An Experiment on the Effects of Using Color to Visualize Requirements Analysis Tasks

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

2023 IEEE 31st International Requirements Engineering Conference (RE) | 979-8-3503-2689-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/RE57278.2023.00023

***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 scientific evaluation paper, we explore the effectiveness and usability of EVO. We conduct an experiment ($n = 32$) to measure any effect of using colors to represent evidence pairs. We find that with minimal training, untrained modelers were able to use the color visualization for decision making. The visualization significantly improves the speed of model comprehension and users found it helpful.**

I. 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 [1], [2] and with families of models [3], where users evaluate collections of models. Given the potential for automating analysis of goal models [4] and connecting them with downstream activities [5], the broader aim of this research program is to improve the *cognitive effectiveness* [6] 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 [7]. 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 [8]. 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.

In prior work, Grubb and Chechik developed automated analysis techniques for Tropos models with evolutionary information [9]. 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) [10]. Varnum et al. completed a preliminary evaluation with an example but did not validate this approach with users [10]. Prior work suggests that color can help individuals interpret certain graph types faster [11], but should be used as a secondary encoding [6].

Contributions. We 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. We report on an IRB-approved between-subjects experiment conducted with 32 undergraduate students. We aim to answer four research questions:

- RQ0 Do modelers across treatment groups perform similarly on basic goal modeling and simulation tasks?
- 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/instrument?

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.

Organization. The remainder of the paper is organized as follows. Sect. II reviews goal modeling and the EVO approach. Sect. III describes our study methodology. We report on the results of our study in Sect. IV, and discuss lessons learned and validity in Sect. V. Finally, we review related work in Sect. VI and conclude in Sect. VII.

2332-6441/23/$31.00 ©2023 IEEE
DOI 10.1109/RE57278.2023.00023

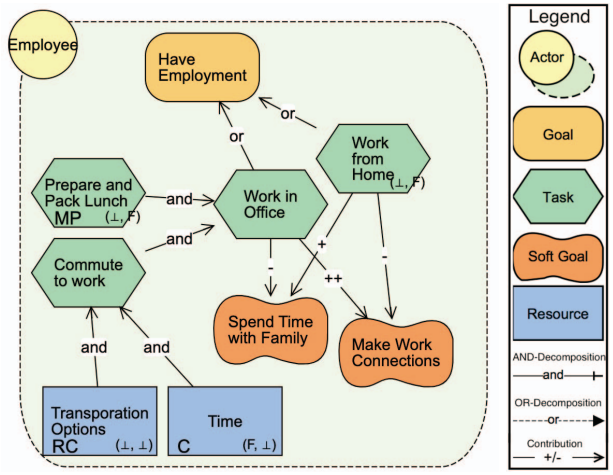

Fig. 1: Employment Model & Goal Modeling Legend

II. Background

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

A. 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, 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.

Fig. 2: Evidence pairs overlayed with EVO color assignments.

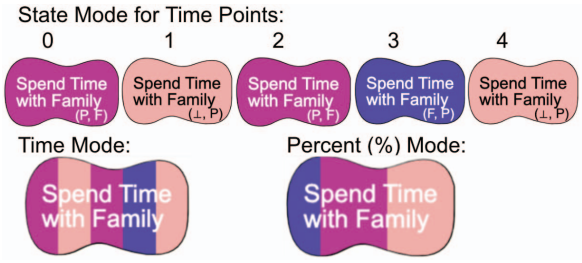

Fig. 3: EVO modes showing only Spend Time with Family.

B. Simulating Models over Time

We use the Evolving Intentions framework [9] 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. Intentions that are not assigned evolving functions receive their valuations via propagation. Thus, 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.

C. EVO: Evaluation Visualization Overlay

As briefly mentioned in Sect. I, Varnum et al. introduced the Evaluation Visualization Overlay (EVO) [10]. 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. 4 gives the initial state of the Summer model (see Sect. III-B for details). In Fig. 4, Have Summer Activity is colored dark red because it has been assigned the $(\perp, F)$ 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.

III. Methodology

In this section, we describe our methodology for conducting this study, which was approved by our institutional review board (IRB). Our supplemental materials are available online[1].

