into subsequent or child nodes. An intention with an and-decomposition requires all of its children to be fulfilled, while an or-decomposition requires only one to be fulfilled. In Fig. 1,

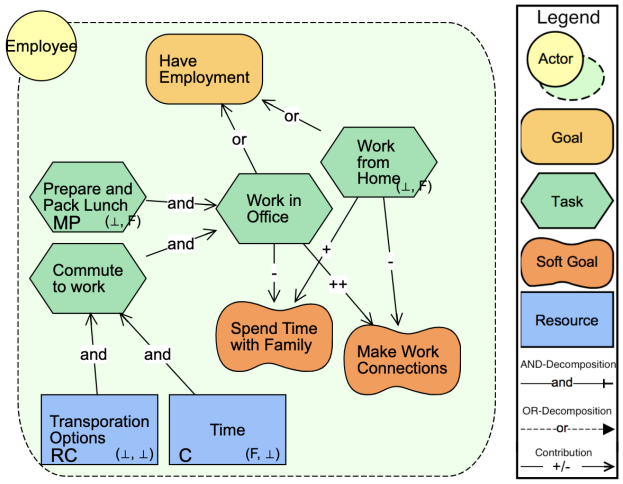


Fig. 1: Employment Model & Goal Modeling Legend

Fig. 2: Evidence Pairs with EVO Color Assignments

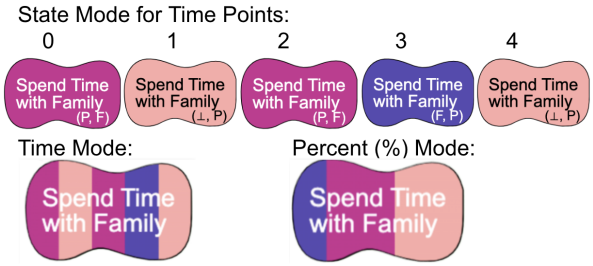


Fig. 3: EVO Modes (State, Time, and Percent) shown on Spend Time with Family Soft Goal

the Employee's only goal is to Have Employment, which is Or-decomposed into two alternate tasks Work from Home and Work in Office. Contribution links (e.g., +, -, ++S, –S) indicate that an intention has influence on another intention. For example, Work in Office (see Fig. 1) propagates all evidence to Make Work Connections via a ++ link, while the - link between Work in Office and Spend Time with Family negates and propagates partial evidence of fulfillment.

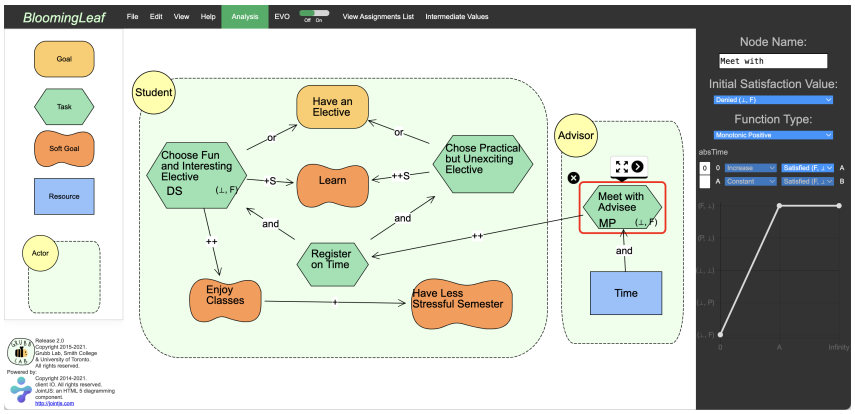
The fulfillment of an intention is evaluated qualitatively using an *evidence pair* (s, d), which separates evidence *for* and *against* the fulfillment of the intention. Both s and d consist of one of three values: F represents full evidence, P represents partial evidence, and $\perp$ represents no evidence, where $\perp \leq P \leq F$. Thus, goals can have one of five initial values: [*Fully*] *Satisfied* $(F, \perp)$, *Partially Satisfied* $(P, \perp)$, *Partially Denied* $(\perp, P)$, [*Fully*] *Denied* $(\perp, F)$, and *None* $(\perp, \perp)$; as well as four conflicting values that may result from propagation: (F, F), (F, P), (P, F), and (P, P). For clarity, we list these evidence pairs in Fig. 2. In Fig. 1, the task Prepare and Pack Lunch is assigned the value *Denied* $(\perp, F)$ because the actor Employee has not yet completed the task.

2.2 Simulating Models over Time

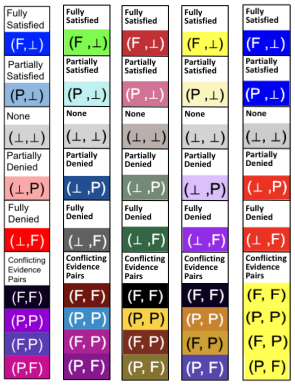
We use the Evolving Intentions framework [22] to simulate how a model's fulfillment changes over time. The framework allows users to specify one or more stepwise functions (called *User-Defined* (*UD*) functions) describing how the evidence pair assignment for an intention changes over time. Over any time interval, the valuation of an intention can *Increase* (*I*), *Decrease* (*D*), remain *Constant* (*C*), or be random or *Stochastic* (*R*).

In Fig. 1, the resource Time remains CONSTANT with the valuation of *Satisfied* $(F, \perp)$ over time. The *MP* label on Prepare and Pack Lunch indicates a *Monotonic Positive* function, meaning that the valuation will become more fulfilled until it is fully satisfied and then it will remain constant with that value. Three other functions that appear in this paper are: (*Denied-Satisfied* (*DS*)) the satisfaction evaluation remains *Denied* $(\perp, F)$ until t and then remains *Satisfied* $(F, \perp)$; (*Stochastic-Constant* (*RC*)) changes in satisfaction evaluation are stochastic or random until t and then remains constant with a given evidence pair; and (*Constant-Stochastic* (*CR*)) the satisfaction evaluation remains constant at a given evidence pair until t and then changes in evaluation are stochastic.

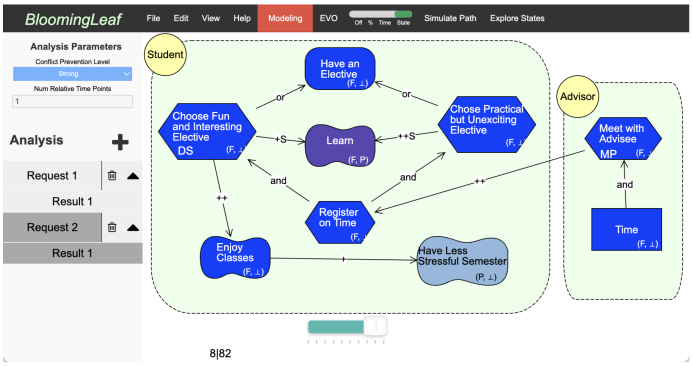
After a path has been simulated, all of the intentions in the model are assigned an evidence pair label for each time point. The time points in the simulated path are ordered and have absolute times (i.e., ticks) associated with them, which allows the path to have variable intervals of real-world time between sampled time points. The mapping of simulation ticks to real-world time is at the discretion of the user. For each time point in the path, the intentions in the model are assigned either directly by their evolving function specifications or indirectly via propagation. When an intention is not constrained via specification or propagation, an evidence pair label is assigned randomly. A contribution of the framework is to allow users to make trade-off decisions about the future states of the model by stepping through each time point in a simulation and reviewing the evidence pair assignments of each intention.



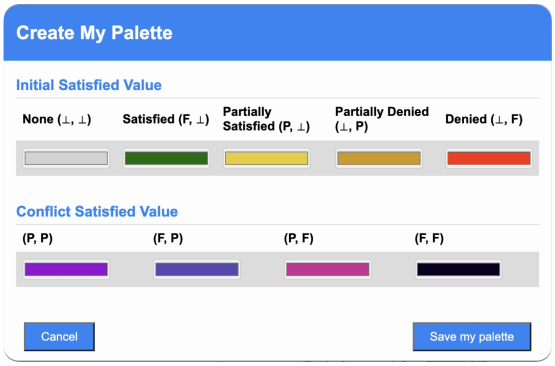
(a) Modeling Mode in BloomingLeaf with the Course Model on the Center Canvas



(b) EVO Color Palettes



(c) Analysis Mode in BloomingLeaf Simulating the Course Model



(d) Sample EVO Custom Palette

Fig. 4: Screenshots of BloomingLeaf. On the left-side, the Course scenario is modeled and analyzed (see Sect. 3.2 for a description of the Course model). The model is created and the initial evidence pairs and evolving functions are assigned in Fig. 4(a), where Meet with Advisee is selected enabling the intention inspector on the right panel. Since *Monotonic Positive* is selected for Meet with Advisee, the right panel also shows a graph describing how the fulfillment will *increase* and then become *constant*. With these initial assignments, the user simulates a path of the Course model evolving in Fig. 4(c), which shows the model at time point 82 (the eighth step in the path). Since Meet with Advisee has transitioned to being constant and *Satisfied* $(F, \bot)$, the EVO State model colors it blue. The right-side illustrates the palette extensions by Ben Ayed et al. [6]. Fig. 4(b) shows the five palettes (left to right): Default, Green-Black, Red-Green, Yellow-Purple, and Color-Blind. Fig. 4(d) displays the custom palette editor, where users can assign a color to each evidence pair.

2.3 EVO: Evaluation Visualization Overlay

As briefly mentioned in Sect. 1, Varnum et al. introduced the Evaluation Visualization Overlay (EVO) [45]. EVO was designed to assist users in understanding evidence pairs. Each evidence pair (s, d) label is assigned a color (see legend in Fig. 2), where blue denotes evidence for (i.e., the s value), red denotes evidence against (i.e., the d value), and purple denotes conflicting evidence. The more saturated (or darker) the color shade, the stronger the evidence (i.e., F is darker than P). Observe that (F, F) is a very dark shade of purple, whereas (P, P) is a lighter shade of purple. For (P, F) there is both blue and red present, making it purple, but because there is more evidence for denial, it is more red-purple, with the inverse being true for (F, P). During modeling activities, when EVO is enabled the color of each intention corresponds to any initial assignment, while unassigned intentions retain their original color (see legend in Fig. 1). This provides an overall visualization of the model's initial state. For example, Fig. 5 gives the initial state of the Summer model (see Sect. 3.2 for details). In Fig. 5, Have Summer Activity is colored dark red because it has been assigned the $(\bot, F)$ label.

The main contribution of EVO is to assist users in evaluating evidence pair assignments across a simulation path. Within the Evolving Intentions framework introduced above, it is difficult for a user to remember all of the different valuations of each intention at each

time point, much less synthesize them all together to act upon the given information. EVO provides three modes to visualize simulations: *State*, *Time*, and *Percent*. To introduce these modes, we consider only the Spend Time with Family intention from Fig. 1. *State* mode shows the current time point of the model, with the background of each intention colored based on their assigned evidence pair. Fig. 3 shows the color and evidence pair assignments for Spend Time with Family at time points 0–4. *Time* mode shows the valuations over the entire path in one view. For example, in Fig. 3, each of the stripes on Spend Time with Family represents the colors of each state shown above. Finally, *Percent* mode colors by overall evaluation percentages, making the background of each intention colored with the percentage of states in the simulation where the intention has each evidence pair assignment. The width of each colored stripe corresponds to the percentage of time points that it holds a specific evidence pair, ordered based on level of fulfillment.

2.4 BloomingLeaf Tool

BloomingLeaf is a browser-based tool for the formal automated analysis of goal models [19]. Subjects in the User study used BloomingLeaf to update and explore simulations of their chosen scenario. The Modeling mode, shown in Fig. 4(a), allows the user to build an initial goal model and assign relevant evaluation labels. If needed, a user can also specify the timeframe over which the model or certain events occur, and define the custom evolution of an intention. The EVO slider in the top toolbar of Fig. 4(a) allows a user to turn on the color overlay.

The Analysis mode allows the user to simulate random paths based on the initial model. Prior to simulating, the user can set a path's conflict prevention level or the number of relative (meaning additional) time points they wish to have. Once ready, clicking "Simulate Path" returns a random path. A user can simulate as many paths as they wish. Fig. 4(c) shows a model simulated over four time points. The user may switch back and forth between Modeling and Analysis mode. The user can use the slider to examine the evolution of each intention at each time point. The EVO slider enables the user to explore the path with EVO turned on in State, Percent, or Time modes, as explained in Sect. 2.3 While the default is the Blue-Red palette, the EVO dropdown menu contains several alternatives which will be explained in Sect. 2.5. Other features in BloomingLeaf are beyond the scope of this paper.

2.5 International Colors

Ben Ayed et al. extended the EVO framework (see Sect. 2.3) with four additional palettes and the option to create a custom palette, in order to make the EVO framework accessible to a wider audience [6]. This work considers the impact of cultural background and color-deficiencies on users, as the default Blue-Red may not be intuitive to all users. A lack of intuitive understanding may require an extra layer of cognitive processing from users, taking from the intended benefits of EVO.

Three of the additional palettes are for users with different cultural meanings for color: Green-Black for an Arab audience; Red-Green for an East Asian audience; and Yellow-Purple for a Brazilian audience. The Color-Blind palette allows accessibility among users with color deficiencies. Users can create a custom palette of their preferences from the "Create My Palette" window, as shown in Fig. 4(d).

3 Methodology

In this section, we describe our methodology for conducting this study, which was approved by the Smith College Institutional Review Board (Protocol #20-026). Our supplemental materials are available online[1].

3.1 Study Design

3.1.1 Experiment Study Design

We begin by describing the design of the Experiment study. Our primary objective in designing these experiments was to measure the effects of EVO. The original EVO proposal was implemented as an extension to BloomingLeaf [21]. We did not intend to evaluate BloomingLeaf; instead, we wanted to test EVO in isolation without the confounding variables of tooling, making our study tool agnostic. Additionally, we wanted to collect timing information in an accurate way. Thus, we designed the study instrument to be completed via Smith College's browser-based Qualtrics$_{®}$ XM platform. We used the BloomingLeaf git repository [21] only for the purpose of creating our study materials and models.

The core of the Experiment is to measure subjects' performance in answering goal modeling questions with and without EVO (see RQ1 and RQ2). Thus, our dependent variables were the subjects' *score* and *time* in answering goal modeling questions (see Tbl. 1 for a full description of the variables).

[1] See *https://doi.org/10.35482/csc.001.2024* for supplement.

Table 1: Study Variables

Independent Variables	
Treatment (EVO)	Whether subjects used or did not use EVO while answering questions.
Experimental Object (Model)	Which model the subjects used while answering questions.
Period (Order)	Whether EVO training took place before or after the review of an experimental object.
Sequence (Treatment Group)	The interaction of treatment, period/order, and experimental object (i.e., Bike or Summer model).
Dependent Variables	
Score	Number of correct answers on each question set. A satisfactory score is at least 70%. (i) Goal Modeling (and Simulation) Training Questions (ii) EVO Training Questions (iii) Bike Model Questions (iv) Summer Model Questions
Time	The time it takes each subject to complete the questions, which are scored. One time for each of Score (i)–(iv) above.
Review Time	The time it takes each subject to review training materials. (i) Goal Modeling (and Simulation) Training (ii) EVO Training

Table 2: Possible Study Designs [46]

(a) Between-Subjects (or Independent Measures) Design

Seq.	Period I
I	Treatment A, Model 1
II	Treatment B, Model 1

(b) Within-Subjects (or Repeated Measures) Design

Seq.	Period I	Period II
I	Treatment A, Model 1	Treatment B, Model 2

(c) Two-Treatment Factorial Crossover Design Where the Experimental Object is a Two-Level Blocking Variable

Seq.	Period I	Period II
I	Treatment A, Model 1	Treatment B, Model 2
II	Treatment B, Model 2	Treatment A, Model 1
III	Treatment A, Model 2	Treatment B, Model 1
IV	Treatment B, Model 1	Treatment A, Model 2

We considered various approaches in designing our experiment and the risks of each approach. Previous research has demonstrated that *task equivalency* is a risk factor in analyzing model comprehensibility [42]. Thus, we first designed a set of questions that could be fairly answered with and without the use of EVO. This limited our ability to test certain aspects of EVO. For example, we did not ask questions that specifically required the use of EVO Percent or Time modes.

