IADIS International Journal on Computer Science and Information Systems
Vol. 7, No. 2, pp. 129-151
ISSN: 1646-3692

EVALUATION OF “SAFETY-DOMINO”: A GRAPHICAL METAPHOR FOR SUPPORTING MINIMAL CUT SET ANALYSIS

Yasmin I. Al-Zokari. *The University of Kaiserslautern, Department of Computer Science, Kaiserslautern, Germany*

Dirk Zeckzer. *The University of Kaiserslautern, Department of Computer Science, Kaiserslautern, Germany*

Peter Dannenmann. *RheinMain University of Applied Science, Department of Engineering, Rüsselsheim, Germany*

Liliana Guzman.
The University of Kaiserslautern, Department of Computer Science, Kaiserslautern, Germany
University of Valparaíso, Chile

Hans Hagen. *The University of Kaiserslautern, Department of Computer Science, Kaiserslautern, Germany*

ABSTRACT

Embedded systems, which are safety critical, are frequently analyzed to find out whether they follow the safety standards and to improve their safety. Fault tree analyses enable the safety analysts to find possible causes of a system’s failure or unsafe behavior. From a fault tree, minimal cut sets are computed giving unique combinations of basic events that cause such a failure. Using minimal cut sets, the safety analysts try to find out how to improve the safety of the system with a restricted amount of resources, e.g., cost and time. The two most important results of minimal cut set analyses are the order (number of basic events) and the failure probability of the minimal cut sets. Small order minimal cut sets having 1-3 basic events are the most important ones, because, e.g., a single point of failure is more likely to occur than multiple failures at the same time. However, there is a lack of tools that visualize both failure probability and order of minimal cut sets. Therefore, we presented in our previous paper “Safety-Domino” the design of a graphical metaphor for visualizing the minimal cut set’s order together with its failure probability. The design decisions were based on perception theory. To find the optimal color scheme for the design, we performed an empirical evaluation with 32 participants. This paper presents the extended version of an earlier paper [AZZD*12] with an extended analysis of the results and additional preference tests. The statistical results of this evaluation are extended to show the tendency and relative accuracy of the participants using the different color schemes. These results indicate that the color scheme we call “gray-opposite” performs slightly better than the color scheme we call “black” for larger problem sizes. However, from the participants’ feedback and the additional preference tests (monitor-based and paper based) the black color was preferred. Therefore, we propose to interpret the results of any evaluation considering the feedback of the participants and including additional preference tests. In additional, we performed correlation tests to know if the demographic features of the participants had an effect to the results found.

KEYWORDS

Safety Visualization, Number Visualization, Basic Events, Minimal Cut Set Order, Minimal Cut Set Failure Probability.

1. INTRODUCTION

In order to ensure that a system or an embedded system is operating safely, several analyses are performed. These analyses are used to detect risks and hazards, and to eliminate, to avoid, and/or to reduce them [BV10]: p. xiii. A safety-critical system, e.g., an autonomous vehicle, could cause different levels of harm to people (i.e., injury, death) or to the environment (e.g., polluting the environment) [Sav]. Minimal Cut Set (MCS) analysis is used for safety analysis. The two major metrics of MCSs used in MCS analysis are their failure probability and their order. However, almost all tools performing safety analysis [AZYZ*11] do not represent these measurements graphically. CakES and ViSSaAn are the first visualization tools tackling this issue by using visualization to ease the analysis and exploration of the system and its hazards. In [AZZD*12] the safety-domino---a visual metaphor for representing the MCSs order---was presented together with the preliminary results of our evaluation thereof. This paper presents the results of this evaluation in detail.

1.1 Safety Analysis Techniques

We briefly introduce the most important concepts of safety analysis underlying our approach in this section. Fault Tree Analysis is a method used in system safety studies [BV10] based on so-called fault trees ([BV10]: p. 54). Tools performing fault tree analysis should be able to suitably display the fault tree analysis results, e.g., Minimal Cut Sets. A basic event (BE) occurs with a certain failure probability (FP) ([BV10]: p. 56ff). Minimal Cut Sets (MCSs) are the unique combinations of basic events [BV07] and are important indicators for safety problems. The number of these BEs in a MCS is the MCS order. Our research is motivated from the safety analyses domain. The safety tasks, MCSs sub-tasks, and tools are discussed in detail in [AZZD*12]. To identify hazards with high failure probability and the possible combinations of components MCSs analysis is used. The most important metrics in this analysis are the failure probability and the order of the MCSs [VGRH81, KZ96]. A detailed description is given in [AZZD*12]. A MCS with order “1” is called a single point of failure and has to be eliminated from the hazard and therefore from the system.