A. Experiment Design

Our primary objective in designing this experiment was to measure the effects of EVO. The original EVO proposal was implemented as an extension to BloomingLeaf [12]. We did not intend to evaluate the usability of 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 our institution's browser-based Qualtrics® XM platform. We used the BloomingLeaf git repository [12] only for the purpose of creating our study materials and models.

In designing this experiment, our main consideration was ensuring that we measured the appropriate elements, and controlled for the risks of variability between subjects' tasks, subjects' natural performance, and any learning, fatigue, or carryover effects (see Sect. V-C). We chose a nested 2x2 design [13], with random treatment group assignment. To measure the impacts of using EVO, we compared measurements of subjects analyzing a model with and without having access to EVO, using two different models. To mitigate any learning effects, we varied the EVO training order. We took measurements of subjects' correctness when answering questions, labeled as score, and how long it took subjects to answer these questions. Thus, our dependent variables were *score* and *time*. Previous investigations have demonstrated that task equivalency is an important factor in analyzing model comprehensibility [8]. We designed our questions to be similar but not identical. To understand any effects that may result from model variation, we test two models in our design.

We explored conducting the study as either a between- or within-subjects comparison. Ideally, our study would be analyzed in-subjects. This would control for natural variations in individual performance, model variability, and EVO ordering. Yet, analyzing this design requires the use of ANOVA, for which we were unsure we could get sufficient subjects. Instead, we planned our analysis to be performed between-subjects, but this has the downside of not being able to control for individual subject variability.

TABLE I: Study Models

Models	Figure	Actors	Intentions	Links	Evolving Functions
Course	n/a	2	9	10	2
Employment	Fig. 1	1	9	10	3
Summer	Fig. 4	1	14	17	8
Bike	Fig. 5	1	16	20	7

B. Materials: Models and Videos

In this study, we used four models: the Employment model (see Fig. 1), the Summer model (see Fig. 4), the Bike model (Fig. 5), and the Course model (not shown for space considerations, see online[1]). We list these models and their associated metrics in Tbl. I. The Course model 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. II-A, 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. 4), 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. 5, the City actor wants to construct bike lanes, 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.

[1]See *https://doi.org/10.35482/csc.002.2023* for supplement.

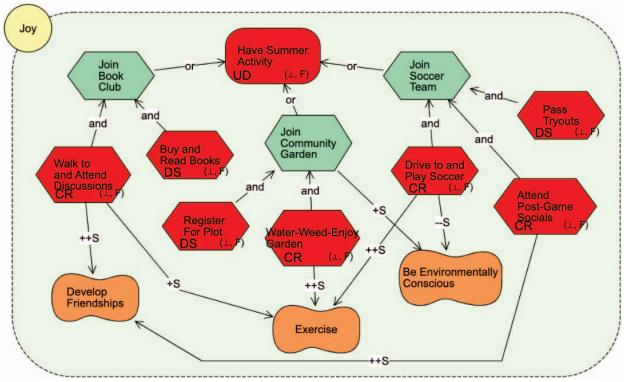

Fig. 4: Summer Model

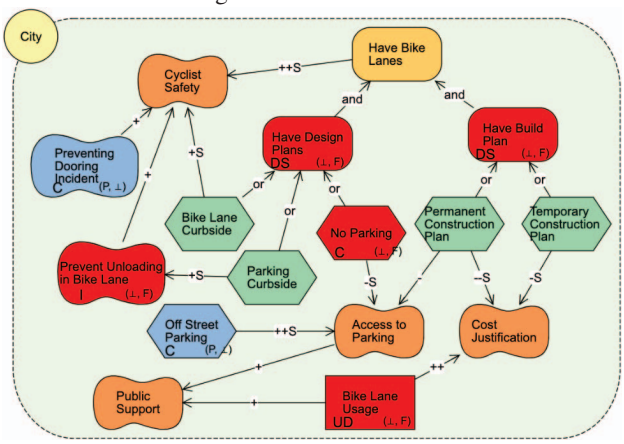

Fig. 5: Bike Model

Subjects were tested on their ability to answer questions about the Bike and Summer models (see Tbl. IV 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 consisted of three 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*.

