2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)

Exploring the Robustness of the Effect of EVO on Intention Valuation through Replication

Yesugen Baatartogtokh
University of Massachusetts
Amherst, MA USA
yesugen@umass.edu

Kaitlyn Cook
Smith College
Northampton, MA USA
kcook93@smith.edu

Alicia M. Grubb
Smith College
Northampton, MA USA
amgrubb@smith.edu

***Abstract*—The development of high-quality software depends on precise and comprehensive requirements that meet the objectives of stakeholders. Goal modeling techniques have been developed to fill this gap by capturing and analyzing stakeholders' needs and allowing them to make trade-off decisions; yet, goal modeling analysis is often difficult for stakeholders to interpret. Recent work found that when subjects are given minimal training on goal modeling and access to a color visualization, called EVO, they are able to use EVO to make goal modeling decisions faster without compromising quality. In this paper, we evaluate the robustness of the empirical evidence for EVO and question the underlying color choices made by the initial designers of EVO. We conduct a pseudo-exact replication ($n = 60$) of the original EVO study, varying the experimental site and the study population. Even in our heterogeneous sample with less a priori familiarity with requirements and goal modeling, we find that individuals using EVO answered the goal-modeling questions significantly faster than those using the control, expanding the external validity of the original results. However, we find some evidence that the chosen color scheme is not intuitive and make recommendations for the goal modeling community.**

***Index Terms*—Requirements, Goal Modeling, Replication**

I. Introduction

Requirements engineering (RE) is important for the development of software [1]. RE methods allow developers to identify, document, and manage the specifications of software systems under consideration, especially in the early stages of a project. For example, goal modeling approaches to RE can help stakeholders reason about trade-offs in the design of software systems and understand the dependencies of such systems [2], [3]. Yet, these approaches have not seen expansive adoption among practitioners, due to the complex nature of building models and understanding their analysis [4].

One approach to assisting users in understanding goal models has been the introduction of color to communicate whether elements in a goal model are satisfied or not, and most goal modeling tools (e.g., [5]–[9]) have implemented this feature. Other tools have used colors to identify aspects of the model (e.g., root/leaf nodes [10], legal requirements [11]). When researchers validate these approaches they analyze them in the context of the larger tool and do not isolate the coloration as a factor. We consider two approaches to color valuations in goal models. A green-red color scheme is quite common, where green and red denote the colors of a traffic light [12] and where blue denotes conflict [7]. A blue-red color scheme has also been used, where blue is satisfied, red is denied, and purple (i.e., blue + red) is conflict [9]. In the blue-red scheme, blue is used to assist color deficient users; yet, we hypothesize this choice makes the scheme less intuitive for a general audience.

While there is substantial research on the use of color in data visualizations [13], there is little empirical evidence to validate the use of color in goal model analysis. Oliveira and Leite provide a theoretical formulation for the use of the green-red palette based on RGB-colormetrics [14]. Baatartogtokh et al. isolated the blue-red palette in an experiment to evaluate whether the addition of colors affected users' ability to answer goal model questions [15]. They found that the coloration feature (called EVO) significantly improved the speed of subjects' decision making, without affecting quality. This state-of-the-art feature shows promise; however, the empirical findings supporting it are limited by the *small sample size* of the original study and may be compromised by *hypothesis guessing* due to the prominence of the researchers at the study site. Also, the results may not extend externally because of the homogeneous characteristics of the sample. Thus, **our aim is to investigate the robustness of the original study findings and to strengthen the body of evidence about EVO**.

Contributions. In this paper, we contribute a *pseudo-exact dependent* replication [16] of the experiment by Baatartogtokh et al. [15]. We change the experimental site and the study population with the goal of exposing and understanding sources of variability that influence the results [17]. We improve on the original study by increasing the size of our sample ($n = 60$), increasing statistical power to detect small- or intermediate-sized EVO effects. Additionally, we collect subjects' color preferences and demographic information to understand whether the EVO color palette is intuitive to users. Finally, we conduct a meta-analysis of Baatartogtokh et al. [15] and the replication study in order to obtain a more precise and generalizable estimate of the effect of EVO.

Research Questions. We replicate the research questions of Baatartogtokh et al. [15] (called the "Smith" experiment):

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?

1558-1225/25/$31.00 ©2025 IEEE
DOI 10.1109/ICSE55347.2025.00143

RQ3 How do subjects rate the study experience and instrument?

We add two research questions to explore variations between studies and evaluate the original color choice:

RQ4 To what extent do subjects associate blue and red with good and bad outcomes, respectively?

RQ5 How do subjects compare across studies? What are the sources of variability, if any, that influence the results?

We confirm the results of Baatartogtokh et al. [15]. Subjects answered questions significantly faster with EVO than without, without any evidence of impact on correctness. We found no statistically significant differences in the impact of EVO on completion speeds or answer correctness between the original experiment and our replication; thus, we contribute pooled, more precise, estimates of EVO's impact on goal modeling. In terms of color choice in EVO, we found that most subjects associate green with good outcomes and red with bad outcomes, while only a few associate blue with good. We argue that the chosen EVO palette is not intuitive for users. These results impact future RE research and the development of goal modeling tools aimed at software practitioners.

Organization. In what follows, we describe our replication design in Sect. III and results in Sect. IV. We explore implications in Sect. V and future directions in Sect. VII. Additionally, we give a brief overview of goal modeling terminology in Sect. II and discuss related literature in Sect. VI.

II. Background

Goal Models. A goal model represents a scenario as a directed graph, where the nodes are stakeholders' 'intentions' and the links represent the relationships between intentions. Intentions are assigned to an 'actor' in the scenario, who is an entity involved in the activity. Fig. 1 shows a goal model fragment in Tropos [18], which was adapted from the original study Bike model [19]. The model displays the City actor and its intentions. *Intentions* consist of four types: goals, tasks, soft goals, and resources. These intentions and their default background color are shown in the legend in Fig. 1. For example, goals are yellow ovals and tasks are green hexagons. Note that the default color of each intention is not the construct under investigation in this study; instead, the default color representation is treated as the control.

Model links consist of two general types: decomposition and contribution. Decomposition links decompose an intention into sub-intentions. In an and decomposition all of the sub-intentions need to be satisfied in order for the parent intention to be satisfied. In Fig. 1, the City's root goal is Have Bike Lanes, which is and-decomposed: both Have Design Plans and Have Build Plans must be satisfied in order for Have Bike Lanes to be satisfied. Only one fulfilled sub-intention is needed to satisfy the parent intention of an or link; thus, only one of No Parking or Bike Lanes Curbside must be satisfied to make Have Design Plans satisfied. By or decomposing goals, stakeholders can evaluate scenarios and make trade-off decisions. Contribution links indicate that a source intention influences the destination intention of the link. Contribution links are denoted as ++, +, −−, or −, where the number of symbols indicates the strength of the influence. For example, the − link towards Cost Justification in Fig. 1 will propagate partial negative evidence, whereas the −− link to the same node will propagate full negative evidence.

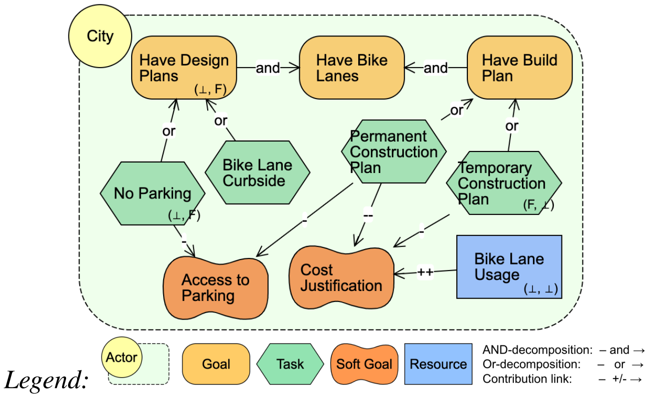


Fig. 1: Fragment of Bike model & goal model legend.