In a simple independent measures (between-subjects) design (see Tbl. 2(a)), we assign half the subjects to use EVO and the other half as a control to answer the same questions over a given model. This design does not account for differences in subject variability. Since we do not compare EVO with another technique (as in [23]), we did not have anything for the control group to learn in place of EVO, and we anticipated that the control subjects would complete the study notably faster, but receive the same compensation as the EVO subjects. Thus, we excluded this design due to the unequal treatment of subjects, which violates institutional norms.

In a simple repeated measures (within-subjects) design (see Tbl. 2(b)), all subjects would first answer questions without EVO before learning and answering questions with EVO. This design mitigates the two limitations of the simple between-subjects but introduces both carryover and learning-by-practice effects, which cannot be separated in this experiment. To ensure *task equivalency* (see above), we had to keep questions similar enough between study periods in a repeated measures design. Subjects may become better at answering goal modeling questions with practice and there may be carryover between periods. To mitigate these risks we considered separating the study periods into multiple temporally distributed sessions, but ruled out this proposal due to the risk of subjects ghosting or dropping out of the study (i.e., mortality threat [50]) and subjects behaving differently at different times (i.e., history threat). Finally, a repeated measures design requires using a second model for subjects to answer questions about. If all subjects use the same model for EVO and a second model as a control, then the model and question difficulty are confounded with EVO usage.

To control for the various risks mentioned above, we chose a two-treatment factorial crossover design where the experimental object (i.e., model) is a two-level blocking variable (see Tbl. 2(c)). While the risks of carryover, learning, and model variability are still present in this design, using a crossover design allows us to control for and measure differences between periods and our experimental object (i.e., model). Further, we considered a balanced crossover design with an addi-

tional treatment period (not shown), but this would have required a third model (i.e., adding to the task equivalency threat) and would substantially increase the length of the study. We chose against this design to follow institutional norms and limit our total session time to a maximum of one hour to mitigate subject fatigue (i.e., tiredness/boredom).

Given our concerns about carryover at design time, we wanted to ensure that we had the ability to analyze the data appropriately in the event of carryover. Therefore, upon analysis, we first check for the presence of carryover. If it is not detected then we continue our analysis within subjects, per our repeated measures design. In the event that carryover is detected, then we convert our analysis to between-subjects and only consider the initial measurement period (see Period I, in Tbl. 2(c)). In this case, we verify that the random assignment of subjects to sequences is sufficiently uniform by checking for the existence of variations between sequences using the scores on the study training.

After first introducing the various study materials in Sect. 3.2, we return to our protocol for the Experiment in Sect. 3.3 and describe it in more concrete terms. See Sect. 6.2 and Sect. 6.3 for a discussion of statistical power and threats to validity of our design, respectively.

3.1.2 User Study Design

As already introduced in Sect. 1, our goal was to experimentally simulate the experience of stakeholders modeling and reasoning with goal models. We designed the User study to be complementary, yet comparable, to the Experiment study. As mentioned above, the Experiment was conducted in isolation of any goal modeling tooling; thus, the User study was designed to take tool usage into account. We used the same subject population and training materials to enable comparison between phases; yet, in this study, we explored individual variations between subjects and collected richer qualitative data about subjects' perceptions of EVO and BloomingLeaf (including options for different EVO modes and palettes).

A critique of the Experiment was that subjects may not have been invested in the outcome of the modeling task and that the modeling and analysis questions (see Tbl. 4) may not be sufficiently realistic of goal modeling activities. We designed our User study with the objective that subjects work on a problem and question that they are personally invested in. Mitigating these issues in a one-hour study of untrained modelers required us to perform some of the initial model creation offline. We added a *pre-study questionnaire*[1] to the study interest form, where subjects described in detail a decision they were currently struggling with (e.g., choosing between opportunities after college). Subjects described the trade-offs they were considering, as well as any of their dependencies. This questionnaire allowed us to create an initial model for the subjects' chosen scenario (see Sect. 5.2 for a discussion of our generated models). Additionally, we asked subjects to self-describe their cultural background and color associations (e.g., color(s) associated with positive outcomes), in order to gauge the interpretability of the various EVO color palettes.

Using BloomingLeaf directly allows us to see how subjects use goal modeling in a real-life decision-making scenario that is applicable to them. Being able to interact with the model and having the freedom to use or not use EVO gave us direct insight into their preferences and reasoning behind their decision-making, which is crucial to our understanding of how usable and useful EVO is.

Given the qualitative nature of the User study, our design included considerations of data analysis. A single researcher lead each in-person session; while a second researcher transcribed the session recordings. Two researchers independently reviewed the qualitative responses from the surveys, as well as session recordings and transcripts, taking notes on observations before meeting to compare findings and review any discrepancies. We did not formally code this data [51] as it was constrained and within our expected observations, instead providing the transcripts and anonymized qualitative data for other researchers to verify.

3.2 Materials: Models and Videos

In this study, we used four models: the Employment model (see Fig. 1), the Summer model (see Fig. 5), the Bike model (Fig. 6), and the Course model (see Fig. 4(a)). We list these models and their associated metrics in Tbl. 3. The Course model (see Fig. 4(a)) describes the process of a student (and their advisor) trying to decide whether the student should take a fun and interesting or practical and unexciting elective in the next semester. In Sect. 2.1, we describe the Employment model (see Fig. 1) to introduce goal model syntax. The model describes an employee, who is debating between working from home or working in an office, with the top-level goal of Have Employment.

In the Summer model (see Fig. 5), the actor Joy wants to have a summer activity, with choices between tasks Join Book Club, Join Community Center, and Join Soccer Team. These tasks are and-decomposed into sets of tasks that must be satisfied. In the Bike model shown in Fig. 6, the City actor wants to construct bike lanes,

Table 3: Study Models

Model	Fig.	Actor	Intention	Link	Evolving Function
Course	4(a)	2	9	10	2
Employment	1	1	9	10	3
Summer	5	1	14	17	8
Bike	6	1	16	20	7

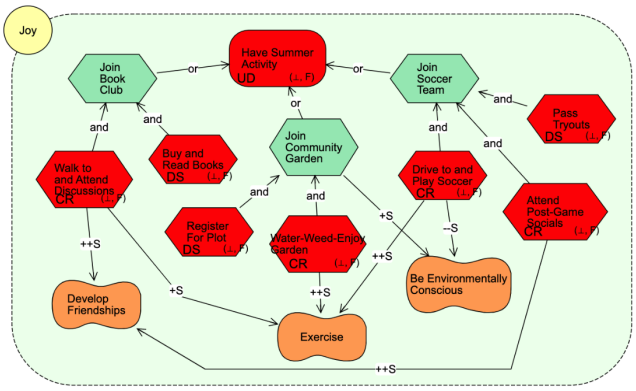


Fig. 5: Summer Model with EVO on in Modeling Mode

with the top-level goal Have Bike Lanes, for which they must have satisfied both sub-goals Have Design Plans and Have Build Plans. These two goals are OR-decomposed into tasks they must choose from.

In the Experiment study, subjects were tested on their ability to answer questions about the Bike and Summer models (see Tbl. 4 for list of questions). We created both an EVO and control version of all models. These models as well as their simulations are available online[1]. While the Bike model has more intentions and links, the evolving functions are simpler than the Summer model.

Our study used four training videos (transcripts available online[1]): (i) *Goal Models in Tropos* (VidGM) reviews goal modeling and explains Tropos evidence pairs and links. (ii) *Introduction to Simulation Over Time* (VidSim) introduces function types and evolving intentions, describing what it means to simulate a model over time. (iii) *EVO* (VidEVO) introduces the *EVO* color scheme for evidence pairs and goes over its three possible modes: *State*, *Time*, and *Percent*. (iv) *Introduction to BloomingLeaf* (VidBL) introduces the basic modeling and simulation features of the BloomingLeaf tool, including usage of the EVO feature. VidBL was only shown to subjects in the User study.

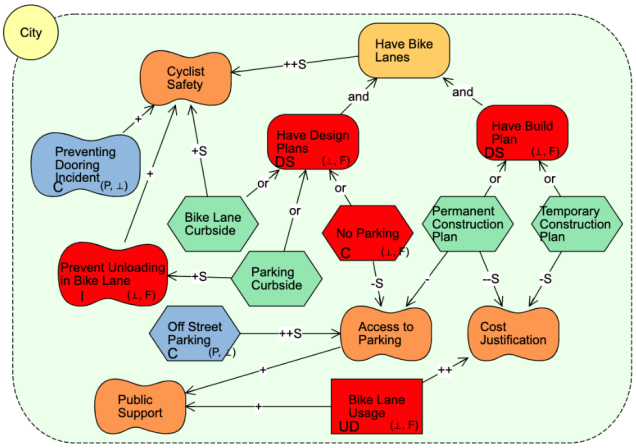


Fig. 6: Bike Model with EVO on in Modeling Mode

3.3 Study Procedures and Protocol

Tbl. 5 lists the steps in our protocol for both the Experiment and User study. We divided our protocol into five parts (i.e., periods), which are listed in the left-most column of Tbl. 5. The four middle columns of Tbl. 5 list the protocol for each treatment group (i.e., sequence) of the Experiment. The right-most column lists the procedure for the User study. Parts 0, 1, and 5 are common across all subjects and protocols. In Part 0, we obtained *informed consent* from all subjects and had them rate their previous experience with goal modeling. In this step, we also had them complete a short (seven question) color deficiency test to ensure subjects met the inclusion criteria (see Sect. 3.4).

In Part 1, subjects completed two training modules, one introducing goal modeling more generally using VidGM, and the other introducing the minimal required subset of the Evolving Intentions framework (using VidSim). We used the Course and Employment models in Part 1 and in the common 'Training: EVO' module in Parts 2 and 3 (see shaded areas in Tbl. 5). Specifically, the Course model was used as part of our training materials, including videos, to introduce new concepts. After each module, subjects were asked questions to test their understanding using the Employment model. The EVO training did not specifically test subjects on their knowledge of the different EVO modes. We took measurements of subjects' correctness when answering questions, labeled as *score* (see Tbl. 1), and how long it took subjects to answer these questions, labeled as *time*. With the exception of the color deficiency test in Part 0, we use the threshold of 70% to determine if a score is considered satisfactory (i.e., pass), per the faculty code at Smith College (see Section VII.G.1.d [13]).

Table 4: Summer and Bike Questions Used in the Experiment Study

Page	Num	Summer Model	Bike Model
P1	Q1	What is the initial satisfaction value of "Pass Tryouts"?	What is the initial satisfaction value of "Prevent Dooring Incident"?
P1	Q2	What is the initial satisfaction value of "Exercise"?	What is the initial satisfaction value of "Bike Lane Usage"?
P1	Q3	Is the initial state of the model more satisfied, denied, or conflicted?	Is the initial state of the model more satisfied, denied, or conflicted?
P2	Q4	For each of the elements listed below, how many times over the simulation does the element become Fully Satisfied? (a) Have Summer Activity, (b) Pass Tryouts, (c) Exercise	For each of the elements listed below, how many times over the simulation does the element become Fully Satisfied? (a) Bike Lane Curbside, (b) Temporary Construction Plan, (c) Public Support
P2	Q5	How does "Join Soccer Team" generally evolve over the simulation?	How does "Public Support" generally evolve over the simulation?
P2	Q6	For each of the following satisfaction values, at which time point in the simulation do the most number of elements have the value. Note: In the event of a tie, choose the later time point (higher number). (a) Fully Satisfied, (b) Fully Denied, (c) Any Conflicted Value	For each of the following satisfaction values, at which time point in the simulation do the most number of elements have the value. Note: In the event of a tie, choose the later time point (higher number). (a) Fully Satisfied, (b) Fully Denied, (c) Any Conflicted Value
P2	Q7	Which intentions are Partially Denied at Time Point 1?	Which intentions are Partially Satisfied at Time Point 1?
P3	Q8	Which intention would you choose to satisfy to make "Exercise" Fully Satisfied?	Which intention would you choose to satisfy to make "Prevent Unloading in Bike Lane" Fully Satisfied?
P4	Q9	On the previous page, we ask the question: 'Which intention would you choose to satisfy to make "Exercise" Fully Satisfied?' You answered [insert Q8 choice]. Please explain your answer to this question.	On the previous page, we ask the question: 'Which intention would you choose to satisfy to make "Prevent Unloading in Bike Lane" Fully Satisfied?' You answered [insert Q8 choice]. Please explain your answer to this question.
P4	Q10	How would assigning "Drive to and Play Soccer" the value Fully Satisfied influence the model?	How would assigning "Parking Curbside" and "Temporary Construction Plan" the value Fully Satisfied influence the model?
P5	Q11	Click here for a PDF to compare three different scenarios of the Summer model. Should you choose to join a book club, community garden, or soccer team?	Click here for a PDF to compare different scenarios of the Bike Lanes model. How should you construct the bike lanes?
P6	Q12	On the previous page, we asked you to compare three different scenarios of the Summer model and answer the question: 'Should you choose to join a book club, community garden, or soccer team?' You answered [insert Q11 choice]. Please explain your answer to the previous question.	On the previous page, we asked you to compare different scenarios of the Bike Lanes model and answer the question: 'How should you construct the bike lanes?' You answered [insert Q11 choice]. Please explain your answer to the previous question.

In Part 5, we debriefed and remunerated subjects, having them reflect on the study. The debriefing varied slightly between studies to reflect the differences in Parts 2-4. Other than the consent form, subjects were not able to take any study materials with them, upon completion.

3.3.1 Conducting the Experiment Study

After subjects completed the common training modules described above, they completed Parts 2–4 (see Tbl. 5), which varied based on the subjects' randomly assigned treatment group (i.e., sequence). All subjects completed the 'Training: EVO' module and answered questions about the Bike and Summer models (see Tbl. 4) after examining each model. What varied is which experimental object (i.e., Bike or Summer model) they answered questions about using EVO and whether they answered questions about a model before or after completing the EVO training (i.e., period/order). We define four treatment groups (listed in Tbl. 5):

`EBk-XSm` Subjects' answered Bike model questions with EVO, then Summer model questions without EVO.

`XSm-EBk` Subjects' answered Summer model questions without EVO, then Bike model questions with EVO.

`ESm-XBk` Subjects' answered Summer model questions with EVO, then Bike model questions without EVO.

Table 5: Study Protocol

Part	Experiment Study - Treatment Groups				User Study
	EVO: Bike		EVO: Summer		
	EBk-XSm	XSm-EBk	ESm-XBk	XBk-ESm	
0	Consent, Color Test, and Subject Background				
1	Training: Goal Modeling and Simulation				
2	Training: EVO	Summer Control	Training: EVO	Bike Control	Training: EVO
3	Bike EVO	Training: EVO	Summer EVO	Training: EVO	Interactive Modeling Session in BloomingLeaf
4	Summer Control	Bike EVO	Bike Control	Summer EVO	
5	Debrief				Debrief

`XBk-ESm` Subjects' answered Bike model questions without EVO, then Summer model questions with EVO.