C. Procedure: Conducting the Experiment

Tbl. II lists the steps in our protocol for each treatment group. Parts 0, 1, and 5 are common across all subjects. 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. III-D). In Part 1, subjects completed two training modules, one introducing goal modeling more generally using VidGM, and the other introducing the minimal required subset

TABLE II: Study Protocol

	Treatment Groups			
	EVO: Bike		EVO: Summer	
Part	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
3	Bike EVO	Training: EVO	Summer EVO	Training: EVO
4	Summer Control	Bike EVO	Bike Control	Summer EVO
5	Debrief			

TABLE III: 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

of the Evolving Intentions framework (using VidSim). We used the Course and Employment models in Part 1 and in the 'Training: EVO' module in Parts 2 and 3 (see Tbl. II). 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. These questions allowed us to establish a baseline for comparison of subjects' performance on goal model tasks. In Part 5, we debriefed and remunerated subjects, having them reflect on the study.

Parts 2–4 (see Tbl. II) varied based on the subjects' randomly assigned treatment group. All subjects completed the 'Training: EVO' module and answered questions about the Bike and Summer models (see Tbl. IV) after examining each model. What varied is which model (i.e., Bike or Summer) they answered questions about using EVO and whether they answered questions about a model before or after completing the EVO training. This allowed us to control for both variations in the models and a learning effect.

D. Experimental Conditions and Subject Information

We conducted the experiment in early 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, one author was on hand to answer any questions after each training module.

We recruited 32 undergraduate students to participate, eight per treatment group. All subjects achieved a perfect score on

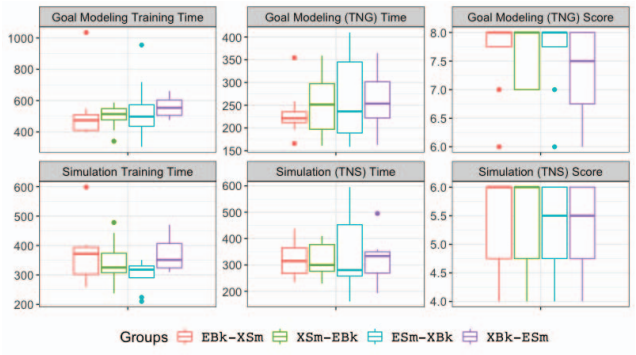

Fig. 6: Scores (counts) and timing data (in seconds) for the goal modeling and simulation training. Maximum TNG score was 8, while maximum TNS score was 6.

the color vision test. During Part 0 of our protocol (see Tbl. II), 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. III 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. The median scores for RE and iStar were low but non-zero. 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. We did not expect subjects to have any familiarity with Tropos or GRL but included them for completeness. Subjects were randomly assigned to treatment groups before demographic information was collected, so we were unable to use this information in group assignments.

IV. Results

In this section, we answer our research questions using data collected in our investigation.

A. RQ0: Establishing a Baseline for Comparison

We begin by answering RQ0: Do modelers across treatment groups perform similarly on basic goal modeling and simulation tasks? All data collected during Part 1 of our protocol (see Tbl. II) 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 (TNG), and then they watched VidSim and answered six questions (plus one qualitative question) about simulating models over time (TNS), see supplement[1] for questions. All answers were scored as correct or incorrect. Fig. 6 reports box plots for subjects' training time, test time, and test scores (from left to right), for both the goal modeling and simulation training. Each box plot is sorted by treatment group and times are reported in seconds. For the goal model training (see first row in Fig. 6), most subjects

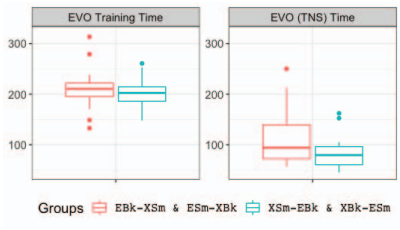

Fig. 7: Timing data (in seconds) for the EVO training.

TABLE V:
EVO training score frequencies, grouped by order (i.e., Part 2 or Part 3, see Tbl. II).