Intention Valuations. Goal models can be assigned evidence pairs which represent the level of evidence for the fulfillment of a given intention. Evidence pairs are represented by (s, d), where s represents the level of evidence for the satisfaction or fulfillment of an intention and d represents the level of evidence against the satisfaction or fulfillment of an intention. Each of s and d can be assigned one of three values: F denoting full evidence, P denoting partial evidence, or $\perp$ denoting no evidence. These combinations of evidence pairs form the five named 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 (F, F), (F, P), (P, F), (P, P). These values can be assigned by the user in the model. In Fig. 1, the user has assigned the task Temporary Construction Plan the value *Satisfied* $(F, \perp)$ because the actor City has already completed the task. During analysis, the initial valuations are propagated throughout the model. In this study, we investigate how the use of color affects subjects' ability to review and interpret the evidence pair assignments of a model.

EVO. As introduced in Sect. I, EVO (Evaluation Visualization Overlay) assigns colors to each of the evidence pairs used for intention valuations [9]. The color blue means that an intention is more satisfied, while red means that the intention is more denied. Thus, as shown in Fig. 2, *Satisfied* is colored blue, *Partially Satisfied* is colored light blue, *Partially Denied* is colored light red, and *Denied* is colored red. *None* is assigned grey. The purpose of assigning colors through EVO is to make evaluating a goal model simpler. Thus, the base colors of intentions are overlayed with colors that represent their valuations. Intentions without assignments retain their base colors. Fig. 3 shows the intentions from the model in Fig. 1 that have assigned evidence pairs, where each intention is sliced showing it with and without EVO. Two additional EVO modes can be used to review a time-based simulation. The Time mode merges all time points into one view, shown as colored stripes over an intention in the order that they

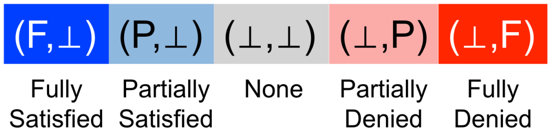


Fig. 2: Legend of evidence pairs with EVO.

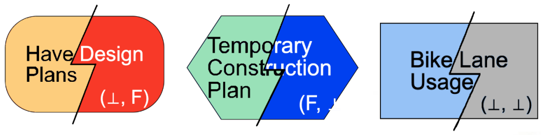


Fig. 3: Fig. 1 Intentions without (left) and with (right) EVO.

occur. The Percent mode also uses colored stripes, where each stripe corresponds to the percentage of the simulation that the intention is assign each evidence pair value.

III. Methodology

Since we replicate the work of Baatartogtokh et al. [15], we do not repeat the original study design in this paper but give a brief overview of it and relevant differences. This study is in compliance with the ACM Publications Policy on Research Involving Human Participants and Subjects and was approved by the Research Ethics Board at the University of Toronto (Protocol ID: 00045411) and the Institutional Review Board at Smith College (Protocol ID: 20-026). Our supplemental materials are available online[1].

Study Rationale. First, we explain our study rationale and trade-offs [20]. While a theoretic replication (see Sect. VI) could have provided more insight into the use of the technique, we argue that the risks introduced in Sect. I need to be mitigated first. The original experiment was conducted at a women's liberal arts college with an undergraduate population of 2,500 [21] that is known for its tight-knit community. Subjects may have gained insight about the research through prior presentations and posters at the college, leading to hypothesis guessing. Additionally, students are instructed in goal modeling at the undergraduate level. As such, it is important to replicate the study at an independent co-educational institution with a larger, potentially more heterogeneous, student body. We thus chose to maintain the original study design and replicate it in a different population. We were able to reuse the study materials and deliver the same training [19] since the exact materials could be used and we do not have the possibility of instructor bias. There were some differences in the study instrument, which we discuss later in this section.

Study Context. Our study was conducted in early 2024 at the University of Toronto, St. George Campus. Subjects were required to be proficient in English and not have a *known* color vision deficiency (i.e., colorblindness). We updated the inclusion criteria for our context, requiring that subjects be computer science (CS) majors or specialists and have completed a second-semester CS course with training on programming and data structures (i.e., CSC111H1 or CSC148H1), which is the same level of study required for the original study at Smith College. We recruited subjects through the CS undergraduate mailing list and targeted course specific mailing lists for eligible classes. We also posted flyers in the building where the software engineering (SE) research lab is located (see supplement[1] for recruitment materials). Subjects completed the instrument in a meeting room in the SE research lab, with one researcher available to answer questions. The researcher sat facing away from the subject's screen, to reduce subject apprehension. All subjects who completed the survey received a $25 CAD Amazon.ca gift card, which was distributed via email in weekly waves.

TABLE I: Study protocol.

	Trial Arms (Sequences)			
Period	`EBk-XSm`	`XSm-EBk`	`ESm-XBk`	`XBk-ESm`
0	Consent, Inclusion Criteria, and Goal Modeling Training			
1	*Training: EVO* Bike EVO	Summer Control	*Training: EVO* Summer EVO	Bike Control
2	Summer Control	*Training: EVO* Bike EVO	Bike Control	*Training: EVO* Summer EVO
3	Collect Demographic Information and Debrief			

Experimental Procedure. The protocol for the in-person lab session is shown in Tbl. I. Subjects were automatically randomly assigned to a trial arm (i.e., sequence). We first verified the subjects' study eligibility, prior knowledge, and color preferences before having the subjects complete training in goal modeling. In Periods 1 and 2, each subject answered the twelve questions listed in Tbl. II relating to either the "Summer" or "Bike" models (see [19] for models). In this crossover design, we varied both the experimental objects (i.e., Bike or Summer model) and whether EVO had been applied to the model or not (i.e., control). Immediately prior to using the EVO treatment, subjects were trained in EVO. For example, `EBk-XSm` (see Tbl. I) was trained in EVO and used EVO to answer the Bike questions in Period 1, then answered the Summer questions in Period 2 without access to EVO. `XSm-EBk` swapped the order of `EBk-XSm` but used the same experimental objects for each treatment. `ESm-XBk` learned EVO in Period 1 but answered Summer questions with EVO, while `XBk-ESm` completed the same in Period 2. All trial arms concluded the session with a short debriefing.

Differences in Study Instrument. We made the improvements recommended by Baatartogtokh et al. [15]. We allowed subjects to select black or dark purple as the color option to represent full conflict (F, F) and updated Q4 and Q8 in Tbl. II. We also asked subjects to briefly describe the context in which they became familiar with any Goal Modeling Languages; however, these responses were limited and not meaningful. This holds for the context of the University of Toronto, as goal modeling is not introduced at the undergraduate level. We collected additional demographic information from subjects, including program and year of study, gender, international student status, country of primary education, and their self described cultural identity. This additional information allowed

[1] See *https://doi.org/10.35482/csc.001.2025* for supplement.

TABLE II: Summer and Bike questions used in the “Smith” and replication experiment.

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	(“Smith” Experiment) For each of the elements listed below, how many times over the simulation does the element become Fully Satisfied? (*Replication Experiment*) *For each of the elements listed below, how many time point(s) over the simulation is the element Fully Satisfied?* (a) Have Summer Activity, (b) Pass Tryouts, (c) Exercise	(“Smith” Experiment) For each of the elements listed below, how many times over the simulation does the element become Fully Satisfied? (*Replication Experiment*) *For each of the elements listed below, how many time point(s) over the simulation is the element 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?	(“Smith” Experiment) Which intention would you choose to satisfy to make “Prevent Unloading in Bike Lane” Fully Satisfied? (*Replication Experiment*) *Which intention would you choose to satisfy to make “Access to Parking” Fully Satisfied?*
P4	Q9	On the previous page, we ask the question: [Q8]. You answered [Q8 choice]. Please explain your answer to this question.	On the previous page, we ask the question: [Q8]. You answered [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: [Q11]. You answered [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: [Q11]. You answered [Q11 choice]. Please explain your answer to the previous question.

us to monitor the diversity of our study population relative to that in Baatartogtokh et al. [15] (e.g., with respect to gender) and to assess whether randomization appropriately balanced study subjects between the four trial arms; see Sect. IV-A for details. Finally, we asked subjects about colors that they associate with good and bad outcomes (see RQ4 in Sect. IV-E).