The current tools are very powerful in modeling fault tress and in computing additional safety measurements. However, they lack a simple and comprehensible graphical representation of those measurements supporting the tasks of the engineers. The reason is that the results are either represented as text or tables. These results are also sometimes incomplete. Thus, the tasks are time consuming and prone to human error ([BV10]: p. xiii). Examples were presented in [AZZD*12, AZYZ*11]. From this we conclude that there is a need for visualizations supporting the safety analysts to solve their tasks more efficiently.

In [AZKS*11, AZZD*12, AZZH11, and AZSZ*11], an MCSs analysis visualization tool was integrated that addresses some of the safety tasks to support an efficient and effective safety analysis of safety critical systems. In this tool, the MCSs with high failure probabilities are represented by red circles, those with moderate failure probability by yellow circles, and those with low failure probability by green circles. The order of the MCS was represented by colored “dominos” (see Section 0, [AZZD*12]). In this paper, we report, an extended evaluation of the color choice for the dominos (Section 2).

1.2 Techniques for Number Visualization

As described in [AZZD*12], numbers between 1 and 6 are the most important ones for MCSs order. Therefore, we studied the different possibilities for representing them. Numbers can be represented in many ways, such as digits, textures, colors, and shapes (more details in [AZZD*12]). However, Digits, textures, colors, and shapes are not suitable for our purpose of representing the MCSs’ order, because of the need of training and the limited visual short term memory of humans. Colors distract the user from the MCS’s failure probability, since people usually focus on the things with highest luminance. As a result, we did not use them for our representation.

For this reason, in [AZZD*12] we adapted the domino representation of numbers to represent the order of the MCSs. This is in line with the results of [Bec66] and [War04], who found that circles are easier to distinguish than other shapes and textures. We tried many masks for positioning the dots (Figure 1). Using random positions of dots is neither easy to understand nor easy to remember. However, the domino positioning of dots is very easy to understand and to remember, because they are symmetric ([War04]: p. 192). Therefore, we decided to use domino positioning for representing the MCSs’ size (Figure 2). As we described in Section 0, in the safety domain, the smaller the MCSs’ size, the more safety critical the MCS. For providing the best result and a correct impression, we considered different dot sizes. Using “equal size” for all dots was not too good, because this gave the opposite impression of importance (Figure 2). Using “small size for smallest numbers” gave the opposite impression of the importance and the representation gets very crowded (Figure 3 (top)), while using “large size for smallest numbers” gave the correct impression of the importance (Figure 3 (middle)). Our final design (Figure 3 (bottom)), the smallest number has the largest size and the orientation of the number “3” is different from that of number “2”. Thus, they are better distinguishable from each other. Initially, all dots were colored black.

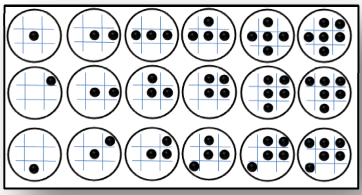


Figure 1. Using dots with random positions to represent numbers.

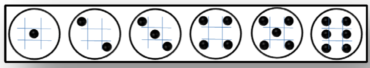


Figure 2. Using domino positions of dots to represent numbers.

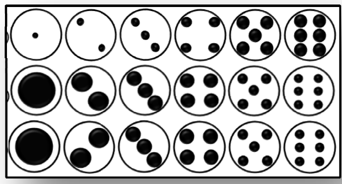


Figure 3. Different sizes for the dots: Top) opposite importance impression Middle) direct importance impression Bottom) the “2” dot orientation is reversed to ease the distinction from the “3” dot

In addition to the different levels of gray evaluated in [AZZD*12], three color schemes were evaluated in Experiment II to find out, which one is the most suitable for the safety-domino representation. The reason of this additional evaluation is to find out if choosing “black” and “Opposite-Gray” other than any color scheme in Experiment I was a successful decision in [AZZD*12].