Once subjects completed the experimental treatment, they completed a common debriefing component, described earlier in this section.

3.3.2 Conducting the User Study

To conduct our interview study, subjects first completed the common training modules described above (Parts 0-1 and the 'Training: EVO' in Part 2, see Tbl. 5 rightmost column). A researcher was available to answer subjects' questions during the training but sat in the room facing the opposite wall with their back towards the subject. Beginning with `S8`, we discouraged subjects from watching the training videos at double speed, as they could miss the evolving functions changing between slides. Prior to that, `S4` and `S7` had watched at least one training video on double speed. We were concerned that this would affect a subject's comprehension of the materials, although these two subjects were able to actively participate in the goal modeling session and evaluate results. After training, the researcher joined the subject at the study table and explained that the session would shift into the 'Interactive Modeling Session in BloomingLeaf' component (see combined Part 3 and 4 in Tbl. 5). We used Release 2.6[2] of BloomingLeaf to conduct the in-person modeling session. The study set-up contained a second computer mouse for the researcher to assist the subject when required. At this time, the researcher began a screen recording (with audio, if subjects provided informed consent to audio recording), and navigated to BloomingLeaf. The subjects were informed that the research team had created the base model based on their pre-study answers. Subjects were informed that they had complete autonomy over the content and appearance of the model. The researcher played the role of expert modeler and engaged the subject in modeling activities, such as:

- Understanding the initial model (including choice of links)
- Adding and removing elements and links
- Assigning and changing initial evidence pairs on the model
- Assigning and changing evolving functions for intentions
- Creating simulation paths with unassigned and absolute time points
- Exploring simulation paths and interpreting the results

Each subject was encouraged to run at least two different simulations with varying evolving functions to explore the alternatives in their model. Apart from initially pointing out the EVO feature as part of an overview of the top toolbar, the researcher did not encourage the subject to use EVO in any capacity. Depending on the comfort level of the subject, the researcher helped complete some modeling tasks in BloomingLeaf. For example, the training did not explain how to create user-defined functions, so when subjects required them, the researcher explained and created the function. In most cases, the 'Interview and Tool Evaluation' component finished when the subject felt that the model was complete and did not want to explore any additional simulations. In three cases, this component was stopped at a natural breaking point due to time considerations.

After completing the interview, the researcher stopped the recording, returned the subject to the Qualtrics questionnaire, and again turned away from the subject to give them privacy. Finally, the subjects completed the debriefing questions (see Part 5 in Tbl. 5).

[2] *https://github.com/amgrubb/BloomingLeaf/releases/tag/v2.6*

Table 6: Subjects' Reported Familiarity with Topics

Subject Group	Median Familiarity (0: None, 10: Complete)				
	English	RE	iStar	Tropos	GRL
`EBk-XSm`	10	0.5	2.5	0	0
`XSm-EBk`	10	0.5	0	0	0
`ESm-XBk`	10	1	0	0	0
`XBk-ESm`	10	0.5	0	0	0
User	10	0	1	0	0

3.4 Experimental Conditions and Subject Information

We conducted the Experiment study in early 2023 and the User study in late 2023. All subjects were required to be proficient in English, be enrolled at Smith College having previously passed 'Programming With Data Structures', and be known to not have a color vision deficiency (i.e., colorblindness), as well as apply to participate in the study. Subjects were excluded if they had a conflict of interest with our lab. Thus, we recruited subjects through a department mailing list and flyers were posted in the science buildings on campus, see supplement[1] for details.

Once subjects applied for the study, they were brought into the lab to complete the one-hour study in-person on our lab machine in a soundproof room. Since the subjects were not required to have training in goal modeling, a researcher was on hand to answer any questions after each training module.

For the Experiment study, we recruited 32 undergraduate students to participate, eight per treatment group (i.e., sequence). We conduct power analysis and discuss our sample size further in Sect. 6.2. We originally recruited twelve subjects for the User study. We excluded one subject during the in-person session (see Sect. 3.5 for a discussion); thus, we report on the results of eleven subjects throughout this article. All subjects in both studies achieved a perfect score on the color vision test. During Part 0 of our protocol (see Tbl. 5), we asked subjects to rate their familiarity with written English, requirements engineering (RE), and three GORE languages (where 0 is no familiarity and 10 is complete familiarity). Tbl. 6 reports the median familiarity score for each treatment group. Subjects rated themselves highly with respect to English. One subject in each of `XSm-EBk`, `ESm-XBk`, and `XBk-ESm` rated their familiarity with English between six and nine, while all other subjects selected ten. Similarly, one subject in the User study rated their familiarity with English as nine, with the remainder rating themselves as ten (see last row of Tbl. 6). In the Experiment study, the median scores for RE and iStar were low but non-zero; while in the User study, the median scores for RE and iStar were zero and one, respectively. Given our target population, we did not expect to find subjects with experience in Tropos or GRL but included these questions for completeness. It is likely that some of our participants completed our course in software engineering, and while RE coverage varies each semester, iStar has been covered recently. Subjects were randomly assigned to treatment groups in the Experiment study before demographic information was collected, so we were unable to use this information in group assignments. Given the data presented above, we determine that the prior knowledge of our subjects are comparable both across treatment groups in the Experiment study and between studies.

We did not collect demographic information (e.g., gender, age, race) because we did not intend to compare outcomes within these categories. For comparison with future replication studies we describe the general demographics of the population from which we recruited subjects. Smith College admits only women to undergraduate programs. Over the past three years, more than 90% of the undergraduate student body was within the age range of 18–22. For the 2023-2024 academic year, the undergraduate student population consisted of 33% students of color, 17% unrepresented minorities, and 13% international students [35].

3.5 User Study Subject Removal

As mentioned above, we excluded the data for one subject in this study after it became apparent during the in-person session that the subject did not have genuine motivation for participating. The subject played the videos at double speed and then did not review any of the training materials, instead skipping ahead to the questions. Then the subject repeatedly asked us for help answering the questions, claiming there was insufficient information. This invalidated the timing data for the study. We decided at this point, to remove the subject from the study. Per our protocol, we completed the remainder of the in-person session and remunerated the subject.

4 Experiment Results

In this section, we answer the research questions from the Experiment study (RQ1-RQ3).

4.1 Preliminaries: Assessing the Presence of Carryover/Learning Effects and Task Equivalency

We begin by considering threats of carryover (or learning by practice) and task equivalency in our study. Re-

Table 7: Tables of Mixed Effects Model for Score Data

Score Data	$\hat{\beta}$	L 2.5%	U 97.5%	p
(Intercept)	12.75	11.84	13.66	**<.001**
evo	-1.13	-2.41	0.16	.09
order	-0.63	-1.91	0.66	.34
expObj	-0.38	-1.66	0.91	.56
*evo*order*	0.88	-0.94	2.69	.34
*evo*expObj*	2.00	0.18	3.82	**.03**
*order*expObj*	0.63	-1.19	2.44	.50
*evo*order*expObj*	-0.75	-2.64	1.14	.43

Items bolded are significant at the $\alpha = .05$ level.

Table 8: Tables of Mixed Effects Model for Time Data

Time Data	$\hat{\beta}$	L 2.5%	U 97.5%	p
(Intercept)	864.38	770.79	957.97	**<.001**
evo	-293.11	-425.46	-160.76	**<.001**
order	-203.05	-335.41	-70.70	**.003**
expObj	-99.07	-231.43	33.28	.14
*evo*order*	-31.32	-218.5	155.86	.74
*evo*expObj*	116.14	-71.04	303.31	.22
*order*expObj*	143.82	-43.36	330.99	.13
*evo*order*expObj*	-78.72	-267.12	109.68	.40

Items bolded are significant at the $\alpha = .05$ level.

call from Tbl. 5 that we collected repeated measurements from subjects in Parts 2–4 to measure their ability to answer questions with and without EVO, using the Bike and Summer models (see Tbl. 1 for variables). We construct a linear mixed effects model for both the question scores and times, allowing for repeated measures within subjects.

In our mixed-effects models, the fixed effects are *evo* (i.e., treatment), *order* (i.e., period), experimental object *expObj* (i.e., Bike or Summer model), and the random effects are for each individual, which allows us to take into account their variation. This is shown in the following equation, where Y_{ij} is the dependent variable (i.e., *time* or *score*) for the i^{th} person during the j^{th} measurement (such that $i = 1, \cdots, 32$ and $j = 1, 2$), where the random effects are variance between subjects $b_i \sim Norm(0, \sigma_b^2)$ and residual error $\epsilon_{ij} \sim Norm(0, \sigma_\epsilon^2)$ (i.e., both following a normal distribution).

$$Y_{ij} = \beta_0 + \beta_1 evo_{ij} + \beta_2 order_{ij} + \beta_3 expObj_{ij} + \beta_4 evo_{ij} order_{ij} + \beta_5 evo_{ij} expObj_{ij} + \beta_6 order_{ij} expObj_{ij} + \beta_7 evo_{ij} order_{ij} expObj_{ij} + b_i + \epsilon_{ij}$$

Tbl. 7 and Tbl. 8 summarize the linear mixed effects models for score and time, respectively. For each, we report effect sizes in terms of $\hat{\beta}$. L is the lower bound on the confidence interval at 2.5% and U is the upper bound at 97.5%. The p-column shows the p-value for each model variable. As mentioned above, the validity of our mixed-effect model depends on the residuals meeting the conditions for normality. We calculated the Pearson residuals for each model and then used the Shapiro-Wilk test to evaluate normality, where normality is detected if the test does not pass the alpha level. We find the residuals for both the score ($p = .13$ for Tbl. 7) and time ($p = .12$ for Tbl. 8) models to be normal, and thus, find our models to be valid for further interpretation. Additionally, in Fig. 7, we provide a Q-Q plot of our residuals for the time data in Tbl. 8. From this, we again see that the time model residuals meet the condition of normality, as the plot shows a scatter of points with minimal deviations from the diagonal line.

Recall that as part of our study design (see Sect. 3.1.1), we wanted to control for the task equivalency threat. In Tbl. 7, the largest effect size and only significant value is the intersection of *evo*expObj*, which increases our suspicion that the experimental object used (i.e., Bike or Summer model) affects the results.

Null Hypothesis 1 *There is no observable difference between subjects' scores with and without the effects of the experimental object used.*

We conduct a Likelihood Ratio Test (LRT) [32] by comparing the mixed effects model in Tbl. 7 with the same mixed effects model without the *expObj* term. We reject Hyp. 1 as we find that there is a statistically significant effect between the scores for the two mixed effects models at the $\alpha = .05$ level ($\chi_5^2 = 13.87; p = .01$). Since we've detected this difference, for the remainder of this section we compare the results for the Bike and Summer experimental objects separately.

Next we consider the presence of a carryover (or learning by practice) effect. We find no indications of this effect in the score data (see Tbl. 7); however, *order* is significant in our time model (see Tbl. 8).

Null Hypothesis 2 *There is no observable difference between subjects' times in each study period (i.e., order).*

We conduct an additional LRT to determine whether any of the terms involving order were necessary, which uncovers the presence of a carryover effect. From this, we reject Hyp. 2 as *order* is found to be statistically significant at the $\alpha = .05$ level ($\chi_4^2 = 33.1; p < 0.001$). Thus, we find evidence of a carryover/learning effect in our experiment. As mentioned in Sect. 3.1.1, since a carryover has been detected, we conduct the remainder of our analysis in this section between-subjects and do not run hypothesis tests on the repeated measure for subjects (i.e., Part 4 in Tbl. 5). For completeness, we include summary data for the entire study.

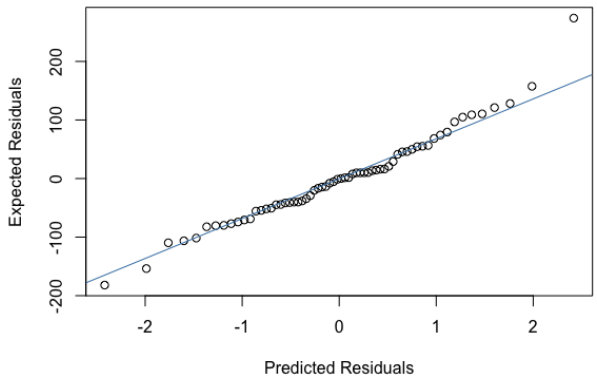


Fig. 7: Q-Q Residual Plot for Tbl. 8

We find evidence of a carryover/learning threat in our repeated measures design and find variations between task outcomes based on the experimental object used.

4.2 Preliminaries: Establishing a Baseline for Comparison Between-Subjects

Since we discovered carryover in Sect. 4.1, we convert our analysis to between-subjects. We begin by establishing that our subject groups are comparable and assessing the subjects' competence in completing goal modeling tasks, enabling further analysis of their data. All data collected during Part 1 of our protocol (see Tbl. 5) was used to establish a baseline both to compare between subjects and evaluate to what extent subjects understood the training.

First, subjects watched VidGM video and answered eight questions about goal modeling, and then they watched VidSim and answered six questions (plus one qualitative question) about simulating goal models over time, see supplement[1] for questions. All answers were scored as correct or incorrect. Fig. 8(a) reports box plots for subjects' material review time, question answering time, and question scores (from left to right). Each box plot is sorted by treatment group and times are reported in seconds. For completeness, we include a plot for the results of the User study in Fig. 8(a), which we discuss in Sect. 5.1. All subjects achieved a satisfactory score (i.e., 70%) on the training questions and thus, their data is included for further analysis. Most subjects spent 13.3–15.8 minutes reviewing the training materials (i.e., rounded first to third quantile[3]), which included a 12.6-minute video), most subjects took 7.7–11.1 minutes to answer the training questions, achieving scores between 12–14 (out of 14). From the box plots, we cannot observe any meaningful difference between treatment groups. For completeness, we used the *Kruskal-Wallis Rank Sum* (KWRS) test [37] to test for any variability between treatment groups.

Null Hypothesis 3 *There is no observable differences between treatment groups with respect to training scores.*

Null Hypothesis 4 *There is no observable differences between treatment groups with respect to the time it took for subjects to answer training questions.*

We failed to reject both null hypotheses at the $\alpha = .05$ level ($\chi^2_3 = 0.60; p = .90$ for Hyp. 3 and $\chi^2_3 = 0.11; p = .99$ for Hyp. 4), meaning that we could not detect a difference between the treatment groups.

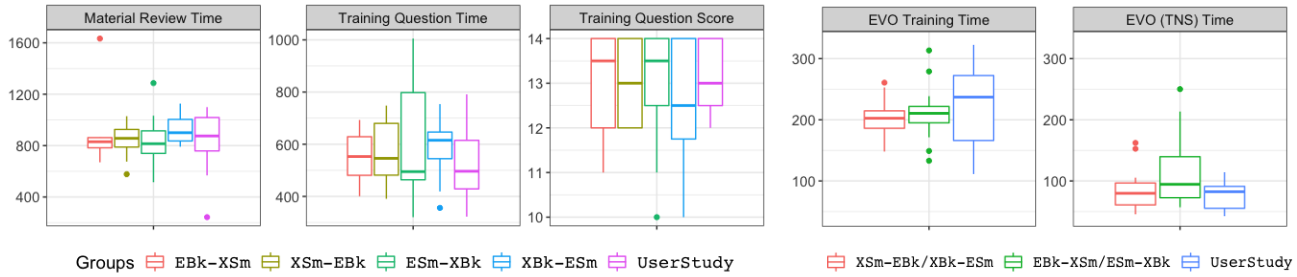
Additionally, subjects were asked to document any questions they had after reviewing the training videos (and associated documents). For the goal modeling training (TNG), eighteen subjects left a substantive question. These questions were most commonly about the evidence pairs, differences in contribution link types, and specific choices made by the modeler of the example. There were two questions about the differences between the training materials and iStar. For the simulation training, fourteen subjects asked a question. The vast majority of them were about choice and usage of evolving functions. Specifically, to explain the behavior of an intention without an assigned evolving function. Anecdotally, based on our experience teaching goal modeling, these questions are consistent with those asked in the classroom. Since subjects were not trained modelers, researchers answered subjects' questions before proceeding to the next part of the study.