Data Processing and Outcomes. We used two primary outcomes for assessing the use of EVO (collected in Periods 1 and 2 of Tbl. I): (a) the speed (in seconds) that it took subjects to complete the goal modeling questions (higher is worse) and (b) the number of correct responses given to those questions (higher is better). The time calculation consisted of adding the times for pages P1, P2, P3, and P5; see left-most column of Tbl. II. To calculate the score, we first performed qualitative coding on questions Q9, Q10, and Q12 (see Tbl. II) using the agreed rubric (see supplement[1]). Questions Q9 and Q12 validated the responses to Q8 and Q11, respectively, and were excluded from the score calculation. Two researchers independently coded each of the responses. We achieved an inter-rater reliability $\kappa = 0.84$ for the Summer questions and $\kappa = 0.87$ for the Bike questions (Cohen’s kappa coefficient). A third researcher joined the discussion to resolve disagreements. Next, we added together the number of correct responses to the remaining questions. Questions Q4 and Q6 were each scored out of 3 points, one point for each subquestion. Subjects’ final scores were thus adjudicated out of 14 possible points.

Additional secondary outcomes included the subjects’ quantitative assessment of the study instrument and materials. Subjects completed four optional material evaluations: one for the initial training (completed at the end of Period 0, Tbl. I), one each for the Summer and Bike models (at the end of the relevant study periods), and one for the EVO training materials (at the end of the period in which the training occurred). For example, `EBk-XSm` evaluated the initial training materials at the end of Period 0, the EVO training materials immediately after the training and the Bike model materials at the end of Period 1, and the Summer model materials at the end of Period 2. Each set of materials was evaluated based on the difficulty of (a) understanding the scenario description, (b) understanding the goal model, and (c) answering the goal modeling questions. Ratings went from 0 (no difficulty) to 10 (complete difficulty).

Statistical Analysis. The target sample size was $n = 56$ subjects (with 14 individuals per trial arm), following the recommendations made in Baatartogtokh et al. [15]. Subject demographics were summarized using medians and interquartile ranges (for numeric characteristics) or frequency distributions (for categorical characteristics), and their balance across the trial arms were assessed using Kruskal-Wallis [KW] tests or Fisher’s exact [FE] tests, respectively. The primary outcomes were evaluated using mixed-effects models with adjustment for the study period in which the measurement was taken, the experimental object, and any other characteristics found to be imbalanced between the four trial arms (see supplement[1] for full model specifications). The subjects’ evaluations of the experimental materials were also assessed using linear mixed-effects models, with additional adjustment for the sequencing of the experimental objects. We considered the study period and the experimental object to be potential modifiers of the effect of EVO; we tested for this by including interactions between each of these factors and the intervention and then evaluating these interactions using likelihood ratio tests [LRT]. Other secondary outcomes were analyzed descriptively.

We used Fisher's exact tests to examine differences in subject characteristics between our study and that of Baatartogtokh et al. [15]. We then reanalyzed the primary outcome data from Baatartogtokh et al. [15] using the mixed-effects models described above and combined both sets of study results using a one-step fixed-effects meta-analysis of the individual subject data (see supplement[1] for details) [22], [23]. We tested for the presence of effect heterogeneity by including interactions between EVO and the study site. In the absence of significant heterogeneity, we reported pooled estimates and 95% confidence intervals for the effects of EVO.

All statistical tests are two-sided and all p-values and confidence intervals are presented at their nominal level, without adjustment for multiple testing. We take $\alpha = 0.05$ as an indication of statistical significance throughout.

IV. Results

In this section, we first give an overview of our subject characteristics and then answer each of our research questions.

A. Subject Characteristics

Sixty-one students at the University of Toronto participated in the experimental study, with 60 of these individuals included in the primary analysis and one excluded due to non-response (see supplement[1] for details). Thirty-three of the subjects (55.0%) self-identified as male and 21 (35.0%) were international students. Subjects most commonly reported Canada (30 subjects; 50% of the sample), China (8; 13.3%), the United States (4; 6.7%), and India (4; 6.7%) as the location for the majority of their primary education. Most subjects (70%) indicated complete familiarity with English but had limited previous exposure to requirements engineering or Goal Modeling Languages: 58 (96.7%) indicated no familiarity with any of iStar, Tropos, or GRL.

These attributes were all balanced between the four trial arms (see supplement[1] for a table of characteristics and p-values for balance). There were likewise no significant differences in the speed at which each of the groups completed the goal modeling (KW, $\chi^2_3 = 2.44$, $p = 0.48$) or EVO (KW, $\chi^2_3 = 1.58$, $p = 0.66$) training modules nor in the number of questions that they answered correctly on the corresponding comprehension checks (KW, $\chi^2_3 = 0.54$, $p = 0.91$ for goal modeling; $\chi^2_3 = 3.59$, $p = 0.31$ for EVO training). A vast majority of the subjects (57; 95%) received a perfect score on the color vision test and all 60 individuals passed (i.e., received a score of 5 or greater out of 7), as required per the study's inclusion criteria.

We find no significant variation in the baseline characteristics of the four trial arms, indicating that the groups are broadly comparable and that no adjustment for baseline characteristics is necessary.

B. RQ1: Learning and Applying EVO to Goal Modeling Tasks

Given the baseline comparability of the four trial arms, we combine data from all 60 subjects when evaluating whether they could successfully learn and apply EVO from the provided training materials. The median time to review these materials was 3.32 minutes (interquartile range [IQR], 3.27 to 3.61 minutes), with a subsequent 1.35 minutes (IQR, 1.09 to 1.77 minutes) spent completing the six EVO training questions. All but one subject passed the EVO training module (defined as answering five or more questions correctly), and 54 subjects (90.0%) received a perfect score. All 60 subjects were included in subsequent analyses. Further, as we discuss in RQ2, all subjects achieved a passing score when using EVO to answer questions about the Bike and Summer models.

Subjects generally found the EVO training materials to be clear. When asked to rate the difficulty they encountered understanding the scenario, understanding the model, and answering the training questions, subjects' median responses were 1.00, 2.00, and 1.00, respectively, on a scale of 0 (no difficulty) to 10 (complete difficulty); forty-four subjects of the 57 who provided difficulty ratings (77.2%) rated all three components at 4.00 or below on the scale. See Sect. IV-D for a discussion of the perceived difficulty of the remaining study materials. As part of the training, subjects were asked to document questions they had about EVO. Six subjects (10%) responded. Subjects asked meta questions about the chosen colors and their relationship with element types and text colors. Two subjects asked questions about distinguishing between the different EVO modes (i.e., Percent/Time), though subsequent materials labeled which model was represented.

RQ1: In the initial training, almost all subjects were able to learn and use EVO to answer questions. Most subjects finished the training in under six minutes and rated it as low difficulty.

C. RQ2: Comparison of Performance with EVO vs. Control

Here we consider the two primary outcomes for the assessment of EVO: the time (in seconds) that it took subjects to complete the goal-modeling questions (listed in Tbl. II), capturing subjects' ability to make decisions quickly, and the number of correct responses provided to those questions, capturing their overall comprehension of the model.

Completion Time. Subjects completed the goal-modeling questions in as little as 92.56 seconds or as much as 1339.32 seconds, with the median completion time across all experimental objects, experimental conditions, and periods being 523.59 seconds (or 8.73 minutes) (see Fig. 4). Completion times were significantly faster during Period 2 than during Period 1, indicating the presence of a learning effect (mean difference, -127.24 seconds; 95% confidence interval [CI], -173.78 to -80.70; $p < 0.001$; Tbl. III). However, the magnitude of this learning effect did not significantly differ between the Bike and Summer models (LRT, $\chi^2_2 = 0.78$, $p = 0.68$) or between the EVO and control conditions (LRT, $\chi^2_2 = 2.91$, $p = 0.23$), indicating that there were no significant carryover effects due to either experimental object or experimental condition sequencing (see supplement[1] for details).