We chose one of the standard known contrasting color schemes [Hun00], p. 53. This is, because they have high contrast and are best for highlighting [PRL12]: p. 256, 280. The contrasting color schemes can be categorized to three groups: complementary, split complementary, and triadic. From the contrasting color schemes, we chose the complementary one, because it has a one-to-one mapping.

We used three complementary color schemes: the invert (complementary), the complementary (traditional complementary)-same, and the complementary (traditional complementary)-opposite. The invert color scheme is based on Goethe’s color wheel also called the artist’s wheel. In our case, the inverted colors of (red, yellow, green) are (green, violet, red) [wiki12]. The traditional complementary is based on Itten [JI60] (see also [War04]: p. 124f) with the harmony color wheel also called the HSV color wheel. Here, the opposite colors on the color wheel of (red, yellow, green) are (cyan, blue, magenta). From this complementary color scheme, we derived two color schemes: complementary-same with complementary hue and the same saturation, and complementary-opposite with complementary hue and opposite saturation. In both cases, we kept the lightness constant (adapting [War04]: p. 136, to our situation).

2. COLOR SCHEME EVALUATION

In order to know whether the gray-opposite color scheme is better supporting the MCS analysis tasks (Section 1.1) than the black color scheme, we performed an empirical evaluation (Experimental Treatment I). To find out whether the additional color schemes presented above are suitable choices, we introduced additional experimental tasks and preference questions (Experimental Treatment II).

2.1 Research Methodology

In the following subsections, we present an extended version of the empirical evaluation reported in [AZZD*12]. We extended the work presented in [AZZD*12] by including the analysis of the relative accuracy as well as the analysis of the user preferences into the evaluation scope.

2.1.1 Research Goals and Hypotheses

In general, our research goal is to compare the efficiency and the accuracy achieved by safety engineers when they analyze MCSs between using gray-opposite colored dots and black colored dots. In particular, efficiency means here the time required by a safety engineer for counting circles (i.e., MCSs) with a predefined number of dots (i.e., MCS order) in a given visualization. Accuracy refers to the tendency of the accuracy and to the relative accuracy. The tendency of the accuracy is the difference between the number of circles counted and the true number of circles with a predefined number of dots in the visualization, while the relative accuracy is the ratio between the found circles and the true number of circles with a predefined number of dots in the visualization. Consequently, we defined the following research hypotheses:

RH1: The efficiency achieved by safety engineers using gray-opposite colored dots in identifying MCSs with small orders differs from the efficiency achieved by safety engineers using black colored dots.

RH2a: Safety engineers using gray-opposite colored dots achieve a better relative accuracy in identifying MCS with small orders than safety engineers using black colored dots.

RH2b: Safety engineers using gray-opposite colored dots achieve a better tendency of accuracy in identifying MCS with small orders than safety engineers using black colored dots.

Here, a “better relative accuracy” implies a value equal or near to 1, while a “better tendency of accuracy” implies a value equal or near to 0. In addition, we also explored user preference, i.e., “easy to count” (easy to use), “comfortable for the eyes” (is not tiring when working for a long time), “appealing” (pleasant to use). Thus, we defined three additional research hypotheses:

RH3: The complementary color schemes are preferred for performing the counting (easy to count) to the black and/or gray-opposite color schemes (for both environments, monitor and paper).

RH4: The complementary color schemes are more comfortable for the eyes than the black and/or gray-opposite color schemes.

RH5: The complementary color schemes are more appealing than black and/or gray-opposite color schemes.