We conclude that all subjects performed satisfactorily on the goal modeling and simulation training, and that subject groups were indistinguishable.

4.3 RQ1: Subjects' Use of EVO

We consider RQ1: To what extent are subjects able to learn EVO, and then use EVO to answer goal modeling questions? Given the results in Sect. 4.1 and Sect. 4.2, we investigate this question between-subjects. In Parts 2–4 (see Tbl. 5), each subject completed the EVO Training module and answered questions about the Bike and Summer models (see Tbl. 4), one using the EVO feature and one without. Thus, we compare the EVO training module and the results of each model separately. We divide RQ1 into two sub-questions: (a) Is our training

[3] In [4], we reported either the minimum and maximum times or the first to third quantile. For better consistency, in this article, we exclusively report times as the first quantile rounded down to the nearest half minute and third quantile rounded up to the nearest half minute.



(a) Goal Modeling Training Timing Data (in Seconds) and Scores (in Counts out of a Maximum of 14)

(b) EVO Training Timing Data (in Seconds)

Fig. 8: Goal Modeling and EVO Training Boxplots

Table 9: EVO Training Score Frequencies for Experiment (Ordered by Group) and for User Study

EVO Training	Score Frequencies		
	0-4	5	6
`EBk-XSm` & `ESm-XBk`	0	4	12
`XSm-EBk` & `XBk-ESm`	0	3	13
User study	0	0	11

sufficient for learning how to use EVO? and (b) To what extent were subjects able to answer questions with and without EVO?

(a) EVO Training. All subjects completed a common EVO training module consisting of six questions. We combined treatment groups `EBk-XSm` & `ESm-XBk` (i.e., EVO training in Part 2, see Tbl. 5) and `XSm-EBk` & `XBk-ESm` (i.e., EVO training in Part 3), to understand if there were any effects in reviewing one of the experimental models (i.e., Bike or Summer) first. Tbl. 9 lists the score data for the EVO training. All subjects achieved a score of 5 or 6 (out of a possible 6), and thus, achieved a satisfactory score (i.e., 70%). Fig. 8(b) shows the box plots for the training and test times for the EVO Module. Fig. 8(b) includes the combined treatment groups, as well as the data from our User study, which we discuss in Sect. 5.1. Most subjects (rounded first to third quantile[3]) took 3–4 minutes to review the training materials and 1–2 minutes for the EVO questions.

Again, subjects were asked to document any questions they had after reviewing the EVO training, with nine subjects asking a question. Questions focused on understanding the simulation results and the differences between the EVO modes. Two subjects asked about the order of the Percent (%) mode, which was further clarified. Thus, subjects learned and demonstrated proficiency in using EVO in under ten minutes.

(b) Answering Questions with EVO. We now review subjects' ability to answer the model questions listed in Tbl. 4. Q4 and Q6 were each scored out of 3, one for each sub-question. Q9 and Q12 were excluded from scores as they were used to validate the answers of Q8 and Q11, respectively. Thus, each model was scored out of 14.

Tbl. 10 lists median scores for each treatment group. Scores ranged between eight and fourteen for the Bike model, with a median score of thirteen. Scores for the Summer model ranged between nine and fourteen, with a median score of twelve. Given these ranges, we note that two scores did not achieve a satisfactory level (i.e., 70% or 10/14). In both of these cases (one each for the Summer and Bike models), the subjects were not using EVO. Thus, when subjects used EVO, they achieved a satisfactory score.

Overall, EVO produced a slightly better median for the Bike model but also a slightly worse median for the Summer model. The questions answered best by subjects were Q1, Q3, and Q5 (see Tbl. 4), with only one subject incorrectly answering each question between both the Bike and Summer models combined. The worst performing question was Q6(b) for the Summer model and Q6(a) for the Bike model.

Given the score data in Tbl. 10, we did not expect to find variations between groups.

Null Hypothesis 5 *There is no observable differences between treatment groups with respect to scores of the Bike model in Part 2 and Part 3.*

Null Hypothesis 6 *There is no observable differences between treatment groups with respect to scores of the Summer model in Part 2 and Part 3.*

We failed to reject both null hypotheses at the $\alpha = .05$ level ($\chi^2_1 = 0.44; p = .50$ for Hyp. 5 and $\chi^2_1 = 3.62; p = .057$ for Hyp. 6), and did not find any statistical difference between treatment groups with respect to the subjects' scores for Bike and Summer model questions.

Table 10: Median Scores (out of Fourteen) for Bike and Summer Questions in Experiment, with Bold Indicating EVO Use

Part 2 and Part 3		
Group	Bike Median	Summer Median
`EBk-XSm`	**13**	
`XSm-EBk`		13
`ESm-XBk`		**12**
`XBk-ESm`	13	
Part 4 (*Repeated Measure*)		
Group	Bike Median	Summer Median
`EBk-XSm`		12.5
`XSm-EBk`	**13.5**	
`ESm-XBk`	12	
`XBk-ESm`		**11.5**

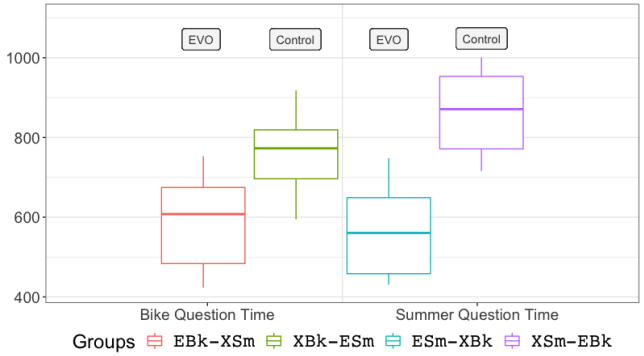


Fig. 9: Timing Data (in Seconds) for Answering Bike and Summer Questions (see Tbl. 4) in Experiment in Part 2 and Part 3, see Tbl. 5

We conclude that subjects were able to learn EVO, and then use EVO to answer goal modeling questions.

4.4 RQ2: Comparing EVO with the Control

Next, we consider RQ2: How does EVO compare with the control in terms of time and subjects' perceptions? We again break this research question into two sub-questions: (a) Does EVO help subjects make decisions faster? and (b) How do subjects perceive EVO?

(a) Bike and Summer Times. To measure subject completion times, we added their times from Pages 1, 2, 3, and 5 (see Tbl. 4). Pages 4 and 6 were excluded because they contained solely free form answers where subjects' time depended on the length of their answer.

The times for both models are comparable, ranging from five to twenty minutes. Fig. 9 gives the box plots

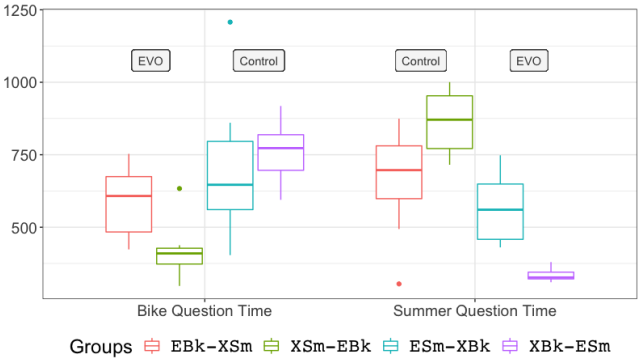


Fig. 10: Timing Data (in Seconds) for both Initial and Repeated Measures (Parts 2–4 in Tbl. 5) of Subjects' Answering Bike and Summer Questions (see Tbl. 4) in Experiment

for the initial measurements (i.e., Parts 2 and 3) of each treatment group (i.e., sequence) for the Bike and Summer model question times. In the Bike model (left side), `EBk-XSm` (red) used EVO to answer the questions and have visibly lower times. In the Summer model (right side), `ESm-XBk` (blue) used EVO to answer the questions and also have visibly lower times.

Null Hypothesis 7 *There is no observable differences between treatment groups with respect to times of the Bike model in Part 2 and Part 3.*

Null Hypothesis 8 *There is no observable differences between treatment groups with respect to times of the Summer model in Part 2 and Part 3.*

We reject both null hypotheses at the $\alpha = .05$ level ($\chi^2_1 = 6.35; p = .012$ for Hyp. 7 and $\chi^2_1 = 9.93; p = .002$ for Hyp. 8). We find the effect size to be large[4] for both ($\eta^2 = 0.38$ for Hyp. 7 and $\eta^2 = 0.64$ for Hyp. 8). Thus, we find the times are significantly faster when subjects used EVO. This finding further supports the results of RQ1 (see Sect. 4.3), as subjects were able to learn and actively use EVO to answer goal modeling questions faster.

For completeness, in Fig. 10, we provide the box plots for each treatment group for the Bike and Summer model, including the repeated measures (i.e., Part 4). Note that the colors and order associated with each group varies between Fig. 9 and Fig. 10. As discussed in Sect. 4.1, our study design is threatened by carry-over and learning effects. This is observed in Fig. 10, where the results are more pronounced when the control group used EVO (i.e., `XSm-EBk` (green) for the Bike model and `XBk-ESm` (purple) for the Summer model. We hypothesize that the interaction of subjects being

[4] A large effect is defined as $\eta^2 >= 0.14$.

Table 11: Subjects' Average (Mean) Rating of Difficulty with Three Aspects in Experiment (Where 0 was No Difficulty and 10 was Complete Difficulty): Understanding the Scenario Description, Understanding the Model, and Answering the Questions

	Scenario	Model	Questions
Phase 1	3.7	5.0	4.8
EVO	2.6	2.6	2.3
Summer	3.4	4.2	4.1
Bike	3.6	4.2	4.6

in the control group and using EVO may contribute to this additional benefit and that with additional training and experience the time savings of using EVO may be more pronounced.

(b) Qualitative Perspectives. Finally, we performed a qualitative analysis on the question, "Compare and contrast the colored views with the non-colored views, which do you prefer? Why?"[1]. All subjects preferred the EVO view over the control. More than half said that EVO was faster and/or easier to use. Other comments include that EVO was more intuitive, better for comparing models, and improved subjects' high-level understanding of the model. While no critiques of EVO were present in this question, we discuss subjects' recommendations for improving EVO in Sect. 4.5.

We conclude that subjects preferred using EVO over the control. Subjects' completion times were faster with EVO.

Table 12: Recommendations for Improvement from Experiment

EVO Improvements
- Add ticks or an outline to time mode. (x4)
- Choose prettier colors (and better fonts). (x2)
- Better contrast between text color and EVO color. (x2)
- Change conflict colors:
- All conflicts the same color.
- (P, P) should be grey, reduce visual noise.
- Use green/yellow for conflicting evidence pairs.
- Left to right arrow on time mode.
- Eliminate possible left-right bias in % mode.
- Colors may not be accessible to all users. (x2)

Goal Modeling Improvements
- Add goal prioritization in models.
- Organize models as decision tree.
- Improve visualization of links (maybe with color).
- Create model-level metrics (in a table).
- Distinguish between OR and XOR links.
- Make evolving functions more explicit.
- Add more possible values for (s, d).

Study Instrument Improvements
- Clarify difference between + and $+S$. (x2)
- Better explain evolving functions.
- Clarify difference between initial state and time point 0. (x2)
- Clarify difference between % and Time mode.
- Organize handout landscape with models left to right.
- Text too crowded/overlap, make images simpler/larger. (x2)
- Change "become Fully Satisfied" wording in Q6.
- (F, F) looks black, not dark purple.
- Add progress bar to questionnaire.

4.5 RQ3: Improvements and Recommendations

Finally, we address RQ3: How do subjects rate the study instruments and experience? To answer this question, we collected optional quantitative ratings after each module and qualitative reports at the end.

For each of Parts 1–4 in Tbl. 5 (i.e., the initial training sequence, the EVO training, the Summer model, and the Bike model), subjects rated their experience completing each part. They were asked to rate their difficulty with the three aspects (where 0 was no difficulty and 10 was complete difficulty): (i) understanding the scenario description, (ii) understanding the model, (iii) answering the questions. Tbl. 11 gives the average difficulty rating for each aspect and each part. Subjects had the most difficulty during the initial training phase, which seems appropriate because subjects had very limited familiarity with RE and goal modeling (see Tbl. 6, discussed in Sect. 3.4). Subjects perceived the Bike scenario and questions as slightly more difficult than the Summer model but perceived the models similarly. The EVO training was rated as the least difficult part, with average scores of 2.3-2.6. While this provides additional data for our assertions in RQ1, comparing between the scores in Tbl. 11 is confounded by the fact that the EVO training was the shortest module and built on the Phase 1 training.

Finally, we ask subjects for suggestions and additional comments. Specifically, to gather suggestions, we asked the question: "What suggestions or changes would you recommend to the developers of this goal modeling language (and tool)?" Tbl. 12 lists the recommendations provided by subjects, organized into three categories: improvements that can be made to EVO, goal modeling, and our study instrumentation.

Subjects made a variety of recommendations about improving the look and feel of EVO—from changing the colors of conflicting evidence pairs to adding ticks to show time points in the Time mode. We are aware

of the accessibility issues associated with color vision deficiencies (see Sect. 2.5) [6].

Since this study was conducted in isolation from tooling and other approaches, many of the goal modeling recommendations have already been investigated by other approaches. For example, goal prioritization, XOR links, model-level metrics, and quantitative valuations have all been investigated by researchers [16, 15,2,8]. We found the recommendation about improving the visual aspects of the links of interest and may pursue this in future work.

Finally, subjects recommended improvements to our study instrument. Subjects recommended clarifying the differences between link types, evolving function types, and the difference between the initial state and time point 0. Specifically, with respect to EVO, one subject thought more explanation was required to understand the difference between % and Time mode. Other comments included adding a progress bar and improving our study handouts and questions. Three subjects (excluded from Tbl. 12) encouraged the developers to implement the EVO feature.

Six subjects provided additional comments. Of these responses, three mentioned that the survey was long/hard, one said that they do not like goal modeling, one thought that (F, F) is the color black, and the final comment explained an inconsistency in the subject's answer to a previous question.

We conclude that subjects rated the study instruments and experience as suitable and not overly difficult; yet, roughly 10% reported that the study was long or hard. Subjects found the initial training most difficult and the EVO training easiest.

5 User Study Results

In this section, we describe the results of our User study and answer research questions RQ4-RQ6. In discussing these results we refer to an individual subject as S1-S11. For example, S5 would refer to subject number five after anonymization.

5.1 Preliminaries: Comparing Subjects Across Studies

In this section, we intend to compare the results between the Experiment and the User study. Thus, we begin by assessing whether subjects across studies perform similarly on the common components (i.e., the training modules). Using the analysis already described in Sect. 4, we divide this inquiry into two questions, which parallel Sect. 4.2 and Sect. 4.3, respectively: (a) Do subjects across studies perform similarly on basic goal model training tasks? and (b) To what extent are subjects able to learn EVO, and how do they compare across studies? As mentioned throughout Sect. 3, we used the same training material for both investigations[5].