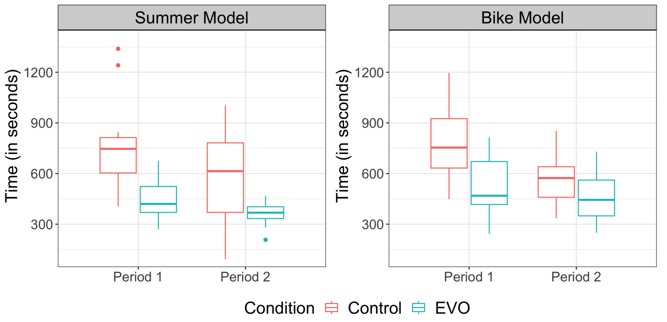


Fig. 4: Distribution of completion times for the experimental goal-modeling exercises, shown separately within each experimental object (i.e., Bike or Summer model), period, and experimental condition (i.e., EVO or control).

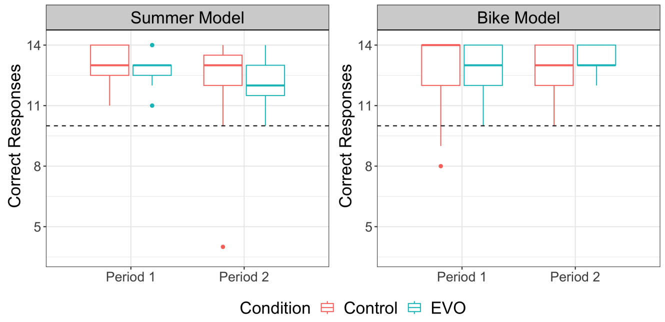


Fig. 5: Distribution of correct responses to the experimental goal-modeling questions, shown separately within each experimental object (i.e., Bike or Summer model), period, and experimental condition (i.e., EVO or control). Responses are scored out of 14 points; the dashed line indicates a satisfactory performance of 10 or more correct responses.

The median completion time under the EVO experimental condition was 419.60 seconds (6.99 minutes) and under the control condition was 655.97 seconds (10.93 minutes), almost a four minute difference (or 36% reduction) in median completion time. After controlling for both the study period and the experimental object encountered, we find that subjects' use of EVO was associated with significantly faster completion times: subjects answered questions 227.07 seconds (3.78 minutes) faster, on average, with EVO than without EVO (95% CI, -273.61 to -180.53; $p < 0.001$; see Tbl. III). There was weak but suggestive evidence of task inequivalence between the Bike and Summer models, with the Bike model questions taking subjects an average of 42.61 seconds longer than the Summer model questions (95% CI, -3.93 to 89.15; $p = 0.07$). The time benefits of EVO did not, however, significantly differ between these two experimental objects (LRT, $\chi^2_2 = 2.08$, $p = 0.35$).

Number of Correct Responses. Although subjects tended to complete goal-modeling questions faster when using EVO, there was no meaningful change in their understanding of the model, as measured by the number of correct responses. Fig. 5 displays the complete distribution of subject scores across all experimental objects, periods, and experimental conditions. The median number of correct responses was 13 out of 14 under both experimental conditions (IQR for both EVO and control, 12 to 14). Taking scores of 10 and above to indicate satisfactory performance on the goal-modeling exercise (i.e., $\geq 70\%$ of model questions correct), we see that all 60 subjects achieved satisfactory performances while using EVO; 57 out of the 60 subjects (95%) also achieved satisfactory performances under the control condition. After accounting for both the study period and the experimental object, we estimate that the odds of correctly answering a goal-modeling question are the same for EVO as for control (adjusted odds ratio $[OR]$, 1.00; 95% CI, 0.71 to 1.42; $p = 0.98$; see Tbl. III).

There were no statistically significant learning effects across the two study periods (adjusted OR, 0.89; 95% CI, 0.63 to 1.25; $p = 0.49$) nor significant differences in the effect of EVO between Periods 1 and 2, i.e., carryover effects (LRT, $\chi^2_2 = 0.09$, $p = 0.96$). There were also no significant differences in the odds of a correct response between the Bike and Summer models (adjusted OR, 1.38; 95% CI, 0.98 to 1.96; $p = 0.07$).

Subjects' Perspectives. In our qualitative analysis, 58 subjects (96%) said that they preferred the color view over the non-color view. 40 subjects mentioned that EVO made the model easier to read or understand, 15 mentioned that it was faster/quicker for answering questions, and 13 mentioned it was better for seeing the model *at a glance*. One subject preferred the non-color view, finding the colors distracting, while another did not have a view preference, stating that EVO was good for a quick glance but was overstimulating at times.

RQ2: We find significant evidence that using EVO is associated with faster completion times, but found no evidence of any impacts on correctness. A supermajority of subjects preferred the EVO over the control for its speed and ease of use.

D. RQ3: Study Experience and Subjects' Perceptions

Subjects assessed four facets of the study materials: the initial goal modeling training materials, the EVO training materials, and the two experimental objects (i.e., the Bike and Summer models). Results regarding the EVO training materials were summarized previously in Sect. IV-B; the remainder of the materials are discussed below.

Goal Modeling Training Materials. All 60 subjects provided quantitative assessments of the initial training sequence. However, two subjects reported having misinterpreted and reversed the difficulty scale when providing ratings of these and all other materials. We removed these individuals from the analysis, so that we had complete material assessments from 58 subjects (96.7%) in Period 0. A sensitivity analysis that includes the two subjects with errors is available online[1].

Subjects found the goal model in the initial training sequence moderately challenging to understand, with a median difficulty rating of 5.0 out of 10 (IQR, 2.25 to 7.00). The goal modeling scenario and corresponding goal modeling questions were somewhat more accessible, with median difficulty ratings of 2.0 (IQR, 1.0 to 4.0) and 3.0 (IQR, 2.0 to 6.0), respectively.

TABLE III: Mixed-effects analysis of the experimental goal-modeling exercise.

	Time to Complete Questions				Odds of a Correct Response			
	Mean Difference (in sec.)	L 2.5%	U 97.5%	P-Value[a]	Odds Ratio	L 2.5%	U 97.5%	P-Value[a]
Treatment				< 0.001				0.98
Control	*[Reference Level]*				*[Reference Level]*			
EVO	-227.07	-273.61	-180.53		1.00	0.71	1.42	
Period				< 0.001				0.49
Period 1	*[Reference Level]*				*[Reference Level]*			
Period 2	-127.24	-173.78	-80.70		0.89	0.63	1.25	
Experimental Object				0.07				0.07
Summer	*[Reference Level]*				*[Reference Level]*			
Bike	42.61	-3.93	89.15		1.38	0.98	1.96	

L 2.5% and U 97.5% refer to the lower and upper bounds, respectively, of the 95% Wald confidence interval for the corresponding effect estimate. [a]P-values calculated using Wald tests on the slopes from the (generalized) linear mixed-effects models.

TABLE IV: Perceived difficulty of the study instrument and goal-modeling tasks, with and without the use of EVO.

	Bike Model[a]		Summer Model[a]		Effect Size of EVO[b]	
	EVO	Control	EVO	Control	Adjusted Diff. in Mean Ratings (95% CI)	P-Value
(a) Understanding the Scenario Description	2.00 (1.00 to 3.25)	3.00 (1.00 to 5.00)	1.00 (1.00 to 2.00)	2.00 (0.00 to 4.00)	-0.50 (-0.94, -0.05)	0.03
(b) Understanding the Model	2.00 (0.75 to 4.00)	5.00 (3.00 to 6.00)	2.00 (1.00 to 5.00)	3.00 (1.00 to 5.00)	-1.01 (-1.61, -0.41)	0.002
(c) Goal-Modeling Questions	3.00 (1.00 to 5.00)	5.00 (2.00 to 7.00)	3.00 (1.00 to 5.00)	3.00 (2.00 to 5.00)	-1.23 (-1.92, -0.55)	0.001

[a]Subjects rated difficulty on a scale from 0 (no difficulty) to 10 (complete difficulty). Summaries refer to the subjects' ratings of the Bike study materials [resp. Summer study materials] and are reported as median (IQR). They are calculated using all available information taken over both study periods, i.e., combining Bike [resp. Summer] ratings across individuals who saw the Bike model first and individuals who saw the Summer model first. [b]Effect sizes and p-values were calculated using a mixed-effects model that adjusted for the experimental object (i.e., Bike or Summer model), the study period, and whether the subject saw the Bike or Summer model first. Confidence intervals are Wald confidence intervals and p-values are from Wald tests of the model slopes.