2.1.2 Research Design

For evaluating these research hypotheses, we designed an experiment with two groups. We used a counter-balanced within participant design with repeated measures to reduce the amount of error from natural variance between participants and learning effects. Considering that not enough safety experts were available, we used a convenient sample [ER11] including 32 professionals of the University of Kaiserslautern performing abstracted and simplified tasks. Participants were randomly assigned to groups. Each participant had to independently perform a fundamental task: Counting the circles having a specific number of dots. Thus, we used three different orders (1, 2, and 3, i.e., critical MCSs order in a hazard), three hazard sizes (12 (small) and 324, 540 (medium)), and two color schemes (black and gray-opposite). The visualizations were presented using a interactive power point presentation. For avoiding learning effects with respect to color schema, we counterbalanced our design forming two groups being exposed to black then gray-opposite and vice versa. The number of dots on each circle was intentionally randomly changed to avoid any learning effects. The first part of the experiment consisted of:

1. *Agree on an informal consent* for voluntarily participating in the empirical evaluation. Theses slides informed the participant regarding the research purpose, procedures, confidentiality and anonymity regarding all collected. It also informed the participant about the use and the future analysis of the collected data.
2. *Perform a color deficiency test*. This test supports the later analysis of possible correlations between errors and color deficiency.
3. *Introducing the experimental tasks* including an example.
4. *Training*: Each participant was asked to count stepwise the circles with 1, 2, and 3 dots in different levels in a visualization having 12 circles (hazard size of 12 MCSs). We used one visualization containing circles with black dots and one visualization containing circles with gray-opposite dots. One half of the participants saw the black dots first and the other half saw the gray-opposite dots first.
5. *Experimental treatment I*: Each participant performed the following tasks independently.
   a. Counting stepwise circles with 1, 2, and 3, black and gray-opposite dots in the red, yellow, green regions (different safety levels) in a problem size of 324 circles (hazard size of 324 MCSs).
   b. Counting stepwise circles with 1, 2, and 3, black and gray-opposite dots in the red, yellow, green regions (different safety levels) in a problem size of 540 circles (hazard size of 540 MCSs).

   In both cases, the number of dots was randomly distributed over different sizes and it was in the range of [1-13] dots. The groups of circles including black and gray-opposite dots were presented in different orders.
6. *Demographic questionnaire:* their age, gender, profession, and highest degree were asked for to allow finding any correlations.

The second part of the experiment focused on eliciting the user preferences. It included:

7. *Experimental treatment II:* Each participant performed the following tasks independently:
   a. *Monitor-based:*

      Seven images---one for each color scheme---were presented on 3 slides, one for a problem size of 12, 324, and 540 circles, respectively. The participants were asked to rank these images regarding their preference for solving the tasks introduced in the experimental treatment I. They should use a 7-point Likert scale (1: you like most to 7: you like least).
   b. *Paper-based:*

Seven images with the largest problem size (540) were presented on separate printed pages to the participants in random order. The participants were asked to evaluate whether the visualizations are "easy to use" (i.e., easy to count), "appealing", and "comfortable for the eyes". The participants should use a 5-point Likert scale (1: very good, 5: very bad) for each item.

After completing these tasks, some participants gave their feedback regarding the visualization and the color used independently. We used open questions for this purpose and wrote down all responses. During the whole experiment, the experimenter and an observer were responsible to taking notes of the participants' comments and questions. A detailed description of this research design is available in [AZZD*12].

2.2 Results

The empirical evaluation was performed as planned in February 2011 at the University of Kaiserslautern.

2.2.1 Sample

The empirical evaluation was conducted in February 2011 including 32 participants (5 females). They were in the age of 21 to 59. Nine participants have color deficiency view. The highest degree achieved by the participants was: B.Sc. (4), M.Sc. (23), Ph.D. (2), other (3). The participants worked in visualization (34.4%), software engineering---having safety knowledge---(15.6%), other computer science (31%), and other disciplines (19.0%). This data was collected after the execution of the experiment for each participant by the last power-point slide.

2.2.2 Data Analysis Procedure

In case, the data is normally distributed, we computed the sample mean $\bar{x}$ and the standard deviation σ. In case, the data is non-normally distributed, the sample median $\tilde{x}$ and the quartiles *Q1 25%,* and *Q3 75%* were computed. We gathered the individual results, {{1 dot, 2 dots, 3 dots}, {black, gray-opposite}, {324 circles, 540 circles}} and used these values to compute $\tilde{x}$ and the quartiles (we transformed all data to get a normal distribution using the functions: log10, sqrt, Reciprocal, arcsin, arctan, sqrtsqrt, but the data distribution did not change to a normal distribution, $p < 0.0001$, for all tests, an example for problem size 324, 1 dot is shown in Figure 4). The time format used in this paper is (*minutes:seconds*). The data analysis was performed using SPSS 17, G*Power, gnumeric, and minitab.