EVO Train.	Score Freq.		
	0-4	5	6
Part 2	0	4	12
Part 3	0	3	13

spent 8–9.5 minutes on the initial training (i.e., rounded first to third quantile), which included a 7.5-minute video), most subjects took 3–5 minutes to answer the TNG questions, achieving scores between 6–8. For the simulation training (see second row), subjects completed the initial training (including a 5-minute video) in 5–6.5 minutes. They then answered the TNS questions in 5–6.5 minutes, achieving scores between 4–6. From the box plots, we cannot observe any meaningful difference between treatment groups. For completeness, we used the *Kruskal-Wallis Rank Sum* (KWRS) test [14] to test for any variability between treatment groups. Our null hypothesis was that the treatment groups performed equally well on the questions, both in terms of score and time. We failed to reject our null hypothesis ($p \not< 0.1$), 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 subjects performed similarly on basic goal modeling and simulation tasks.

B. RQ1: Subjects' Use of EVO

Second, we consider RQ1: To what extent are subjects able to learn EVO, and then use EVO to answer goal modeling questions? Given our RQ0 results, we investigate this question between-subjects using a nested 2x2 design. In Parts 2–4 (see Tbl. II), each subject completed the EVO Training module and answered questions about the Bike and Summer models (see Tbl. IV), 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)

TABLE IV: Summer and Bike Questions

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.

Is our training 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 matched treatment groups `EBk-XSm` & `ESm-XBk` (i.e., EVO training in Part 2, see Tbl. II) 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. V lists the score data for the EVO training. All subjects achieved a score of 5 or 6 (out of a possible 6), and the groups are not distinguishable. Fig. 7 shows the box plots for the training and test times for the EVO Module. Subjects took between two and five and a half minutes to review the training materials and between one and four and a half minutes for the EVO questions. Our null hypothesis is that there is no significant variation between groups. We fail to reject this hypothesis (KWES, $p \not< 0.1$), unable to detect variations between groups.

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. IV. 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. VI 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. 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. IV), 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. The phrasing of Q6 can be improved (see Sect. V-A for a discussion). Given the score data in Tbl. VI, we did not expect to find variations between groups (i.e., our null hypothesis) and, in fact, did not find any statistical difference between treatment groups (i.e., KWES, $p \not< 0.1$) with respect to the subjects' scores for Bike and Summer model questions.

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

C. 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?

TABLE VI: Median scores (out of fourteen) for Bike and Summer questions. Bold indicates subject group used EVO.

Group	Bike Median	Summer Median
EBk–XSm	**13**	12.5
XSm–EBk	**13.5**	13
ESm–XBk	12	**12**
XBk–ESm	13	**11.5**

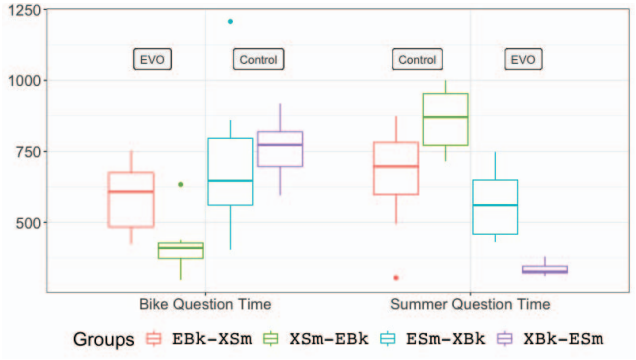

Fig. 8: Timing Data (in seconds) for answering Bike and Summer questions (see Tbl. IV).

(a) Bike and Summer Times. To measure subject completion times, we added their times from Pages 1, 2, 3, and 5 (see Tbl. IV). 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. 8 gives the box plot for each treatment group for the Bike and Summer model question times. In the Bike model (left side), `EBk–XSm` (red) and `XSm–EBk` (green) used EVO to answer the questions and visibly lower time. Again, our null hypothesis is that there is no difference between treatment groups. Using the KWES test, we find the times for the Bike model to be significantly faster ($p < 0.01$). In the Summer model (right side), `ESm–XBk` (blue) and `XBk–ESm` (purple) used EVO to answer the questions and also have visibly lower time. Again using the KWES test, we find the times for the Summer model to be significantly faster ($p < 0.001$).