(a) Goal Modeling Training. To understand the results of the modeling (and simulation) training (see Part 1 in Tbl. 5), we use the same methodology described in Sect. 4.2. Most subjects spent 12.6–16.97 minutes reviewing the training materials (i.e., rounded first to third quantile), and 7.14–10.25 minutes answering the training questions. Subjects scores ranged between 12–14 (out of 14). Thus, all subjects achieved a satisfactory score (i.e., 70%) on the training questions. Seven subjects asked at least one question about the goal modeling training, with four asking about contribution links, and one asking about the direction of decomposition arrows. Further, one subject asked about the origin of the $\perp$ symbol and another asked about how to avoid conflicting evidence pairs. Finally, one subject asked about the behavior of the Have a Less Stressful Semester element. Subjects also left comments. These comments varied but indicated subjects wanted to know more about modeling in practice, how the simulation works, and how individual functions affect the simulation results.

Fig. 11 shows the box-plots for the training times. Additionally, Fig. 11 contains the timing data for the EVO training, which we discuss later in this subsection. The initial training times and scores are similar to those from the Experiment. The goal model training time box-plot contains one unexpected data point, reporting that a subject took less than twenty seconds for training, but the researcher in the room does not recall any participant skipping the training. We keep this data point for transparency though we note that it does not affect the variance between groups.

Similarly, we compare the User subjects with each treatment group of the Experiment in Fig. 8(a). Note that the ordering of the plots varies between Fig. 8(a) and Fig. 11. The training times in the User study were comparable to each treatment group (see Fig. 8(a) left plot). Additionally, the completion times and scores for the training questions (see middle and right plot in Fig. 8(a)) were similar to those in the Experiment.

Null Hypothesis 9 *There is no significant variation between the study samples (i.e., the Experiment or User*

[5] Per the recommendations of the Experiment study (see Tbl. 12), we allowed subjects to choose black or dark purple as the color that represents full conflict (F, F) in the EVO training questions.

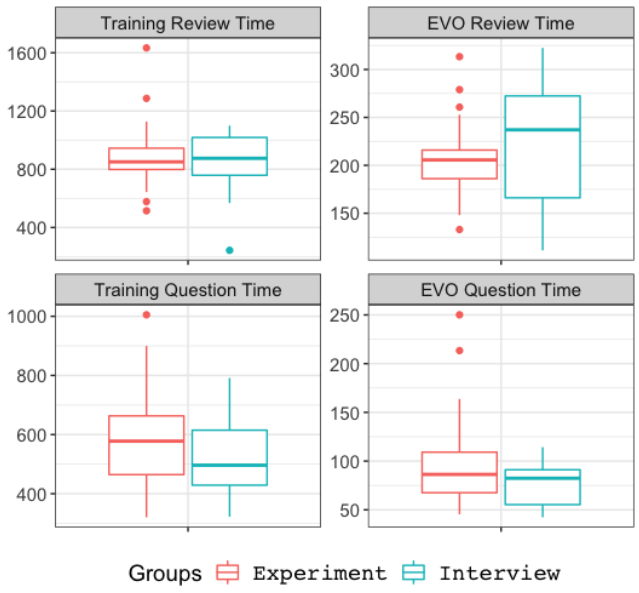


Fig. 11: Goal Modeling, Simulation, and EVO Modules' Timing Data in Seconds: Time to Review Training Materials in the Top Row and Data to Answer the Training Questions in the Bottom Row

study) in terms of performance (i.e., score) on the training.

Null Hypothesis 10 *There is no significant variation between the study samples (i.e., the Experiment or User study) in terms of completion times for the training questions.*

We failed to reject both null hypotheses at the $\alpha = .05$ level ($\chi^2_1 = 0.08; p = .78$ for Hyp. 9 and $\chi^2_1 = 0.95; p = 33$ for Hyp. 10), unable to detect variations between samples.

(b) EVO Training. Next we investigate to what extent subjects learned EVO during Part 2 of the study (see protocol in Tbl. 5). We find that all subjects successfully answer the EVO questions, obtaining a perfect score (see Tbl. 9). This means that subjects understood the default EVO color palette and how to interpret colored intentions when modeling. Most subjects took 2.5–5 minutes to review the EVO training, and took between less than a minute and two minutes to answer the EVO questions. Six subjects left a comment or question about the EVO training. Three asked about the modes (e.g., percent vs. time), and one asked about the color of Have a Less Stressful Semester. One subject inquired whether the conflicting colors made them hard to distinguish at a glance, while the final subject asked a more general question about whether, in practice, negative aspects of a scenario are added to the model. As mentioned in Sect. 3.1.1, subjects were not specifically given questions to test using Time or Percent modes.

Fig. 11 (right-hand side) contains the timing data for both studies. The top-right box-plot in Fig. 11 shows that the User study has a larger range of times for the EVO training, than the Experiment. This data is also present in Fig. 8(b), where the Experiment is subdivided into those groups that completed the EVO training module earlier or later in the protocol.

Null Hypothesis 11 *There is no significant variation between the study samples (i.e., the Experiment or User study) in terms of performance (i.e., score) on the EVO training.*

Null Hypothesis 12 *There is no significant variation between the study samples (i.e., the Experiment or User study) in terms of completion times for the EVO training questions.*

We failed to reject both null hypotheses at the $\alpha = .05$ level ($\chi^2_1 = 2.81; p = .09$ for Hyp. 11 and $\chi^2_1 = 1.71; p = 19$ for Hyp. 12), unable to detect variations between samples.

We found that User study subjects were able to learn to interpret the basics of goal modeling and the EVO palette. We conclude that subjects performed similarly on all training material for both the Experiment and User study.

5.2 RQ4: Subjects' Experience Modeling with BloomingLeaf

We now consider RQ4: To what extent did subjects engage in goal modeling and decision-making activities using BloomingLeaf? We divide RQ4 into two subquestions: (a) How do the subjects participate in extending a base goal model? and (b) To what extent were subjects able to evaluate goal modeling simulation results? The data for this research question was collected during the in-person modeling session (see Parts 3 and 4 of our protocol in Tbl. 5).

(a) Initial Goal Modeling. We evaluated how subjects participate in goal model building activities on the model drafts. Prior to the in-person goal modeling session, we used the subjects' answers to the pre-study questionnaire to create initial drafts of their models. Subjects picked a scenario from three provided prompts (based on the work by Cebula et al. [9]) or had the option to pick their own. Nine subjects explored a scenario where they were "choosing between jobs/opportunities after college" and two subjects (S1 and S6) looked at "choosing a club or organization to participate in". To

Table 13: Subject Data for User Study

Subject	Initial (Pre) Model			Change Events			Valuations Assigned	Functions Assigned	Simulations Generated
	Intentions	Links	Actors	Intentions	Links	Actors			
S1	14	18	3	1	4	0	9	0	5
S2	14	12	3	5	4	1	5	2	3
S3	19	25	3	0	2	0	2	1	3
S4	37	36	3	5	5	0	12	12	3
S5	19	21	4	2	4	0	4	5	2
S6	15	27	2	2	6	0	4	4	2
S7	23	25	2	0	3	0	8	7	2
S8	20	16	3	1	27	0	6	6	4
S9	33	50	3	1	7	0	10	10	2
S10	26	27	3	10	1	0	4	4	2
S11	20	21	3	0	4	1	3	4	3

maintain confidentiality, we named the main actor in each model *'Self'* instead of the subject's name. We refrained from making links where the relationship between nodes was unclear, leaving it up to the subject to flesh out.

During the interactive modeling session, we asked subjects to understand, evaluate, and extend our initial draft of their scenario model. Initially, subjects varied in their level of comfort with directly editing the model. Some subjects (i.e., S2 and S4) appeared to be more apprehensive about modifying the model (or making a mistake). Beginning with S4 and after, we assured subjects that it was their model and they could break it and make mistakes. We believe this permission led later subjects (e.g., S8) to immediately edit their model.

Tbl. 13 lists the information about each subject's model. The initial (pre) and final (post) scenario models for each subject are available in our online appendix[1]. The "Initial Model" columns of Tbl. 13 contain the number of links, intentions, and actors made by researchers. All subjects made alterations to the initial model. The "Change Events" columns in Tbl. 13 list the changes (counting additions, deletions, and other changes) made by subjects. As is evident from Tbl. 13, if there is a change event for an intention, there is likely to also be a change event for a link. While the initial and final models for some subjects (e.g., S5, see online[1]) have the same number of intentions or links at the end as in the beginning, this does not mean that the model was unchanged. For example, S4 created a new intention and link but then decided to delete it. S5 added a new task and deleted an existing soft goal. All subjects implemented changes to the model links, as unclear relationships were not represented in the initial scenario model. Most subjects changed the model before simulation. An exception to this is S5, who first simulated the model before making any changes. Many subjects went back and forth repeatedly simulating and editing the model. These changes show that all subjects were actively engaged in the modeling process.

When necessary, the researcher assisted subjects in adding User-Defined functions or absolute time assignments to the model, which was beyond the scope of the study training. This allowed the subjects to represent their model more accurately. S6 and S8 both used two user-defined functions. S5, S6, S8, and S9 used absolute time points and constrained evolving functions with absolute time points. Additionally, researchers assisted S4 and S9 with assigning evolving functions, though given the size of their models, this was done to save time (e.g., S9 was choosing between seven trade-off options). Thus, subjects were able to use goal modeling and the BloomingLeaf tool, and with the assistance of researchers, subjects were able to use advanced features.

(b) Analysis of Simulation Results. All subjects ran at least two simulations (see "Simulations Generated" column in Tbl. 13 for full list). Additionally, each subject added at least one initial evaluation or evolving function prior to running a simulation (see "Valuations Assigned" and "Functions Assigned" columns in Tbl. 13 for counts). In some cases, the researcher had assigned initial value(s) to the base model and asked the subject to evaluate them. This occurred with S2, S3, S4, S5, S6, S7, and S11. Some subjects asked to generate a simulation with the functions that were present in the initial model before adding their own evolving functions. The first subject, S1, created four simulation paths with only one time point in the path and wished for additional time points. The researcher then assisted

S1 in adding additional time points. Adding additional time points was not sufficiently covered in our training materials; thus, to mitigate this issue for the remainder of the subjects, the researcher added at least three time points to each simulation request.

Overall, subjects were able to interpret the simulations, assigning meaning to the results and evaluating multiple scenarios. Subjects asked the most questions about the meaning of conflicting values, which the researchers clarified. In some cases, subjects had to be corrected when misinterpreting evidence pair labels (e.g., swapping the s and d values, see Sect. 2.1), which we discuss in Sect. 5.3.

Eight out of eleven sessions ended "conclusively", meaning that the subject was able to make a realistic decision based on the model and simulation results. Two sessions (S8 and S9) ended due to time constraints. However, both subjects expressed an interest in continuing to simulate results and had ideas for a future direction. Between filling out the per-study questionnaire and the in-person modeling session, S6 had already made a decision; thus, S6 used the modeling session to validate their decision and see the results of predictions in the model's evolutionary information.

Finally, the subjects varied in whether they mapped the simulation result to real-world time. For example, S5 mapped each tick in the timeline to one month. This difference may be related to whether the subject assigned absolute times to transitions in the evolving functions. Subjects S6, S8, and S9 used absolute assignments in their model. S6 specified that the time points were over three months in a semester while S9 described the first few time points as the current semester and later time points for becoming a lawyer. However, S8 assigned absolute time points in their model but did not specify a real world mapping (e.g., weeks, months, year). The time points were used as markers for whether they worked more or less hours in a given week.

We conclude that subjects were able to actively participate in goal modeling activities in BloomingLeaf to evaluate real-life scenarios and make decisions. All subjects made changes to the initial model and evaluated at least two simulation results, and most sessions ended with the subject making a decision about their chosen scenario.

5.3 RQ5: Subjects' EVO Use

We answer RQ5: How do subjects perceive and use EVO during an in-person goal modeling session? We again divide this into two questions: (a) How did participants use EVO? and (b) To what extent did participants find the EVO extension beneficial?

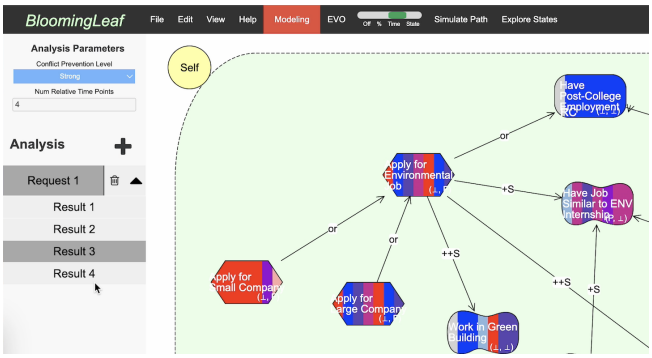


Fig. 12: Screenshot of S10 from User Study using EVO Time Mode to Understand BloomingLeaf Results

(a) EVO Usage. Out of our eleven subjects, eight turned on EVO when analyzing simulation results for at least a portion of the goal modeling session. Three subjects (i.e., S3, S7, and S8) did not turn on EVO at all. Of the eight that used EVO, all began with the default Blue-Red palette. Only S1 changed the palette, by turning on the Red-Green palette after using the default. However, they expressed confusion that red was satisfied and green was denied and created their own custom palette (shown in Fig. 4(d)), where satisfied is dark green and denied is red. S1 had indicated in their pre-study questionnaire that these are the colors they associate with good and bad things happening, respectively. The subject used the custom palette for the remainder of the session. The remaining subjects used the default palette for the entirety of the session.

In terms of EVO mode usage, all EVO subjects used State mode in some capacity when evaluating a result in Analysis mode. S1 used Percent mode briefly before switching to State mode, after expressing confusion at the stripes, and then used Time mode briefly while evaluating a simulation result. S10 used both State and Time modes at different points in their analysis, using Time mode to get a high-level overview of different Analysis results before using State mode to select a particular result and time point as their chosen path. Further, S10 was able to compare simulation paths using Time mode. Fig. 12 shows a screenshot of BloomingLeaf in Analysis mode, with S10 clicking between Results 1-4 and reviewing the simulation path in time mode. Only a fragment of the model is shown in Fig. 12, but S10 reviewed the entire model when switching between results. Overall, subjects appeared to prefer EVO state mode in order to walk through individual time points.

Of the three subjects that did not use EVO in their sessions, S8 was able to navigate the model's results and

evidence pair values well. However, S3 and S7 expressed confusion in reviewing evidence pairs. S3 asked for clarification on conflicting values and whether one is more satisfied or more denied, while S7 mistook a partially satisfied value for partially denied, misremembering the order of the formal notation. Using EVO may have assisted these subjects in this case.

Overall, when subjects did not use EVO, they focused on one node at a time. For example, in the screencast recordings (e.g., S7, S8) we saw subjects moving their mouse back and forth between each node they examined. Subjects that did not use EVO tended to talk about one or two intentions at a time, whereas subjects that used EVO talked about the model as a whole. For example, S4 did not initially use EVO in analyzing a simulation result. Their model was large, with 37 intentions and 36 links initially. While evaluating the model without EVO, S4 explained that they were focusing only on the satisfaction values of the four task nodes or 'options' for their decision. Turning on EVO in the middle of the session allowed S4 to evaluate other intentions as well, including their soft goals. In this case, using EVO to evaluate the results of a large model allowed the subject to broaden their focus and evaluate the entirety of the model, not just parts of it. However, we note that when a few subjects used EVO, they compared to what extent the intentions in the model were fulfilled (e.g., all, most, or some) without examining the relative importance of each intention. Thus, there may be a positive relationship between the size of the model and the benefits of using EVO, but there may also be other issues when reviewing large models.