Time: The time was automatically measured by clicking a next button. The time was always measured for all three colors (red, yellow, green; safety levels) together.

Accuracy: We first computed the sum of all results over all colors (red, yellow, green) for each participant. This is done, because the time was automatically measured by clicking a "next" button in each slide that contains these three colors (safety levels: (unacceptable) red, (tolerable) yellow, and (negligible) green). Then, we used this value to compute the sample $\bar{x}$ $(or\ \tilde{x}$)), σ (or *quartiles*).

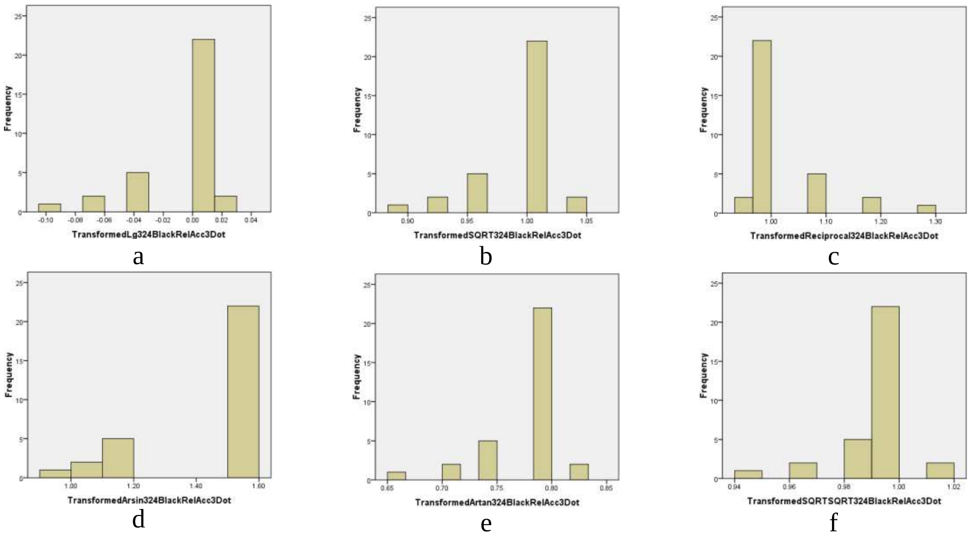


Figure 4. The data transformation tests for the problem size 324 of black 1 dot.
a) log10 b) sqrt c) reciprocal d) arcsin e) arctan f) sqrtsqrt

For the total results, {{black, gray-opposite}, {324 circles, 540 circles}}, we summed (tendency) the individual results of each participant over the dots (1 dot, 2 dots, 3 dots). We did not use any absolute values for the tendency to find out if the participants tend to overlook (negative results) or to over-count (positive results) the number of circles with a specific number of dots. The tendency equation is shown in Eq.1. In addition, we computed the relative accuracy of the results to assess the quality of our domino visualization representation and to determine which color scheme provides the most accurate results. The relative accuracy equation is shown in Eq.2. This equation is interpreted as followed: if the relative accuracy is 1 then the quality is good and if the relative accuracy is 0 then the quality is bad. Further, we measured the correlation between each of color deficiency, age, and gender, and the relative accuracy.

$$Tendency = \sum \#found - \#correct \quad \text{Eq.1}$$

$$RelativeAccuracy = {}^{\#found}/_{\#correct} \quad \text{Eq.2}$$

2.2.3 Training

In the training, we used the smallest problem size (12 circles). There were 32 participants, from which only four made one error each. All errors were in the first slide due to counting the number of dots in the circles instead of counting the number of circles that contain the dots, and due to one participant missing a circle.