Upon further inspection of Fig. 8, we observe a possible learning effect—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). Yet, when we conduct a pair-wise comparison based on treatment group order and EVO, we do not find a significant difference with respect to order but we do find one with respect to using EVO; thus, we hypothesize that the interaction of subjects being in the control group and using EVO may contribute to this additional benefit. Therefore, we found a significant effect between the treatment groups with respect to the time required to answer the Bike and Summer questions.

(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. IV-D.

TABLE VII: Average (mean) subjects' rating of their difficulty with three study aspects (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

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

D. 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. II (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. VII 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. III, discussed in Sect. III-D). 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. VII 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. VIII 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 red-blue color vision deficiencies (see Sect. VII for details).

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 [15], [16], [17], [18]. 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. VIII) 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.

V. Discussion

Next, we describe our lessons learned, compare the bike and summer model, and discuss the validity of our experiment.

A. Lessons Learned and Implications for Research

Subject Background and Recruitment. We developed this 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. The purpose of the pilot was 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. IV. We discovered these values varied dramatically based on how much text subjects entered in the free form questions. As listed in Tbl. IV, 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 $20 USD. We launched three separate iterations of the study. Our first 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. III-D. 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. We had sufficient volunteers for our in-person version and felt this was an important lesson learned.

TABLE VIII: Recommendations for Improvement

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.

Improvements to the Study Instrument. We reviewed the questions and supplemental information from the study by Hadar et al. [7] 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 Q6 (for both the Bike and Summer models, see Tbl. IV), 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. IV) 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 a future iteration of this

instrument, we would change the selected intention for the Bike model and remove the *none of the above* option.

In our analysis, 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.

Statistical Methods. Given our per group sample size, any statistical test will have lower power to make conclusions (see Sect. V-C and online[1]). In Sect. IV, we used the KWRS test to evaluate if there are distinct groupings within our sample data [14]. 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.

B. Comparing Bike and Summer Models

As introduced in Sect. III-A, we explored our research questions between-subjects. In Sect. IV, we found a statistically significant difference between using EVO and the control in the time it took subjects to answer questions about both the Bike and Summer model. Yet, in this test, we cannot directly compare the times associated with the Bike and Summer model or control for individual subject variability.

We briefly explore variations of the time it took subjects to answer the test questions (i.e., our *dependent variable*). We compare test times given three factors (*independent variables*): (i) whether the subject used EVO, (ii) whether it was the first or second measurement for that subject, and (iii) whether the measurement was of the Bike or Summer model. In order to identify which factors are significant, we compared within subjects by fitting multiple linear mixed-effects models and then conducted a model comparison with repeated measures data using a likelihood ratio test (i.e., ANOVA) [19]. We used a linear mixed-effects model to account for non-independence (i.e., there were two measurements for each subject).

Comparing the full model to one with interactions between factors showed that the interaction terms in the model are not significant ($p > 0.05$). We found the EVO factor to be significant ($p < 0.001$), meaning that within-subjects there was a difference in the time it took subjects to answer questions with EVO as opposed to without EVO. The order of whether subjects were given the control or the treatment first was significant ($p < 0.001$), implying that there was a learning effect over time. Which model was measured was not significant ($p > 0.05$), meaning that there is no significant difference in the times for the Summer and Bike models.

Since there is no significant difference between models and no interaction effect, we can analyze this as a two-way ANOVA where using EVO and order of EVO presentation are the two factors. Using a statistical power test for repeated measures ANOVA within-subjects with a medium effect size, we found that the minimum sample size using G*Power [20] for our experiment was 56. Thus, we have low statistical power. We did not find any difference between the Bike and Summer models and found the presence of a learning effect within subjects.

C. Threats to Validity

We discuss threats to validity using the categories in [13].

Conclusion Validity. Our main threat in this experiment is low sample size. Having 32 subjects spanning four treatment groups is considered a low sample size. Thus, we chose to conduct our main analysis between-subjects to mitigate this threat. We may have experienced a reliability of measures threat, as subjects asked questions about the wording of Q6 (see Sect. V-A). 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. III-B), and maintained our laboratory setup throughout the study period, to ensure a consistent in-person experience. 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. III). 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.