(b) EVO Perceptions. In the study debrief, all subjects stated that they liked the color view. This included the subjects who did not use it. S3 wrote that the "colored view would be a bit easier than trying to squint and see the evidence pairs that may be hiding in corners". S3 expanded on this further by saying that some text and evidence pairs had overlapped, making it difficult to read; but, they did not state why they did not choose to use EVO despite seeing its benefits. S7 stated they preferred the color view because it is easier to look at, but also said that they are still digesting the modeling language and had only just learned the tool when asked about what changes they would recommend. Thus, their decision to not use EVO may have been due to feeling overwhelmed by using a new tool. It is possible that more experienced goal modelers or BloomingLeaf users would be comfortable with using EVO immediately. S8 also liked the colored views, writing it is "clearer to tell what is happening" in the study debrief. S8 also wrote that they realized they had not used EVO during the in-person modeling session, saying "Maybe it would've been too overwhelming and [they were] able to synthesize the data". This relates to S7's comment on having just learned to use the tool. As mentioned above, the advantages of employing EVO on small to medium-sized models may be limited.

Of the subjects that used EVO during the in-person session, all responded positively. S2 expressed that they liked seeing the colors. S11 referred to an intention's evidence pairs by color, saying "more blue or more red" instead of "more satisfied or more denied". Similarly, S9 stated in the study debrief that they wanted to avoid red outcomes. S10 stated that they enjoyed seeing EVO in Time mode to get a consolidated view, indicating it made the most sense to them.

Other feedback included that the formal notation takes more time to read, and EVO is easier than "squinting". EVO allowed the subjects to see valuations immediately, making analysis results easier to navigate. S5 noted that initially, the color was confusing, but once they learned, it made reading and model comprehension simpler. This may relate to S8's choice to not use EVO despite recognizing its value. Finally, S6 noted that they liked Time mode, but not Percent.

> We conclude that all subjects responded positively to EVO, with a preference for State (and Time) mode. Subjects who did not use EVO in the in-person session saw its benefit as well, though initial hesitation among beginners may have deterred them.

5.4 RQ6: Subjects' Ratings

Finally, we ask RQ6: How do subjects assess the in-person session and BloomingLeaf? In Sect. 4.5, we discussed how the Experiment subjects rated the online Qualtrics training. In the User study, subjects again rated the difficulty of the initial training sequence and EVO training across three aspects (where 0 was no difficulty and 10 was complete difficulty): (i) understanding the scenario description, (ii) understanding the model, (iii) answering the questions. Tbl. 14 lists the average difficulty ratings for both parts. This data is consistent with the Experiment study. Subjects had more difficulty with the first training than with the EVO training. In looking at this data more, two subjects rated both training modules exceptionally difficult, giving all three aspects of the EVO training a rating of 9 or 10 out of 10. Yet, unlike the Experiment, the qualitative data did not give any indication that any participant found the study difficult. It is possible that these two subjects misread the scale. We did not require subjects to rate the in-person modeling session, which was an oversight in our study design. If we conducted this or

Table 14: Subjects' Average (Mean) Rating of Difficulty with Three Aspects in User Study (Where 0 was No Difficulty and 10 was Complete Difficulty): Understanding the Scenario Description, Understanding the Model, and Answering the Questions

	Scenario	Model	Questions
Phase 1	3.7	4.7	4.5
EVO	2.7	2.9	2.5

a similar study again, we would have subjects rate the difficulty of using BloomingLeaf and participating in the modeling session. We asked subjects whether the session impacted their decision-making process at all. Specifically, we asked subjects to "[r]eflect on how participating in this study may have altered [their] thinking about [their] chosen topic"[1]. Only one of the subjects in the User study (S8) changed their opinion about their scenario topic, though most did not have a preferred option prior to the study. S8 found that they could achieve their goals by doing the options one after the other, rather than thinking of them as an exclusive "or". Subjects who did not change their opinion gave a variety of reasons for how the modeling session altered their thinking. For example, S1 said that they can "spend less time thinking about other missed opportunities or closed pathways after considering [their] decision through goal modeling". Thus, it appears that the session helped S1 with decisiveness. S4 stated that "the study supported [their] initial prediction of [their] decision. But now that it confirmed [their] prediction, [they] feel more confident in the decisions [they will] make in the future". S5 felt that goal modeling helped clear their mind, made them more aware of possible obstacles, and reduced their anxiety. S6 stated that while their views have not changed, "it's been very very helpful to see all of [their] internal calculations modeled on 'paper' in a way that [they have] never really considered before". Lastly, S11 wrote that the session "helped [them] draw connections between [their] goals and tasks and it helped [them] to understand the impact that each of [their] decisions have on [their] goals". Thus, it seems that overall, the session supported subjects' decision-making process. However, S10 pointed out that while "it is great to have all the influences" in the model, it gets "messier and messier as we add more and more tasks and soft goals".

We asked subjects to suggest improvements to the developers of BloomingLeaf and recommend changes to the underlying goal modeling language. Additionally, we asked them which aspects of BloomingLeaf were the easiest and hardest to use, and which features they used the most often. Tbl. 15 summarizes the subjects' responses to these questions.

Table 15: Subjects' Reflections on User Study

BloomingLeaf and Language Improvements
- Improve visualization of links (maybe with color).* (x2) - Make gear icon size consistent across all links. - Automatically layout and hide model elements. - Automatically adjust node ensure no overlapping text. - Add tool tips or embedded tutorial. - Improve explanation of Percent mode.* - Make all conflicting values the same color.* - Improve the process of editing and re-running a model. - Distinguish "Simulate" button from menu tab.
Easiest Features in BloomingLeaf
- Adding intentions to model. (x5) - EVO colors. (x3) - Adding links to a model. (x3) - Evaluating results through time points. (x3) - Editing the goal model. - Switching between Analysis and Modeling modes.
Hardest Features in BloomingLeaf
- Adding evolving functions.* (x3) - Making and editing links.* (x3) - Editing the shape of an intention. - Creating a simulation. - Understanding how time points are generated. - Using the items in the "Settings" bar. - Switching between Analysis and Modeling modes. - Setting constraints between time points. - Understanding the model.
Features Most Used during the In-person Session
- Running simulations. (x5) - Adding or changing links. (x4) - Adding or changing intentions. (x3) - EVO State mode. (x3) - EVO Time mode. (x2) - Exploring analysis results with the slider. (x2) - Setting initial values for intentions. (x2) - Modeling mode. (x2)

In terms of improving the user experience of BloomingLeaf, we found that we should focus future work on improving the links. Three subjects found adding links to be easy, while three found this hard. Three subjects recommended improvements to the visualization or use of the links. On the whole, subjects found EVO easy to use; yet, they either do not see the value in percent mode or find it confusing.

We expected and confirmed that subjects would respond that one of the hardest parts of BloomingLeaf was understanding the evolving functions and simulation results. Since the primary focus of our study was EVO, we did not dedicate time for in-depth training

on the evolving functions. For example, S5 noted that adding constraints on time points was difficult, which was not covered in the initial training. Additionally, some subjects found the process of editing and generating simulations confusing, with one subject (S1) suggesting a tutorial where the viewer can hover over elements themselves, instead of just watching a video. We are actively working on the automatic layout of models [48] and hiding model elements [7], which were both suggested again in this study (see Tbl. 15). These results are consistent with the recommendations from the Experiment (see Tbl. 12), although the improvements suggested by User study subjects were the direct result of using BloomingLeaf. In Tbl. 15, we denote items that were also recommended in the Experiment with an asterisk ('*').

Eight subjects left additional comments. Four (i.e., S2, S5, S8, and S9) mentioned that the experience was cool or fun, saying that it helped clear the mind and that they saw the benefit of making a model. Subject S5 noted that they are still not sure how to transfer an idea into a goal model, and the flexibility and openness of the choices make it overwhelming and difficult to navigate initially. S6 suggested it would be helpful to explain how the random assignment of nodes works. S9 said that the difference between Percent and Time mode was unclear, but that Time mode made more sense. Lastly, S11 said they enjoyed participating and wanted to continue editing the model.

We conclude that subjects found the modeling session helpful in mapping out their goals and making decisions. It helped subjects confirm their initial predictions or evaluate scenarios in a broader light.

6 Discussion

In this section, we first synthesize the results across our two studies. We then describe our lessons learned, compare the Bike and Summer model for the Experiment, and discuss threats to the validity of our investigation.

6.1 Synthesis of Results Across Studies

Through the dual investigation of the Experiment and User studies, we found that subjects all had a positive response to using EVO. The Experiment revealed that EVO allowed subjects to make decisions faster without impacting model comprehension. Supplementing this with the User study showed that most subjects chose to use EVO when using BloomingLeaf. The three subjects who chose not to use EVO wrote in the User study debrief that they still preferred having color as opposed to no color. We hypothesize that these three subjects did not use EVO (despite stating their preference for it) due to feeling overwhelmed by seeing the model and BloomingLeaf tool for the first time, or felt that their model was understandable enough without EVO. A future iteration of this study would include asking the subjects why they chose to use or not use any feature.

The results of our investigation are limited to the study context (see Sect. 6.3.4 for additional discussion). Given that we explore the use of EVO with untrained modelers, we cannot assert whether these results would hold for trained modelers. Perhaps trained modelers are already familiar with the evaluation labels used in the underlying language and would find the colors distracting. Perhaps they would find more benefit because they have an understanding of the analysis and tooling, and can easily look at the model and analysis results as a whole. A future study should explore how experienced modelers interact with EVO and BloomingLeaf.

Similarly, there is a risk that our results do not scale to realistically sized models. In the Experiment, we chose the models listed in Tbl. 3 because they are in domains suitable to a general audience and of a size that could be understood within the allotted time (i.e., 9–16 intentions, with a single actor). In the User study, we created larger more realistic models (i.e., 14–37 intentions, 2–4 actors) for each subject based on their interests. Yet, we cannot make broader claims about the effectiveness of EVO on industrially-sized models. We aim to explore the scalability of goal models and EVO to see whether there is a point in model size or simulation complexity at which subjects find value in using EVO. Further, we hypothesize that there may be an interaction effect between model size and subject expertise. For example, experienced modelers may find no value in using EVO on smaller models, where they can check the labels of intentions quickly and may find significant value in EVO when working with larger models. An eye-tracking study may help us differentiate the ways in which experienced and novice modelers interact with models of different sizes, as we can keep track of how many intentions, and which intentions, they focus on most.

In the Experiment, we found that using EVO does not impact whether a subject answered goal modeling questions correctly; yet, we observed in the User study that subjects focused on different parts of the model depending on whether EVO was turned on or not. For example, S4 appeared to focus on the main intentions in the model without EVO but evaluated the model in its entirety with EVO. Thus, we have anecdotal evidence that there may, in fact, be a difference in model comprehension, but we cannot say whether or how this result

generalizes. A future eye-tracking study investigating model comprehension may validate this observation.

Most subjects used the default Blue-Red palette. Only one subject explored non-default palettes, choosing to first use the Red-Green palette before creating their own custom Green-Red palette. This study has exposed the need for a Green-Red palette. Ten out of eleven subjects associate green with "good" outcomes and nine out of eleven subjects associate red with "bad" outcomes. Only two subjects associate blue with "good" outcomes[6]. This was surprising given our subjects had diverse cultural associations[1]. The training videos used the default palette, which may have contributed to the fact that subjects used the Blue-Red palette despite associating other colors with good events. In particular, the coloring scheme for conflicting values was explicitly explained. Not introducing the other palettes with the same level of depth may have discouraged subjects from using it. Thus, if the default palette was not intuitive at first, being trained in its use means that subjects were able to learn and use it easily without feeling the need to use alternate palettes. In the future, we would also inquire about subjects' reasoning for their palette choice.

Subjects in the Experiment were able to answer goal modeling and simulation questions both with and without EVO. In the User study, subjects were able to extend the initial goal model and evaluate its results. They were able to make their own interpretations and decisions based on the results. Thus, we learned from this that goal modeling and the BloomingLeaf tool are accessible to subjects after being trained in its use. More complex implementations, such as User-Defined functions, were used with the assistance of the researcher at hand. This allowed subjects to represent their scenarios and capture their ideas more accurately. The response from the User study was overall positive, both in terms of overall experience and modeling specifically. One subject in the Experiment study stated that they did not like goal modeling (see Sect. 4.5). This may have been due to the fact that the tasks in the experiment were restrictive, making the goal modeling experience less personal or enjoyable. The User study allowed subjects more freedom to explore what they wished and likely gave for a more satisfying experience. The User study allowed us to explore how subjects use goal modeling in a more realistic and dynamic setting, as opposed to viewing static pictures. However, we cannot verify this interpretation because we did not ask subjects for an explicit rating of the User study experience. Most subjects provided their thoughts on the experience without explicit prompting. Overall, subjects were able to grasp the concept and purpose of goal modeling and were able to use it to examine various scenarios.

[6] Subjects could choose more than one color.

6.2 Lessons Learned and Implications for Research

Subject Background and Recruitment. We developed the Experiment study instrument over a six-month period. We first iterated the instrument with individuals in our lab, then completed a small pilot with four subjects to evaluate the quality of our instrument and understand what timing data was generated from our Qualtrics® XM platform. The pilot helped us improve the quality of the data we collected. We added opportunities for subjects to take breaks and originally collected one timing value for Q1-12 in Tbl. 4. We discovered these values varied dramatically based on how much text subjects entered in the free form questions. As listed in Tbl. 4, we separated these questions across six pages (see Page column) and added timing information to each page. It was extremely difficult to recruit subjects for a survey that took a full hour. Due to Smith College policies and U.S. tax legislation, we were not able to offer remuneration in an amount over $25 USD. We launched three separate iterations of the Experiment. The first two iterations were meant to be completed by the subjects in their own time using the Qualtrics link. In our first iteration, we emailed researchers within the goal modeling community and targeted trained modelers. We received five responses and of these, only one completed the study instrument. Our second attempt was to recruit subjects within a large software engineering class with Tropos instruction at another institution, again receiving only one completed response. After two unsuccessful attempts, we pivoted to an in-person lab study. We updated our protocol to include additional training and recruited students as described in Sect. 3.4. There may be a cognitive difference between participating in a one-hour in-person lab session as opposed to completing a one-hour online survey, even when remuneration amounts are the same. In the future, a hybrid approach could be considered. Scheduling an online session where a researcher is present to answer questions, but allowing the subject to complete the questionnaire without having to travel may have allowed us to recruit more subjects for the first two attempts. We had sufficient volunteers for our in-person version of the Experiment and felt this was an important lesson learned.

We developed the User study instrument and materials over five weeks. We completed the study in one attempt and used the same recruitment methods and

population as the Experiment. We re-used the training materials from the Experiment to maintain consistency across studies but fleshed out the pre-study questionnaire and added additional questions to the debriefing section, including a demographic question. Demographic questions were important as this is a user-centered study and we wanted to explore whether cultural background related to EVO palette choice. The length of the pre-study questionnaire may have discouraged potential subjects from completing the sign-up process for the study, as it required subjects to input lengthy answers and think about a scenario. However, this was necessary to generate the drafts of the subjects' model so that the in-person session could be spent editing and simulating results. We created an additional three minute training video to introduce BloomingLeaf, being cognizant of video length to limit additional training. We had scheduled session blocks to be an hour and fifteen minutes (the length of a Smith College class) but aimed to ensure that the entire session did not extend beyond an hour as subjects in the Experiment had expressed fatigue towards the end.