This is in line with our observation in Sect. IV-A that a supermajority of subjects in our replication study had no prior familiarity with goal modeling, such that the initial training sequence represented their first encounter with the field.

Experimental Materials. One subject opted not to complete the Period 2 materials assessment, so that we had complete and error-free ratings from 58 subjects (96.7%) in Period 1 and 57 subjects (95%) in Period 2.

Subjects generally found the experimental tasks to be less difficult than the initial training sequence: across each of the three dimensions subjects were asked to rate (understanding the scenario description, understanding the corresponding goal model, and answering the goal modeling questions), the median difficulty rating was 3.0 out of 10 in Period 1 (IQR, 1.0 to 5.0) and 3.0 out of 10 in Period 2 (IQR, 1.0 to 6.0); there were no significant differences in mean difficulty ratings across the two periods (mean difference in ratings, 0.22 points; 95% CI, -0.10 to 0.54; $p = 0.18$; see supplement[1]).

However, the Bike model did pose a greater challenge to subjects than the Summer model. Under the default coloring scheme, subjects found the Bike scenario and goal model to be more difficult to understand than those for the Summer experimental object and the Bike model questions to be more difficult to answer than the Summer questions (see Tbl. IV). The difficulty of the Bike model also carried over into how the subjects perceived the other study materials, with those who saw the Bike model in Period 1 finding the study materials to be significantly more challenging than those who saw the Summer model in Period 1 (mean difference in difficulty rating, 1.26 points out of 10; 95% CI, 0.20 to 2.32; $p = 0.02$).

Building on the qualitative results from Sect. IV-C, we find that the use of EVO meaningfully impacted subjects' ratings of the experimental materials and their difficulty (see Tbl. IV). Subjects found the scenario descriptions and models to be significantly easier to understand when encountered with EVO: among subjects who faced the same goal model in the same study period, those who saw the model with EVO rated the scenario 0.50 points less difficult to understand (95% CI, -0.94 to -0.05; $p = 0.03$) and the goal model 1.01 points less difficult to understand (95% CI, -1.61 to -0.41; $p = 0.001$), on average, than did those who saw the model with the default coloring scheme. They also rated the corresponding goal-modeling questions as significantly easier to answer (mean difference in ratings, 1.23 point decrease in difficulty; 95% CI, -1.92 to -0.55; $p < 0.001$).

RQ3: Subjects generally found the study instruments and experience to be moderately, though not overly, challenging, with the initial training materials rated as the most difficult to understand. Subjects found the Bike model materials to be significantly more challenging than the Summer model materials, with carryover effects on their perception of the remaining study instruments. After controlling for these differences and carryover effects, we found that subjects rated the experimental materials significantly easier to understand

and use effectively when they saw the materials with EVO.

E. RQ4: Color Associations and Preferences

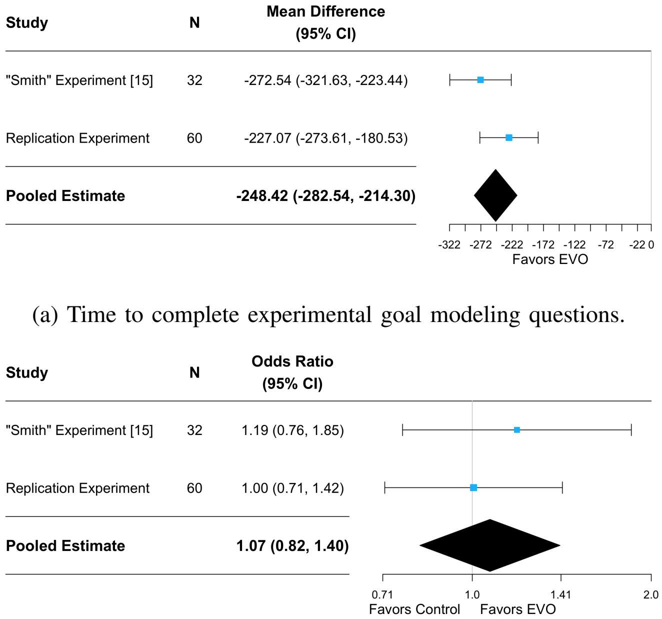
Prior to beginning the experiment, subjects were asked to list the color(s) that they associated with positive outcomes (i.e., good things happening) and with negative outcomes (i.e., bad things happening). Each of the following colors was identified by at least one subject as having positive associations: green (identified by 93.3% of the sample), yellow/gold (11.7%), blue (3.3%), pink (3.3%), red (3.3%), orange (1.7%), purple (1.7%), and white (1.7%). Responses for colors associated with negative outcomes were less varied. 93.3% of the subjects identified red as a negative color, 16.7% identified black/gray, and 1.7% identified each of yellow, orange, blue, and green. Notably, every individual who named red as a negative color also named green as a positive color; only four individuals (6.7%) did not self-report a green $\iff$ positive / red $\iff$ negative visual scale. Two of these subjects completed their primary education in China, one in Romania, and one in Canada. The subjects identified culturally as Chinese, Romanian, and Westernized Asian, respectively.

In the color-blind friendly palette used by EVO, blue coloring indicates fulfilled intentions (what one might consider to be *positive* outcomes), and red coloring indicates denied intentions (what one might consider to be *negative* outcomes). This kind of visual alignment (in which blue is positive and red is negative) was voluntarily mentioned by only two subjects (3.3%), both of whom also listed green as a positive color. See Sect. V-C for further discussion.

RQ4: We found most subjects associate red with bad outcomes. Green had the most prevalent association with good outcomes, with only a couple of subjects selecting blue.

F. RQ5: Comparison and Meta-Analysis of the "Smith" Experiment and Replication Experiment

The present study was undertaken in order to replicate the work of the "Smith" Experiment [15] in a different educational context and subject population, with the goal of exploring the robustness of the original study's findings. Tbl. V provides a formal comparison of the two study populations, as well as the conclusions reached by each of the two studies. As expected, subjects in the replication study had less *a priori* familiarity with requirements engineering and with goal modeling than their counterparts at Smith—University of Toronto students were 91.2% less likely than Smith students to report familiarity with any of iStar, Tropos, or GRL (3.3% vs. 37.5%; FE, $p < 0.001$). Subjects at the University of Toronto were also significantly less likely to report complete familiarity with English (FE, $p = 0.04$). Other key demographic differences include the high proportion of male subjects in the replication study (at 55.0% of the sample), while Smith College only admits individuals who identify as women. The "Smith" experiment [15] did not collect information on subjects' location of primary education or color preferences.



Study	N	Mean Difference (95% CI)
"Smith" Experiment [15]	32	-272.54 (-321.63, -223.44)
Replication Experiment	60	-227.07 (-273.61, -180.53)
Pooled Estimate		**-248.42 (-282.54, -214.30)**

(a) Time to complete experimental goal modeling questions.

Study	N	Odds Ratio (95% CI)
"Smith" Experiment [15]	32	1.19 (0.76, 1.85)
Replication Experiment	60	1.00 (0.71, 1.42)
Pooled Estimate		**1.07 (0.82, 1.40)**

(b) Odds of a correct response to a question.

Fig. 6: Forest plots of two studies investigating the effect of EVO on (a) speed and (b) comprehension of an experimental goal modeling task. The size of each blue square represents the relative weight of each study and the black diamond represents the combined effect estimate. The gray vertical line represents no effect.

Despite the heterogeneity of the two study populations, we find no statistically significant differences in the effect of EVO on the speed with which subjects completed the goal-modeling questions (LRT, $\chi^2_1 = 1.72$, $p = 0.19$) or on the odds that they answered the goal-modeling questions correctly (LRT, $\chi^2_1 = 0.35$, $p = 0.56$) (see full table online[1]). The pooled effect sizes are shown in Fig. 6. After accounting for both period (i.e., learning) effects and the experimental object, we again find that individuals using EVO answered the goal-modeling questions significantly faster than those using the default color scheme (mean difference, -248.42 seconds; 95% CI, -282.54 to -214.30; $p < 0.001$), without any significant effect on the likelihood that they answered those questions correctly (adjusted OR, 1.07; 95% CI, 0.82 to 1.40; $p = 0.62$). This absolute reduction in mean completion time translated to an estimated 30.0% to 41.4% reduction in the time spent on the goal-modeling task, depending on the study, study period, and experimental object subjects encountered (see supplement[1]).