Internal Validity. 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 in our 2x2 design; yet, the Bike model questions may have been slightly harder (see Sect. IV-B). With this design, there is still a risk of carryover effects [21]. Our voluntary study with cash remuneration may have experienced a selection effect. To our knowledge, no subjects used BloomingLeaf or EVO prior to the study.

Construct Validity. We conducted multiple pilot mini-studies (not discussed in this paper) 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. IV and isolated qualitative questions. 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. As always, we have threats of *hypothesis guessing* and *evaluation apprehension*. Some subjects expressed nervousness asking if they needed to review data structures or read about goal modeling before participating. Some students who took a software engineering course may have scored better overall; yet, our common training protocol may have limited this threat.

External Validity. Our setting was not reflective of the use of EVO in the "real world". We conducted the experiment one-on-one in our lab using a survey, instead of embedding EVO within a goal modeling tool (e.g., BloomingLeaf). 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. III), these results may, in fact, generalize. Additional experiments with different populations, problem domains, and larger models for scalability are required.

VI. Related Work

Recent work has critiqued the adaptability of GORE approaches [22]. In this paper, we address this gap by improving the interpretability of Tropos evidence pairs. As already introduced in Sect. I, Hadar et al. [7] and Siqueira [8] 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 [17], while TimedGRL used color in heat maps to visualize evolving GRL models [1]. Varnum et al. proposed using colors to help stakeholders interpret the evidence pairs used in Tropos for intention evaluations [10]. 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 [23]. 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.

We built on the methodology of similar studies in RE for our between-subjects experiment and followed the guidance in [13] and [24]. Winkler et al. reported on a between-subjects 2x2 design similar to ours with sixteen subjects [25]. 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. Ghazi et al. reported a study comparing two navigation techniques for requirements modeling tools [26]. 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 [27]. 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.

VII. Conclusions and Future Work

In this paper, we explored how using EVO to visualize evidence pairs impacts an individual's ability to reason with goal models that evolve over time. To do so, we conducted an IRB-approved between-subjects experiment with 32 undergraduate students. We found that when given a consistent training protocol for goal modeling and simulation, each treatment group performed equally well on the initial training modules, establishing a baseline for comparison between treatment groups. Subjects were able to learn EVO in under ten minutes and use the extension to make decisions. From this 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. Finally, our subjects, without prior training in GORE, were able to complete the instrument without much difficulty. 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].

Given the empirical evidence of the effectiveness of EVO presented in this paper, we encourage the original authors to continue their development of EVO within BloomingLeaf. Additionally, as mentioned in Sect. VI, the selected colors of blue, red, and purple should be validated. Our subjects proposed several alternatives for conflicting colors in Tbl. VIII. We are investigating these alternative color palettes, as well as palettes for colorblind users. In future work, we intend to replicate our study in order to establish external validity (see Sect. V-C), both with subjects in a different context and using EVO embedded within BloomingLeaf and other goal modeling tools. Additionally, future work includes conducting case studies of real groups in early-phase RE using EVO. Other work included extending and validating the EVO feature with other types of analysis.

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.

REFERENCES

[1] Aprajita, "TimedGRL: Specifying Goal Models Over Time," Master's thesis, McGill University, 2017.

[2] A. M. Grubb, "Evolving Intentions: Support for Modeling and Reasoning about Requirements that Change over Time," Ph.D. dissertation, University of Toronto, 2019.

[3] S. Alwidian and D. Amyot, ""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)*, 2020, pp. 252–262.

[4] G. Mathew, T. Menzies, N. Ernst, and J. Klein, "Shorter Reasoning About Larger Requirements Models," in *Proceedings of the 25th IEEE International Requirements Engineering Conference (RE)*, 2017.

[5] J. Horkoff, T. Li, F.-L. Li, M. Salnitri, E. Cardoso, P. Giorgini, J. Mylopoulos, and J. Pimentel, "Taking Goal Models Downstream: A Systematic Roadmap," in *Proceedings of the 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 2014, pp. 1–12.