Improvements to the Study Instrument. We reviewed the questions and supplemental information from the study by Hadar et al. [23] and iteratively developed our study instrument. We encourage other researchers to use and adapt our survey instruments; thus, we report potential areas for improvement.

For example, in question Q4 of the Experiment study (for both the Bike and Summer models, see Tbl. 4), we asked "how many times over the simulation does the element become Fully Satisfied" which would have been better rephrased as, "how many time point(s) over the simulation is the element Fully Satisfied". It was sometimes difficult to achieve task equivalency. For example, the tasks in question Q8 (see Tbl. 4) are not exactly matched between models. The correct Q8 answer for Bike model was *none of the above* because no intentions fulfill Prevent Unloading in Bike Lane. To satisfy Exercise in the Summer model requires either Water-Weed-Enjoy Garden or Drive to and Play Soccer, but we did not include Drive to and Play Soccer as an option, intending subjects to select Water-Weed-Enjoy Garden. Since the Bike model had a *none of the above*, we included the same for the Summer question, yet this resulted in subjects choosing it because they wanted to select Drive to and Play Soccer.

In our analysis of the Experiment, we were unable to detect any differences between scores on the models with or without EVO. Future work is required to determine whether our study instrument is sufficiently discriminatory. One of the aspects we iterated on was the length and complexity of the questions we asked in this study. We opted for a balance in these factors to ensure that subjects would complete the study in one hour, which we agreed upon as a reasonable upper bound.

User Study Reflections. We developed the User study questions to be flexible in order to respond to how subjects used the model and tool. Thus, depending on the subject, we did not feel the need to ask all questions that were listed in the study protocol. As the researcher was present to guide the in-person session, it was important to establish a comfortable rapport with the subject to ensure their communication and participation. We had the same researcher lead each in-person session to maintain consistency. After the first few subjects, we learned that we needed to reassure subjects that the model was fully theirs to modify and that they should not be hesitant to make changes or mistakes. We verbally encouraged subjects to make changes and comments throughout the session.

The shortest modeling session lasted 17 minutes and the longest lasted 35 minutes. While we aimed to complete the entire session in an hour, the time subjects took to complete the training portion of the study varied resulting in less time for the modeling session; thus, we were not able to spend the same amount of time modeling with every subject. However, since we collected qualitative information and stopped when the subjects felt comfortable, the variation in time was acceptable. For example, S10 stayed past the hour time limit, as they wished to continue modeling the scenario. We also observed that depending on the completeness of the initial model, subjects spent more or less time in the modeling mode. With more complete models, we were able to spend more time on analysis.

During the session, we experienced some difficulties with the BloomingLeaf tool, such as the model being over-constrained during the model-editing portion with the subject. In these cases, the researcher was on hand to resolve these issues mid-session, which took away time from actively engaging with the subject. However, we did not discourage subjects from adding too many constraints, as we wanted to observe their behavior and see how they would use the tool on their own. We did assist subjects in understanding link directions, adding User-Defined functions, and setting absolute time assignments in the model. Covering these additional BloomingLeaf features not mentioned in the training showed the flexibility of the tool in representing subjects' needs and also displayed the level at which subjects were engaged in the session. Some subjects were more focused on extending and understanding the model, while others on modeling results and decision-making. We encouraged both. Overall, we felt that it

was important to support the subject in their decision-making and to not restrict their thought process.

BloomingLeaf Design. Conducting the User study discussed in Sect. 5 allowed us to observe users interacting with BloomingLeaf and gave us, as researchers, insights into improving its usage. As mentioned by S1, it would be helpful to have tool tips or an embedded tutorial in the tool.

Most subjects had issues creating links in BloomingLeaf and it was difficult for us to explain this verbally. We had to verbally and visually demonstrate that to create a link the user needs to click and drag then release the mouse button once a red box appears around the destination intention. Most users naturally clicked the link icon and then moved to the destination to click again, which created an erroneous link. S6 explained that they were familiar with modeling in Lucidchart (see *lucidchart.com*), so their intuition did not match BloomingLeaf.

To modify the type of relationship (e.g., and, +) users must hover over the link with their mouse and click the gear icon once it appears. This behavior was not intuitive to subjects as many wanted to click the link to enable the link inspector panel. Yet, clicking the links creates a bend-point in the link. This was only problematic when the subjects thought they had clicked a link because the link inspector panel was shown for a previously modified link; thus, it would be helpful if the link inspector listed the source and destination intentions. Further, the link inspector could give users insights about the meaning of relationships; specifically, the strength of contributions. While there is no direct mapping between the `--` link and the *breaks* link in iStar [12], using these words may help users distinguish between the `-` and `--` link types.

Finally, we can reduce the likelihood of conflicting valuations for intentions. Our backend algorithm creates a simulation path by randomly choosing satisfaction s and denial d values for each intention based on the constraints specified in the model (see Sect. 2.1). When relationships cause only evidence for [resp. against] the fulfillment of an intention to be propagated, the backend calculates the s [resp. d] value and randomly assigns the d [resp. s] value, which causes conflicting evidence pairs to be assigned. We intend to update the backend, to prioritize selecting no evidence $\perp$ when there is no evidence from propagation. For S9, we experimented with adding a resource called None to propagate *None* $(\perp, \perp)$ and reduce the number of conflicting evidence pairs in the model. This workaround was very helpful but looked awkward.

Statistical Methods. Given our per group and study sample size, any statistical test will have lower power to make conclusions (see Sect. 6.3). In Sect. 4 and Sect. 5, we used the KWRS test to evaluate if there are distinct groupings within our sample data and compare between study samples, respectively [37]. It is important to note that the KWRS test is an omnibus test statistic, meaning that while it determines if there is a statistically significant difference between groups, it does not infer more than that, i.e. it does not determine which groups differ. The KWRS test is valuable for small sample sized data because it does not make assumptions about the distribution of the data and is not influenced by data points that vary greatly in magnitude, which is useful for time data. However, not making assumptions means that non-parametric tests such as the KWRS are less powerful than parametric tests, which often assume a normal distribution. Additionally, we use between-subjects analysis using KWRS, which is less powerful than within-subjects analysis since individual variation is not removed. Where appropriate, we evaluate the effect size of our KWRS analysis through the eta-squared (η^2) test for Kruskal-Wallis which is calculated from the H-statistic (i.e., χ^2) [44].

Our analysis was done between-subjects using KWRS due to the presence of a carryover effect within our repeated measures (see Sect. 4.1). We evaluated for a carryover effect using a linear mixed effects model and found *order* (i.e., period) to be significant. Due to this, we drop the second period of the Experiment study and analyze the first period between subjects, though we include data on the repeated measures for completeness.

Finally, we evaluate our sample size using a statistical power test for repeated measures with a medium effect size and find that the minimum sample size using G*Power [14] for our experiment was 56. Thus, the Experiment has low statistical power. In future studies, we recommend recruiting at least 56 subjects. Overall, the limiting factor for our analysis was sample size, which reduces the power of our conclusions.

6.3 Threats to Validity

We discuss threats to validity using the categories in [50].

6.3.1 Conclusion Validity

We wrote scripts to analyze our data wherever possible and automatically recorded page completion times to ensure reliable measurements. Qualitative data was randomized before review and categorization. Different authors conducted the in-person and data analysis components to reduce researcher bias. To mitigate

variations in treatment implementation, we standardized the experimental setup by using our online platform, videos, and pdf handouts to ensure that the subjects had equivalent training materials (see Sect. 3.2), and maintained our laboratory setup throughout the study period, to ensure a consistent in-person experience across both studies.

In the statistical analysis, there may be a risk of Type I errors ("false positives") as we do not adjust our p-value for multiple testing, instead choosing to test our hypotheses at a set α value of .05. While our between-subjects analysis means that individual differences may threaten validity, we do not believe there is a random heterogeneity of subjects risk, since our population was homogeneous, having similar knowledge, abilities, and previous experience with English, Tropos, and RE (see Tbl. 6). In a future study, we would collect data about subjects' year in the undergraduate program (e.g., first-year, seniors) to further mitigate this risk.

Experiment. The main threat in the Experiment is low sample size. Having 32 subjects spanning four treatment groups is considered a low sample size, with a minimum sample size according to G*Power analysis being 56. We may have experienced a reliability of measures threat, as subjects asked questions about the wording of Q6 (see Sect. 6.2).

User Study. There is a risk of a reliability of treatment implementation threat for the User study. Each subject had a customized model for their scenario and was asked different questions based on their interests. We made slight improvements throughout the study to mitigate evaluation apprehension (see Sect. 6.3.3). To partially mitigate this threat, we had two researchers jointly create the initial scenario models for each subject. These models all followed a similar structure, where the subject's primary goal was decomposed into possible tasks and these tasks contributed to the subject's soft goals.

6.3.2 Internal Validity

In both studies, our voluntary recruitment strategy combined with a cash remuneration may have caused a selection effect.

Experiment. We explicitly designed our study to control for a learning effect or maturation risk (i.e., where one group learns a treatment faster than another). We gave opportunities for subjects to take breaks if they were fatigued and shortened the instrument wherever possible. We controlled for an instrumentation effect of the experimental objects in our crossover design; yet, the Bike model questions may have been slightly harder (see Sect. 4.3). With this design, there is still a risk of carryover effects [46]. To our knowledge, no subjects used BloomingLeaf or EVO prior to the study.

User Study. As this study was run at the same institution with the same undergraduate population six months after the Experiment, we may have had subjects who participated in both studies. This means that subjects may have had prior exposure to the goal modeling, simulation, and EVO training materials. However, Tbl. 6 suggests all subjects had similar knowledge, abilities, and previous experience with English, Tropos, and RE. Thus, this threat may have been mitigated by the time between studies (i.e., an entire class year graduated). We may also have single group threats, as we cannot tell if using EVO assisted subjects in their analysis or if they would have performed similarly without it. We aimed to mitigate this by asking subjects whether they preferred the color view or not. We do not consider a maturation effect to be a threat in the User study, as learning was part of the process. Given the possibility of repeated subjects, we may have experienced an additional selection effect.

6.3.3 Construct Validity

For both studies, some students who took a software engineering course may have scored better overall; yet, our common training protocol may have limited this threat. We collected data in multiple forms (e.g., scores and times) and asked different types of questions to mitigate mono-method and mono-operation biases. Additionally, the results of the Experiment and User study taken together mitigate this threat.

As always, we have threats of *hypothesis guessing* and *evaluation apprehension*. In both studies, some subjects expressed nervousness asking if they needed to review data structures or read about goal modeling before participating. For example, `S7` in the User study appeared uncomfortable with the open structure of the questions in the modeling session.

Experiment. We conducted multiple pilot mini-studies (not discussed in this article) to ensure that our study instrument was measuring our intended constructs. In one such study, we found that our unit of time measure was inaccurate because it included too many questions; hence, we divided the questions across multiple pages as listed in Tbl. 4 and isolated qualitative questions.

User Study. Subjects may behaved differently in the in-person modeling session due to knowing they are part of an experiment or may have felt the need to make a decision in order to end the session in compliance with experimenter expectancies. The personal nature of the modeling scenario may have created an additional risk of experimenter expectancies [11]. We tried to reduce

these threats by keeping our questions open-ended with no expectation of a positive response from subjects.

6.3.4 External Validity

Given the contrived nature of our study setting (i.e., one-on-one in our lab), our study was not reflective of the use of EVO or goal modeling and the BloomingLeaf tool in the "real world". We conducted the Experiment using a survey instead of embedding EVO within BloomingLeaf, while the User study had subjects use BloomingLeaf in-person.

The length of the modeling session in the User study, ranging from 17–35 minutes, was shorter than a typical modeling session with stakeholders. In the real world, the evaluation of models would have taken place over multiple modeling sessions, allowing stakeholders to revisit past decisions. Thus, our User study was not reflective of how a typical stakeholder session would go due to time constraints. Additionally, due to constraints over participant time, we were unable to validate EVO on large models that are more reflective of "real world" scenarios.

Our homogeneous population of undergraduate students means that we cannot generalize to the broader RE population, but given the limited prior knowledge of our subjects (see Tbl. 6), these results may, in fact, generalize. As already introduced in Sect. 6.1, additional experiments with different populations, problem domains, and larger models for scalability are required.

7 Related Work

Recent work has critiqued the adaptability of GORE approaches [31]. In this paper, we address this gap by improving the interpretability of Tropos evidence pairs. As already introduced in Sect. 1, Hadar et al. [23] and Siqueira [42] studied the comprehensibility of Tropos models with respect to Use Case models. While it is difficult to compare our results with these studies because we only evaluate Tropos models, this work was influential in the design of our study and the importance of controlling for the use of different models, while investigating the performance of subjects on analysis tasks.

Using color as a technique to improve visualizations of goal models has been a topic of recent interest within the community. Amyot et al. used colors to visualize analysis results in the jUCMNav tool for URN [2], while TimedGRL used color in heat maps to visualize evolving GRL models [3]. Both used green and red to denote the satisfaction and denial of intentions, respectively, based on the colors of a traffic light. Varnum et al. proposed using colors to help stakeholders interpret the evidence pairs used in Tropos for intention evaluations [45] (see Sect. 2.3 for details). At the same time, Oliveira and Leite proposed mapping the primary colors onto NFR soft goal labels and contribution links, allowing color values to be quantitatively calculated and propagated throughout the model [36]. Varnum et al. used a static set of colors; whereas, Oliveira and Leite use a large range of colors calculated dynamically.

In reviewing these approaches, we chose to first validate the coloring approach of Varnum et al. because of its static nature, which made it easier to evaluate experimentally and understand whether color was an effective approach. Further research is required to validate the choice of colors in both approaches, and whether the dynamic nature of Oliveira and Leite's approach causes an additional cognitive load that reduces the overall effectiveness.

As introduced in Sect. 2.5, Ben Ayed et al. [6] extended the work of Varnum et al. [45] to allow users to choose the color palette beyond the default blue-red palette. In the User study (see Sect. 5), only one subject created their own palette (shown in Fig. 4(d)), creating a green-red palette similar to a traffic light used by jUCMNav [2] and TimedGRL [3].

We built on the methodology of similar studies in RE for our Experiment (see Sect. 3 and Sect. 4), and followed the guidance in [41], [46], and [50]. Winkler et al. reported on a between-subjects crossover similar to ours with sixteen subjects [49]. The authors assumed that the treatment group had increased precision and a reduction in time to complete the tasks due to working with direct output from the tool; whereas, the control group completed the task manually. We attempted to control for differences in tool usage by providing both groups with direct output from BloomingLeaf. Noel et al. conducted an experiment with 28 undergraduate students, also using a crossover design. In their design, they specified the modeling method used as a factor and the experimental problem used as a blocking variable to isolate task influence [34]. Ghazi et al. reported a study comparing two navigation techniques for requirements modeling tools [17]. They used time limits to motivate the participants to work as fast as they would on real tasks in industry, giving the subjects about five minutes to try out the tool. However, this may force subjects to work faster, which may result in worse results. To prevent this, we let the subjects take the time needed to review the training documents since our population comprised new learners. Santos et al. presented a quasi-experiment to explore the interpretability of iStar models given different concrete syntax [38]. Subjects were tasked with identifying defects in a goal model, a task we did not include in our study as it may have

been too difficult for new learners and increased their fatigue.