RQ5: Despite significant differences in the study populations at the University of Toronto and Smith College, we found no significant differences in the benefits provided by EVO. Combining data from the two studies produced more precise estimates of EVO's impact on speed and comprehension; these estimates also generalize to a more diverse student population.

V. Discussion

A. Study Training and Debriefing

During the training modules and debriefing, subjects asked questions and made recommendations. We discussed the EVO

TABLE V: Comparison of the two studies included in the meta-analysis.

(a) Study Population and Characteristics

	"Smith" Experiment [15]	Replication Experiment	P-Value[a]
Country	United States	Canada	
Institutional Description	Small women's private liberal arts college	Large coeducational public research university	
Sample Size	32	60	
Subject Characteristics			
Complete Familiarity with English[b]	29 (90.6%)	42 (70%)	0.04
Some Familiarity with RE[c]	17 (53.1%)	22 (36.7%)	0.18
Some Familiarity with iStar, Tropos, or GRL[d]	12 (37.5%)	2 (3.3%)	< 0.001
Identify as Male	—[e]	33 (55.0%)	—

[a] P-values calculated using Fisher's exact test.
[b] Subjects rated familiarity on a scale from 0 (no familiarity) to 10 (complete familiarity). Reported proportions are the percent of subjects who answered 10.
[c] Reported proportions are the percent of subjects who gave a score greater than 0.
[d] Reported proportions are the percent of subjects who gave a score greater than 0 for at least one of the listed languages.
[e] Information not collected by Baatartogtokh et al. [15].

(b) Results

	"Smith" Experiment [15]	Replication Experiment	Comparison
RQ1: To what extent are subjects able to learn and apply EVO?			
	100% (32/32) passed EVO training 78% (25/32) achieved a perfect score	98.3% (59/60) passed EVO training 90% (54/60) achieved a perfect score	Similar result
RQ2: How does EVO compare to control in terms of:			
Completion time?	Significantly faster completion of bike ($p = 0.012$) and summer ($p = 0.002$) models	Significantly faster completion, holding model and period constant (time benefit: 3.78 min.; $p < 0.001$)	Similar result; replication also estimates magnitude of gain
Correct responses?	Correct response rate not significantly different for bike ($p = 0.50$) or summer ($p = 0.06$) model	Odds of correct response not significantly different, holding model and period constant (OR: 1.00; $p = 0.98$)	Similar result; replication also estimates magnitude of difference
Subject preferences?	100% (32/32) prefer EVO	96.7% (58/60) prefer EVO	Similar result
RQ3: How do subjects rate the study experience and instrument?			
	Qualitative analysis only	Significantly less difficult when encountered with EVO ($p < 0.001$)	Novel replication result
RQ4: To what extent do subjects associate blue (resp. red) with good (resp. bad) outcomes?			
	—	3.3% (2/60) subjects self-reported this visual scale	Novel replication result

training specific questions in Sect. IV-B.

After the goal modeling training, 37 subjects (61%) asked a question. Most subjects asked for further clarification on the contribution links and how valuations were assigned to the model. Specifically, twelve subjects asked about the difference between symmetric and asymmetric contribution types, while six asked about the strength of the influence (i.e., + vs. ++) in contributions. One asked what it meant to invert evidence and another asked for contribution link specific examples. Seven subjects were confused about how evidence pairs or links are assigned to the model, with one asking if they are allowed to make subjective decisions. One questioned the choice of valuations used in the model. Three subjects asked what it means for there to be a conflicting evidence pair in the model. Seven subjects also asked about how valuations are propagated throughout the model. Another subject asked how an individual is able to make decisions with evidence pairs if they are able to prioritize one goal over the other.

During debriefing, 52 subjects (86% response rate) left 1-5 substantive comments for "the developers of the goal modeling language (and tool)". Five subjects left comments about improving the study materials in some way (e.g., figure size, video captioning). In Tbl. VI, we list common recommendations, i.e., recommendations that occurred three or more times in the dataset. As mentioned above, almost all subjects preferred the color view, yet ten (16%) recommended changing the actual colors used. Twelve (20%) recommended changes to the visualization of the Precent and Time modes.

Although anecdotal, we observe from this data that subjects have trouble understanding contribution links; but, we cannot separate between whether the problem lies in the difficulty of understanding the foundational language or inadequacies in the training provided in this study. Future replication studies of EVO should improve the training materials and videos to better explain contribution links. We also recommend investigating improvements to the underlying notation.

B. Recommendations for Tool Developers

Given the recommendations in Tbl. VI, especially those also present in the "Smith" Experiment, we recommend that tool developers investigate their impact. For example, adding ticks to the Percent/Time mode is a straight forward change; yet, it is uncertain whether it would improve readability or make the model cluttered. Subjects proposed improving the visualization of links, which could be addressed using color as suggested, by varying line thickness or patterns (i.e., dashed/solid lines) to indicate contribution strength, or with different labels for link types. Finally, subjects recommended adding goal prioritization, highlighting, and model slicing. Given the recent work on presence conditions in goal models [24], model slicing appears to be feasible and should be explored to allow users to focus in on an individual actor or on a set of links in the model.

TABLE VI: Common recommendations for improvement.

EVO Improvements
- Change the color palette (not intuitive). (6)
- Change the conflict colors only. (4)
- Add numbers to Percent/Time mode. * (7)
- Add ticks to Percent/Time mode. (5)
Goal Modeling Improvements
- Add goal prioritization, highlighting, or model slicing. (7)
- Improve visualization of links (maybe with color). (5)
- Make contribution links type difference more intuitive. (5)
- Change the shape or default color of intentions. * (5)
- Make model text more readable. (5)

Bracketed numbers indicate counts of occurrence.
'*' indicates recommendations not present in Baatartogtokh et al. [15].

At a more foundational level, subjects suggested modifying the shape or default color of intentions. Intention types are represented using two encodings: shape as the primary visual variable and color as the secondary variable. In contrast, intention valuations are represented solely as text, unless EVO is enabled, upon which color becomes the primary variable. This may cause confusion with the secondary variable for type, as there is an overlap in the meaning of an intention's color, which may reduce the the cognitive effectiveness of goal modeling. We recommend changing the intention type colors to further differentiate between intention type and valuation so that encodings do not overlap with each other. While prior research advises using color as a secondary encoding [25], EVO's reliance on it as a primary variable introduces tradeoffs that require further investigation.

C. Blue vs. Green

In RQ2 (see Sect. IV-C), we found that all but two subjects preferred EVO to the control; yet, in Sect. V-A, ten subjects (16%) recommended changing the colors in EVO. As introduced in Sect. I, most goal modeling tools use green to denote good outcomes. In RQ4 (Sect. IV-E), we found that most subjects in our study associate green, and not blue, with good outcomes. While we may suspect that these preferences may differ in populations with high prevalence of green-red color-blindness [26], the default colors of EVO need to be revisited. The blue-red palette is more visible for color vision deficient users but less intuitive for the rest of the population.

D. Recommendations for Researchers

The replication package created by Baatartogtokh et al. [19] was thorough but incomplete. The repository did not have a data-dictionary for the final data files or rubric for the qualitative analysis, so we created both. The statistical analysis spanned multiple python and R files which made it confusing to understand and update. We recommend minimizing the complexity of data aggregation procedures.

As already mentioned in Sect. V-A, subjects had significant problems understanding the contribution link types. While this was mentioned in the original experiment [15], it seemed more prominent in our replication. In hindsight, we should have updated the training module that discussed contribution links; however, explaining the syntax and semantics of twelve different contribution types in Tropos seems to be a possible area of future research. We recommend that future experiments similar to ours use only the four symmetric contribution types.