[6] D. Moody, "The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering," *IEEE Transactions on Software Engineering*, vol. 35, no. 6, pp. 756–779, 2009.

[7] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi, "Comparing the Comprehensibility of Requirements Models Expressed in Use Case and Tropos: Results from a Family of Experiments," *Information and Software Technology*, vol. 55, no. 10, pp. 1823–1843, 2013.

[8] F. L. Siqueira, "Comparing the comprehensibility of requirements models: An experiment replication," *Information and Software Technology*, vol. 96, pp. 1–13, 2018.

[9] A. M. Grubb and M. Chechik, "Formal Reasoning for Analyzing Goal Models that Evolve over Time," *Requirements Engineering*, vol. 26, no. 3, pp. 423–457, 2021.

[10] M. H. Varnum, K. M. B. Spencer, and A. M. Grubb, "Towards an Evaluation Visualization with Color," in *Proceedings of the 13th International i* Workshop (iStar)*, 2020, pp. 79–84.

[11] G. L. Lohse, "A Cognitive Model for Understanding Graphical Perception," *Human–Computer Interaction*, vol. 8, no. 4, pp. 353–388, 1993.

[12] A. M. Grubb and M. Chechik, "BloomingLeaf: A Formal Tool for Requirements Evolution over Time," in *Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference (RE): Poster & Tools Demos*, 2018, pp. 490–491.

[13] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer Berlin Heidelberg, 2012.

[14] R. P. Runyon, *Nonparametric Statistics: A Contemporary Approach*. Addison-Wesley, 1977.

[15] X. Franch, G. Grau, and C. Quer, "A Framework for the Definition of Metrics for Actor-Dependency Models," in *Proceedings of the 12th IEEE International Requirements Engineering Conference (RE)*, 2004, pp. 348–349.

[16] X. Franch, "On the Quantitative Analysis of Agent-oriented Models," in *Proceedings of the International Conference on Advanced Information Systems Engineering (CAiSE)*, 2006, pp. 495–509.

[17] D. Amyot, S. Ghanavati, J. Horkoff, G. Mussbacher, L. Peyton, and E. Yu, "Evaluating Goal Models Within the Goal-Oriented Requirement Language," *International Journal of Intelligent Systems*, vol. 25, no. 8, pp. 841–877, 2010.

[18] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos, "Tropos: An Agent-Oriented Software Development Methodology," *Autonomous Agents and Multi-Agent Systems*, vol. 8, no. 3, pp. 203–236, May 2004.

[19] L. Meier, *ANOVA and Mixed Models: A Short Introduction Using R*. CRC Press, 2022.

[20] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences," *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, 2007.

[21] S. Vegas, C. Apa, and N. Juristo, "Crossover Designs in Software Engineering Experiments: Benefits and Perils," *IEEE Transactions on Software Engineering*, vol. 42, no. 2, pp. 120–135, 2016.

[22] A. Mavin, P. Wilkinson, S. Teufl, H. Femmer, J. Eckhardt, and J. Mund, "Does Goal-Oriented Requirements Engineering Achieve Its Goal?" in *Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017, pp. 174–183.

[23] R. F. Oliveira and J. C. S. do Prado Leite, "Using Colorimetric Concepts for the Evaluation of Goal Models," in *Proceedings of the 10th International Model-Driven Requirements Engineering Workshop (MoDRE)*, 2020, pp. 39–48.

[24] F. Shull, J. Singer, and D. I. Sjøberg, *Guide to Advanced Empirical Software Engineering*. Springer-Verlag New York, Inc., 2007.

[25] J. P. Winkler, J. Grönberg, and A. Vogelsang, "Optimizing for Recall in Automatic Requirements Classification: An Empirical Study," in *Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019, pp. 40–50.

[26] P. Ghazi and M. Glinz, "An Experimental Comparison of Two Navigation Techniques for Requirements Modeling Tools," in *Proceedings of the 2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018, pp. 240–250.

[27] M. Santos, C. Gralha, M. Goulão, J. Araújo, and A. Moreira, "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)*, 2018, pp. 228–239.