For the User study, we conducted an *experimental simulation* [43] to mimic stakeholders engaging in GORE activities with a trained modeler. In designing the modeling session, we reviewed the work of Horkoff and Yu [27], who examined interactive analysis of iStar and the later extension of iStar by Horkoff et al. to incorporate creativity triggers [26]. We also reviewed the work on eliciting contribution relationships by Liakos et al. [28]. We compared each of these approaches with the methodology proposed for the Evolving Intentions framework [22]. Additionally, we examined and built upon the existing literature on semi-structured interviews and interviews in requirements elicitation. Hadar et al. looked at the importance of domain knowledge in requirements elicitation interviews [24]. Zaremba and Liaskos described the importance of effective probing methods to elicit subject responses [52]. We chose to use subject-defined scenarios to explore how subjects make decisions in a personal context, based on the RE literature and feedback from the Experiment study.

We built on the work of Cebula et al., who studied how eight novice Tropos modelers create goal models for decision-making [9]. For a given scenario, four created a model by hand, while the other four used BloomingLeaf. The researchers also constructed a model using a pre-study questionnaire for the same scenario. The subjects were then asked to compare the subject-generated and researcher-generated models and choose their preferred one. The subjects' preference for models were mixed, showing that researcher-generated base models are adequate and can be used (and in some cases, preferred) by users making decisions. While Cebula et al. tried to create a near-complete model based on user responses, we took a different approach and tried to create the minimum base model subjects could expand upon in order to see how users interact with it. They also found that subjects were able to understand and use goal models, as well as extend the researcher-generated model and answer questions about major trade-offs.

Outside the field of RE, Clarke and Duimering [10] explored how users experience video games through a behavioral study. They interviewed eleven subjects both online and in-person using the "echo method" [39], with open-ended questions to elicit free responses from participants on various topics. Their questions were designed to investigate both social and technical aspects of the gaming experience, which is similar to our exploration of BloomingLeaf usage and decision-making. While Clarke and Duimering were able to conduct interviews over the internet, we conducted all of our sessions in-person to assist subjects in modeling and tool usage. Using role-playing, Shabtai etal [40] presents a behavioral study of users using a web-based interactive program, wherein they play the role of a banker approving loans. However, as noted previously, we refrained from fictional scenarios to explore the personalized aspects of decision-making.

8 Conclusions and Future Work

In this paper, we explored how using EVO to visualize evidence pairs impacts an individual's ability to reason and make decisions with goal models that evolve over time. To do so, we conducted a two-phased IRB-approved investigation, first with an Experiment with 32 undergraduate students and second with a User study with 11 students (from the same population). Using a consistent training protocol, we observed similar performance across treatment groups in the Experiment and between the samples of the Experiment and User study. Each set of groups demonstrated comparable proficiency in the initial training modules, establishing a baseline for comparison. Subjects were able to learn EVO in under ten minutes and use the extension to make decisions.

From the Experiment, we concluded that subjects were able to answer goal modeling comprehension questions with EVO faster than without EVO but we did not find a significant difference between the scores of subjects who answered questions with and without EVO. Thus, there was no evidence that EVO has an impact on an individual's understanding of goal models. However, subjects had a positive response to EVO and all preferred the EVO view over the control, with most saying that EVO was faster or easier to use. In the User study, most subjects preferred to analyze goal models and simulation results with EVO (specifically State and Time mode). Subjects who completed the in-person session without EVO recognized its benefits as well. Thus, subjects across both studies had a preference for using color. Finally, our subjects, without prior training in GORE, were able to complete the Experiment instrument without much difficulty. Subjects in the User study were able to extend and personalize a goal model as well as evaluate and draw conclusions from its results, demonstrating the applicability of BloomingLeaf. While there was no difference in an individual's understanding of goal models, the preference for EVO from the Experiment and User studies, as well as positive comments from subjects in both studies, may suggest that the value of EVO lies in an improved user experience as opposed to a quantifiable improvement in understanding. By demonstrating the impacts of EVO, we increase the potential of automated analysis techniques

in Tropos. We share our materials as part of our open-science package[1].

In future work, we continue to develop EVO and BloomingLeaf, by implementing the suggestions provided in Tbl. 12 and Tbl. 15. We can improve the experience for new users by implementing an embedded tutorial into BloomingLeaf and adding a Green-Red (i.e., traffic signal) color palette to EVO. Future work will explore removing the Percent mode, improving the multiple color palettes, and validating the use of EVO when users create their own simulation paths [5].

In future work, we intend to replicate our Experiment study in order to establish external validity (see Sect. 6.1 and Sect. 6.3.4) with subjects in a different context. For example, it would be helpful to repeat these studies with trained modelers and practitioners in industry. Further, since this study was conducted at a women's college, it would be immediately beneficial to replicate these results at a co-educational institution. Replicating our User study with more comprehensive debriefing questions would be beneficial to understanding subjects' choices during the in-person session, as well as in establishing a comparison between studies. It would also be helpful to run a similar study where subjects participate over multiple sessions. Additionally, future work includes conducting case studies of real groups in early-phase RE using EVO. Other work includes investigating the scalability of model analysis with EVO and whether there is an increased benefit to EVO with larger models. For example, we did not observe variations in subjects' score in the Experiment with or without the use of EVO. Perhaps with more challenging questions or models larger than 30 elements, any effect of EVO would become evident. Finally, we would like to explore the interaction between model size and subjects' level of experience with GORE on the effectiveness of EVO, as well as what subjects focus on when modeling and reviewing simulation results (with and without EVO) via an eye-tracking study.

Acknowledgments. We thank our study participants. Thanks to Kaitlyn Cook for assisting in our statistical analysis. This material is based upon work supported by the National Science Foundation under Award No. 2104732.

Data Availability Statement. Datasets generated and/or analyzed during the studies are available at *https://doi.org/10.35482/csc.001.2024*.

References

1. Alwidian, S., Amyot, D.: "Union is Power": Analyzing Families of Goal Models Using Union Models. In: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems (MODELS), pp. 252–262 (2020)
2. Amyot, D., Ghanavati, S., Horkoff, J., Mussbacher, G., Peyton, L., Yu, E.: Evaluating Goal Models Within the Goal-Oriented Requirement Language. International Journal of Intelligent Systems **25**(8), 841–877 (2010)
3. Aprajita: TimedGRL: Specifying Goal Models Over Time. Master's thesis, McGill University (2017)
4. Baatartogtokh, Y., Foster, I., Grubb, A.M.: An Experiment on the Effects of using Color to Visualize Requirements Analysis Tasks. In: Proceedings of the IEEE 31st International Requirements Engineering Conference, pp. 146–156 (2023)
5. Baatartogtokh, Y., Foster, I., Grubb, A.M.: Visualizations for User-supported State Space Exploration of Requirements Models. In: Proceedings of the IEEE 31th International Requirements Engineering Conference, pp. 281–286 (2023)
6. Ben Ayed, C., Halili, S., Tan, Y., Grubb, A.M.: Toward Internationalization and Accessibility of Color-based Model Interpretation. In: Proceedings of the 16th International iStar Workshop (2023)
7. Bi, X., Grubb, A.M.: Incorporating Presence Conditions into Goal Models that Evolve Over Time. In: Proceedings of the 13th International Model-Driven Requirements Engineering Workshop, pp. 272–276 (2023)
8. Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J.: Tropos: An Agent-Oriented Software Development Methodology. Autonomous Agents and Multi-Agent Systems **8**(3), 203–236 (2004)
9. Cebula, N., Diao, L., Grubb, A.M.: A Preliminary Investigation of the Utility of Goal Model Construction. In: Proceedings of the 13th International i* Workshop, pp. 67–72 (2020)
10. Clarke, D., Duimering, P.R.: How Computer Gamers Experience the Game Situation: A Behavioral Study. Computers in Entertainment **4**(3), 6–es (2006)
11. Cohene, T., Easterbrook, S.: Contextual Risk Analysis for Interview Design. In: Proceedings of the 13th IEEE International Conference on Requirements Engineering, pp. 95–104 (2005)
12. Dalpiaz, F., Franch, X., Horkoff, J.: iStar 2.0 Language Guide. arXiv:1605.07767 (2016)
13. Faculty Council: Code of Faculty Governance at Smith College. Tech. rep., Smith College (2023). Available at *https://www.smith.edu/sites/default/files/media/Documents/Provost/Code_of_Faculty_Governance.pdf*, accessed 03/29/2024.
14. Faul, F., Erdfelder, E., Lang, A.G., Buchner, A.: G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. Behavior Research Methods **39**(2), 175–191 (2007)
15. Franch, X.: On the Quantitative Analysis of Agent-oriented Models. In: Proceedings of the International Conference on Advanced Information Systems Engineering (CAiSE), pp. 495–509 (2006)
16. Franch, X., Grau, G., Quer, C.: A Framework for the Definition of Metrics for Actor-Dependency Models. In: Proceedings of the 12th IEEE International Requirements Engineering Conference (RE), pp. 348–349 (2004)
17. Ghazi, P., Glinz, M.: An Experimental Comparison of Two Navigation Techniques for Requirements Modeling Tools. In: Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference (RE), pp. 240–250 (2018). DOI 10.1109/RE.2018.00032

18. Grubb, A.M.: Reflection on Evolutionary Decision Making with Goal Modeling via Empirical Studies. In: Proceedings of RE'18, pp. 376–381 (2018)
19. Grubb, A.M.: Reflection on Evolutionary Decision Making with Goal Modeling via Empirical Studies. In: Proc. of RE'18, pp. 376–381 (2018)
20. Grubb, A.M.: Evolving Intentions: Support for Modeling and Reasoning about Requirements that Change over Time. Ph.D. thesis, University of Toronto (2019)
21. Grubb, A.M., Chechik, M.: BloomingLeaf: A Formal Tool for Requirements Evolution over Time. In: Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference (RE): Poster & Tools Demos, pp. 490–491 (2018)
22. Grubb, A.M., Chechik, M.: Formal Reasoning for Analyzing Goal Models that Evolve over Time. Requirements Engineering **26**(3), 423–457 (2021)
23. 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)
24. Hadar, I., Soffer, P., Kenzi, K.: The Role of Domain Knowledge in Requirements Elicitation via Interviews: An Exploratory Study. Requirements Engineering **19**, 143–159 (2014)
25. Horkoff, J., Li, T., Li, F.L., Salnitri, M., Cardoso, E., Giorgini, P., Mylopoulos, J., Pimentel, J.: Taking Goal Models Downstream: A Systematic Roadmap. In: Proceedings of the 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), pp. 1–12 (2014)
26. Horkoff, J., Maiden, N.A.M., Lockerbie, J.: Creativity and Goal Modeling for Software Requirements Engineering. In: Proceedings of C&C'15, pp. 165–168 (2015)
27. Horkoff, J., Yu, E.: Interactive Goal Model Analysis For Early Requirements Engineering. Req. Eng. **21**(1), 29–61 (2016)
28. Liaskos, S., Jalman, R., Aranda, J.: On Eliciting Contribution Measures in Goal Models. In: 20th IEEE International Requirements Engineering Conference (RE), pp. 221–230 (2012)
29. Lohse, G.L.: A Cognitive Model for Understanding Graphical Perception. Human–Computer Interaction **8**(4), 353–388 (1993)
30. Mathew, G., Menzies, T., Ernst, N., Klein, J.: "SHORT"er Reasoning About Larger Requirements Models. In: Proceedings of the 25th IEEE International Requirements Engineering Conference (RE), pp. 154–163 (2017)
31. Mavin, A., Wilkinson, P., Teufl, S., Femmer, H., Eckhardt, J., Mund, J.: Does Goal-Oriented Requirements Engineering Achieve Its Goal? In: Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference (RE), pp. 174–183 (2017)
32. Meier, L.: ANOVA and Mixed Models: A Short Introduction Using R. CRC Press (2022)
33. Moody, D.: 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
34. Noel, R., Panach, J.I., Pastor, O.: Including Business Strategy in Model-Driven Methods: An Experiment. Requirements Engineering **28**(3), 411–440 (2023)
35. Office of Institutional Research: Common Data Set 2022-2023. Tech. rep., Smith College (2023). Available at *https://www.smith.edu/your-campus/offices-services/institutional-research/data-about-smith*, accessed 03/09/2024.
36. Oliveira, R.F., do Prado Leite, J.C.S.: Using Colorimetric Concepts for the Evaluation of Goal Models. In: Proceedings of the 10th International Model-Driven Requirements Engineering Workshop (MoDRE), pp. 39–48 (2020)
37. Runyon, R.P.: Nonparametric Statistics: A Contemporary Approach. Addison-Wesley (1977)
38. Santos, M., Gralha, C., Goulão, M., Araújo, J., Moreira, A.: On the Impact of Semantic Transparency on Understanding and Reviewing Social Goal Models. In: Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference (RE), pp. 228–239 (2018). DOI 10.1109/RE.2018.00031
39. Schaefer, B.A., Bavelas, J., Bavelas, A.: Using Echo Technique to Construct Student-generated Faculty Evaluation Questionnaires. Teaching of Psychology **7**(2), 83–86 (1980)
40. Shabtai, A., Bercovitch, M., Rokach, L., Gal, Y.K., Elovici, Y., Shmueli, E.: Behavioral Study of Users When Interacting with Active Honeytokens. ACM Transactions on Information and System Security **18**(3), 1–21 (2016)
41. Shull, F., Singer, J., Sjøberg, D.I.: Guide to Advanced Empirical Software Engineering. Springer-Verlag New York, Inc. (2007)
42. Siqueira, F.L.: Comparing the Comprehensibility of Requirements Models: An Experiment Replication. Information and Software Technology **96**, 1–13 (2018)
43. Stol, K.J., Fitzgerald, B.: The ABC of Software Engineering Research. ACM Trans. Softw. Eng. Methodol. **27**(3), 1–51 (2018). DOI 10.1145/3241743
44. Tomczak, M., Tomczak-Łukaszewska, E.: The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size. Trends in Sport Sciences **21**, 19–25 (2014)
45. Varnum, M.H., Spencer, K.M.B., Grubb, A.M.: Towards an Evaluation Visualization with Color. In: Proceedings of the 13th International i* Workshop (iStar), pp. 79–84 (2020)
46. Vegas, S., Apa, C., Juristo, N.: Crossover Designs in Software Engineering Experiments: Benefits and Perils. IEEE Transactions on Software Engineering **42**(2), 120–135 (2016). DOI 10.1109/TSE.2015.2467378
47. Vermeeren, A.P.O.S., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K.: User Experience Evaluation Methods: Current State and Development Needs. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, 10, pp. 521–530. Association for Computing Machinery (2010)
48. Wang, Y.L., Grubb, A.M.: Towards a General Solution for Layout of Visual Goal Models with Actors. In: Proceedings of the IEEE 28th International Requirements Engineering Conference (2020)
49. Winkler, J.P., Grönberg, J., Vogelsang, A.: Optimizing for Recall in Automatic Requirements Classification: An Empirical Study. In: Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference (RE), pp. 40–50 (2019). DOI 10.1109/RE.2019.00016
50. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering. Springer Berlin Heidelberg (2012)
51. Yin, R.K.: Case Study Research: Design and Methods (3rd Edition). Sage (2003)
52. Zaremba, O., Liaskos, S.: Towards a Typology of Questions for Requirements Elicitation Interviews. In: Proceedings of the 29th IEEE International Requirements Engineering Conference (RE), pp. 384–389 (2021)