E. Implications for Practitioners

This study is part of a larger effort to improve the applicability and adoption of goal modeling [4] and is one part of the "improve and measure" cycle. Our replication allows for a higher level of confidence in the benefits of EVO for the goal modeling community, including tool developers. Our findings reveal the need to balance intuitiveness and inclusiveness in choosing colors for highlighting elements in goal models.

Thus far, our evaluation of EVO is limited to undergraduate students. While this is not ideal, we have limited access to industrial collaborators and welcome the future involvement of practitioners. Yet, by studying student outcomes, we can improve goal modeling tools for future industrial deployment. Students represent the next generation of industrial practitioners, and by training students in goal modeling, we can have future industrial impact. In particular, we believe that EVO's ability to reduce the cognitive load of using and making decisions with goal models enhances learning, which in turn makes the industrial adoption of goal modeling more realistic in the future.

F. Threats to validity

Conclusion Validity. To reduce researcher bias, separate authors conducted the in-person session, data preparation, and statistical analysis. The lab environment and handouts were kept consistent to minimize variation in subject experience. We found no significant differences in trial arms based on measured subject attributes and training performance (see Sect. IV-A for details); however, there may be other unmeasured factors or sources of heterogeneity in subjects' comfort with goal modeling. See Sect. III for our detailed discussion of statistical analysis that may affect conclusion validity.

Internal Validity. We mitigated fatigue effects by allotting untimed break points for subjects. As in the original study, the study design carries the risk of carryover and learning-by-practice effects. Additionally, there is the instrumentation effect as the Bike model was detected to be harder than the Summer model. We mitigated these threats by controlling for these differences in our statistical analysis.

Construct Validity. We discuss potential threats from the dependence between our replication and the original study. Both studies used the same method of measuring subjects' perceptions of EVO, so there is a threat of mono-method bias. A future evaluation using different methods would help ensure that the results of this replication are not due to similar biases from the original. There may be instrumentation bias, as we used the same training materials and questions as the original study and so potentially replicated any limitations or biases. However, we mitigated this threat by making the changes listed in Sect. III. Finally, there is the risk of hypothesis guessing and evaluation apprehension as our subjects were novices.

External Validity. The experimental setting is not reflective of how stakeholders would use goal modeling and EVO in reality, as the models and scenarios were simple and not personalized due to constraints on time. As our subjects are undergraduate students with no prior experience in RE and goal modeling, our results do not reflect how experts would use EVO.

VI. Related Work

Color and Reasoning in Goal Modeling. The green-red color scheme was first used in goal modeling to highlight the valuations of URN models in the jUCMNav tool [6], [7]. This same color scheme was used for iStar evaluation labels in OpenOME [5]. In the context of quantitative non-functional requirements, Oliveira and Leite [14] used the green-red color scheme and colormetrics to calculate an exact hue based on the propagated valuations in the model, and colored links to indicate the polarity of propagation. Outside of analysis, color has also been used to identify aspects of a goal model, such as root and leaf nodes for iStar models in OpenOME [5]. In piStar, users can assign each intention any color of their choice [27]. Sartoli et al. mapped colors to the model elements, where purposes are purple and obligations are green [11]. Perera et al. introduced level-wise value evaluation, with colors indicating which "level" or phase of development a design choice was made at [28]. As mentioned in Sect. II, Varnum et al. introduced EVO and the blue-red color scheme with the aim of making Tropos evidence pairs easier to interpret [9]. BenAyed et al. extended EVO to introduce four more palettes and a customizable palette option to improve the accessibility of EVO [29], which was explored in [30] through a user study. The experimental validation of EVO by Baatartogtokh et al. [15] is the basis for our replication study.

Visualizations in SE. This study may impact the broader field of software visualizations [31], [32]. To aid decision-making for future software development, prior work used colors to represent properties such as code size within a 3D visualization of software release histories, helping stakeholders better understand the system [33]. Feigenspan et al. [34] used color to highlight program backgrounds to improve code comprehension and found that subjects favored the use of color, though color choice is important. Building on these insights, our findings suggest that users preferentially view green, rather than blue, as indicative of positive outcomes, providing further guidance for color choices in SE visualizations. This preference may also extend to other types of visualizations, such as dependency graphs (e.g., [35]).

Replications in SE. Replications are an important but overlooked aspect of empirical software engineering [36]. Sjøberg et al. found only 18% of surveyed experiments to be replications. In a subsequent study focused on replications, da Silva et al. found an increase in replications after the publication of Sjøberg et al., with the majority being internal replications [37]. As mentioned in Sect. III, SE researchers have long discussed the merits of replication types. Basili et al. [38] and Shull et al. [16] argue for using reference materials and conducting exact replications, while Miller [39] and Kitchenham [40] argue for theoretic replications using different methods and materials to explore the original hypotheses. We heed Kitchenham's warning and based our study on a strong experimental design with a detailed study guide (see [19]); thus, our results enable the original and future researchers to continue building a theory of the impacts of EVO. We contributed to theory building by extending the external validity of the original findings. Future theoretic replications would complement our analysis.

Mendonça et al. proposed the FIRE approach consisting of two cycles [41]: an internal cycle to improve on the original study, and an external cycle to independently validate the findings. Gómez et al. proposed five elements that vary between SE experiments and their replication: site, experimenters, apparatus, operationalizations, and population properties [17]. Our study used a different site and different population properties while keeping the apparatus, experimental objects, and operationalization the same. In subsequent work, Juristo et al. differentiate between close replications (i.e., pseudo-exact) and differentiated replication [42]. Finally, Wieringa argued that researchers should differentiate between replications that extend statistical inference from samples to populations and those replications that use cases to test the underlying theory under analysis [43], with this study being the former.

VII. Conclusions and Future Work

In this paper, we replicated the experiment by Baatartogtokh et al. [15]. We conclude that using EVO to answer goal modeling questions is associated with faster completion times, but found no evidence of any impacts on correctness. Subjects had a positive response to EVO, and preferred it over the control. We also found no significant differences between the results of the original experiment and our replication, allowing us to calculate pooled estimates of EVO's impact. In particular, subjects using EVO completed the goal modeling questions 4.14 minutes (between 30.0% and 41.1%) faster, on average, than the control. Additionally, we found that most subjects associated green with good outcomes, while only a few associate blue (the default color in EVO) with good outcomes. Thus, we recommend changing the default color palette of EVO to align with subject expectations and other goal modeling tools.

In future work, the EVO experiment should be replicated with trained goal modelers and practitioners to verify if the results hold among non-novices, as it is unclear if experts would benefit from the use of EVO, as well as different experimental objects (i.e., larger models). We also recommend a head-to-head comparison of the blue-red and the green-red palettes, with color vision normal and deficient users.

Acknowledgments

We thank our study subjects for their participation. We thank Marsha Chechik for collaborating on this project and hosting us at the University of Toronto in order to conduct this study.

References

[1] K. Pohl, *Requirements Engineering: Fundamentals, Principles, and Techniques*, 1st ed. Springer Berlin, Heidelberg, July 2010.

[2] A. van Lamsweerde, *Requirements Engineering - From System Goals to UML Models to Software Specifications*, 1st ed. Wiley, January 2009.

[3] E. Yu, P. Giorgini, N. Maiden, and J. Mylopoulos, *Social Modeling for Requirements Engineering: An Introduction*. MIT Press, October 2010.

[4] 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'17)*, 2017, pp. 174–183.

[5] J. Horkoff, Y. Yu, and E. Yu, "OpenOME: An Open-source Goal and Agent-Oriented Model Drawing and Analysis Tool," in *Proceedings of 5th International I* Workshop (iStar'11)*, 2011, pp. 154–156.

[6] G. Mussbacher, J. Whittle, and D. Amyot, "Semantic-Based Interaction Detection in Aspect-Oriented Scenarios," in *Proceedings of the IEEE 17th International Requirements Engineering Conference (RE'09)*, 2009, pp. 203–212.

[7] G. Mussbacher and D. Amyot, "On Modeling Interactions of Early Aspects with Goals," in *Proceedings of Early Aspects at ICSE: Workshops in Aspect-Oriented Requirements Engineering and Architecture Design (EA'09)*, 2009, pp. 14–19.

[8] M. Salnitri, E. Paja, M. Poggianella, and P. Giorgini, "STS-Tool 3.0: Maintaining Security in Socio-Technical Systems," in *Proceedings of the International Conference on Advanced Information Systems Engineering (CAiSE'15)*, 2015, pp. 205–212.

[9] 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'20)*, 2020, pp. 79–84.

[10] J. Horkoff and E. Yu, "Finding Solutions in Goal Models: An Interactive Backward Reasoning Approach," in *Proceedings of the 18th IEEE International Requirements Engineering Conference (RE'10)*, 2010, pp. 59–75.

[11] S. Sartoli, S. Ghanavati, and A. S. Namin, "Towards Variability-Aware Legal-GRL Framework for Modeling Compliance Requirements," in *Proceedings of the IEEE 7th International Workshop on Evolving Security and Privacy Requirements Engineering (ESPRE'20)*, 2020, pp. 7–12.

[12] Industry, Joint, "Vehicle Traffic Control Signal Heads-Light Emitting Diode (LED) Circular Signal Supplement," Institute of Transportation Engineers, Washington (DC), Tech. Rep., 2005.

[13] S. Silva, B. Sousa Santos, and J. Madeira, "Using Color in Visualization: A Survey," *Computers and Graphics*, vol. 35, no. 2, pp. 320–333, 2011.

[14] 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'20)*, 2020, pp. 39–48.

[15] Y. Baatartogtokh, I. Foster, and A. M. Grubb, "An Experiment on the Effects of using Color to Visualize Requirements Analysis Tasks," in *Proceedings of the IEEE 31st International Requirements Engineering Conference (RE'23)*, 2023, pp. 146–156.

[16] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo, "The Role of Replications in Empirical Software Engineering," *Empirical Software Engineering*, vol. 13, pp. 211–218, 2008.

[17] O. S. Gómez, N. Juristo, and S. Vegas, "Replications Types in Experimental Disciplines," in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'10)*, 2010, pp. 1–10.

[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] Y. Baatartogtokh, I. Foster, and A. M. Grubb, "An Experiment on the Effects of using Color to Visualize Requirements Analysis Tasks: Supplemental Material," Computer Science: Faculty Publications, Smith College, Northampton, MA, pp. 146–156, 2023. [Online]. Available: https://doi.org/10.35482/csc.002.2023

[20] M. P. Robillard, D. M. Arya, N. A. Ernst, J. L. C. Guo, M. Lamothe, M. Nassif, N. Novielli, A. Serebrenik, I. Steinmacher, and K.-J. Stol, "Communicating Study Design Trade-offs in Software Engineering," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 5, June 2024. [Online]. Available: https://doi.org/10.1145/3649598

[21] Office of Institutional Research, "Common Data Set 2022-2023," Smith College, Tech. Rep., 2023, Available at *https://www.smith.edu/your-campus/offices-services/institutional-research/data-about-smith*, accessed 03/09/2024.

[22] R. D. Riley, P. C. Lambert, J. A. Staessen, J. Wang, F. Gueyffier, L. Thijs, and F. Boutitie, "Meta-Analysis of Continuous Outcomes Combining Individual Patient Data and Aggregate Data," *Statistics in Medicine*, vol. 27, no. 11, pp. 1870–1893, 2008.

[23] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-Analysis of Individual Participant Data: Rationale, Conduct, and Reporting," *BMJ*, vol. 340, 2010.

[24] X. Bi and A. M. Grubb, "Incorporating Presence Conditions into Goal Models that Evolve Over Time," in *Proceedings of the 13th International Model-Driven Requirements Engineering Workshop (MoDRE'23)*, 2023, pp. 272–276.

[25] 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.

[26] J. Birch, "Worldwide Prevalence of Red-green Color Deficiency," *Journal of the Optical Society of America A*, vol. 29, no. 3, pp. 313–320, 2012.

[27] J. Pimentel and J. Castro, "piStar Tool – A Pluggable Online Tool for Goal Modeling," in *Proceedings of the IEEE 26th International Requirements Engineering Conference (RE'18): Posters & Tool Demos*, 2018, pp. 498–499.

[28] H. Perera, G. Mussbacher, W. Hussain, R. Ara Shams, A. Nurwidyantoro, and J. Whittle, "Continual Human Value Analysis in Software Development: A Goal Model Based Approach," in *Proceedings of the IEEE 28th International Requirements Engineering Conference (RE'20)*, 2020, pp. 192–203.

[29] C. Ben Ayed, S. Halili, Y. Tan, and A. M. Grubb, "Toward Internationalization and Accessibility of Color-based Model Interpretation," in *Proceedings of the 16th International iStar Workshop (iStar'23)*, 2023.

[30] Y. Baatartogtokh, I. Foster, and A. M. Grubb, "A Splash of Color: A Dual Dive into the Effects of EVO on Decision-making with Goal Models," *Requirements Engineering*, vol. 29, no. 3, pp. 371–402, 2024.

[31] N. Chotisarn, L. Merino, X. Zheng, S. Lonapalawong, T. Zhang, M. Xu, and W. Chen, "A Systematic Literature Review of Modern Software Visualization," *Journal of Visualization*, vol. 23, pp. 539–558, 2020.

[32] L. Merino, M. Ghafari, C. Anslow, and O. Nierstrasz, "A Systematic Literature Review of Software Visualization Evaluation," *Journal of Systems and Software*, vol. 144, pp. 165–180, 2018.

[33] H. Gall, M. Jazayeri, and C. Riva, "Visualizing Software Release Histories: The Use of Color and Third Dimension," in *Proceedings of the IEEE 15th International Conference on Software Maintenance (ICSM'99)*, 1999, pp. 99–108.

[34] Feigenspan, Janet and Kästner, Christian and Apel, Sven and Liebig, Jörg and Schulze, Michael and Dachselt, Raimund and Papendieck, Maria and Leich, Thomas and Saake, Gunter, "Do Background Colors Improve Program Comprehension in the ifdef Hell?" *Empirical Software Engineering*, vol. 18, no. 4, pp. 699 – 745, 2013.

[35] A. E. Hassan and R. C. Holt, "ADG: Annotated Dependency Graphs for Software Understanding," in *Proceedings of the IEEE International Workshop on Visualizing Software for Understanding and Analysis (VISSOFT'03)*, 2003, pp. 41–45.

[36] A. Brooks, M. Roper, M. Wood, J. Daly, and J. Miller, "Replication's Role in Software Engineering," *Guide to Advanced Empirical Software Engineering*, pp. 365–379, 2008.

[37] Fabio Q. B. da Silva, Marcos Suassuna, A. César C. França, Alicia M. Grubb, Tatiana B. Gouveia, Cleviton V. F. Monteiro, and Igor Ebrahim dos Santos, "Replication of Empirical Studies in Software Engineering Research: A Systematic Mapping Study," *Journal of Empirical Software Engineering*, vol. 19, no. 3, pp. 501–557, June 2014.

[38] V. R. Basili, F. Shull, and F. Lanubile, "Building Knowledge Through Families of Experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 456–473, 1999.

[39] J. Miller, "Replicating Software Engineering Experiments: A Poisoned Chalice or the Holy Grail," *Information and Software Technology*, vol. 47, no. 4, pp. 233–244, 2005.

[40] B. Kitchenham, "The Role of Replications in Empirical Software Engineering—A Word of Warning," *Empirical Software Engineering*, vol. 13, pp. 219–221, 2008.

[41] M. G. Mendonça, J. C. Maldonado, M. C. De Oliveira, J. Carver, S. C. Fabbri, F. Shull, G. H. Travassos, E. N. Höhn, and V. R. Basili, "A Framework for Software Engineering Experimental Replications," in *Proceedings of the IEEE 13th International Conference on Engineering of Complex Computer Systems (ICECCS'08)*, 2008, pp. 203–212.

[42] N. Juristo and S. Vegas, "The Role of Non-exact Replications in Software Engineering Experiments," *Empirical Software Engineering*, vol. 16, pp. 295–324, 2011.

[43] R. Wieringa, "Empirical Research Methods for Technology Validation: Scaling up to Practice," *Journal of Systems and Software*, vol. 95, pp. 19–31, 2014.