Time: We analyzed the time at two abstraction levels: a detailed time “per dot number, i.e., 1, 2, 3 dots”, and an overview “per color scheme, i.e., black, gray-opposite”. The results of the descriptive analysis (Figure 5) show:

A. The participants took less time for the black scheme than for the gray-opposite scheme for performing the tasks with 1 dot, took equal time for 2 dots, and took more time for 3 dots.
B. The participants took overall less time for the black scheme than for the gray-opposite scheme for performing the tasks.

For determining if the participants using the black color scheme perform significantly faster than the participants using the gray-opposite scheme, we used the *Sign* Test. We selected this statistical test because the time variables are asymmetric (Figure 6) and not normally distributed ($N < 50$; Shapiro-Wilk normality test ($\alpha = 0.05$): $p < 0.000$ for all dot numbers and its total in the black color scheme and $p < 0.001$ for 1 dot, 2 dots, and 3 dots in the gray-opposite, and $p < 0.05$ for its total).

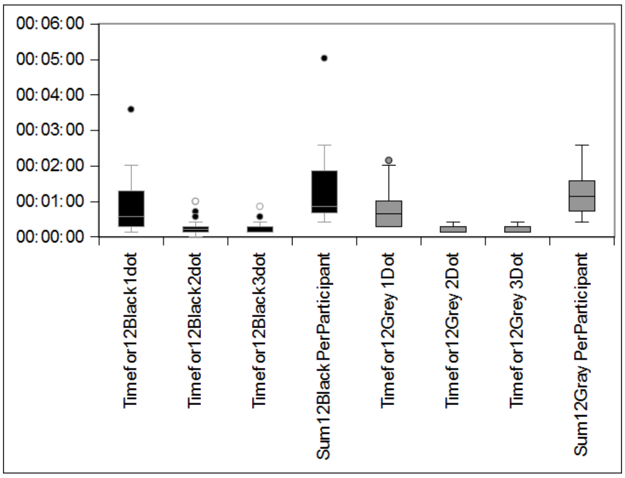
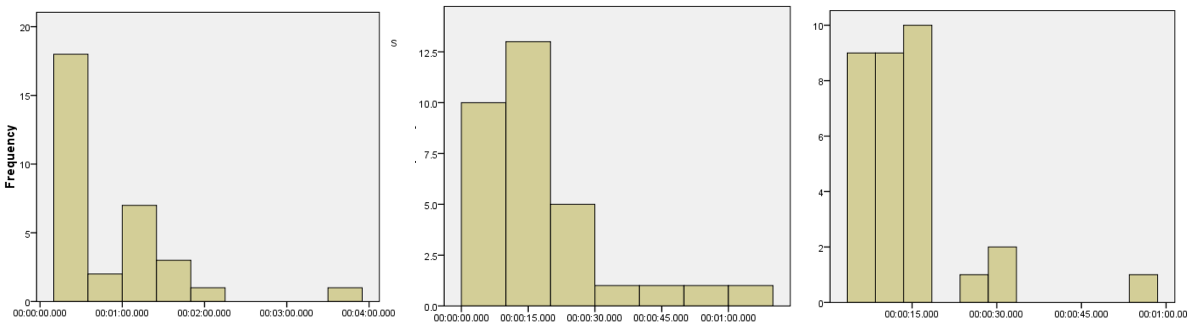


Figure 5. Box plots of the descriptive analysis of the training time.



Black 1 dot. Black 2 dots. Black 3 dots.

Figure 6. Examples of the data distribution of the training time black color scheme.

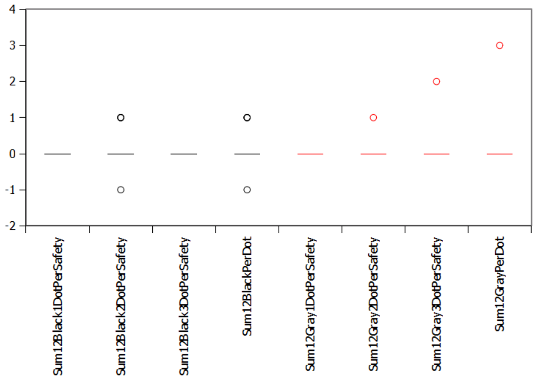


Figure 7. Box plot of the training tendency.

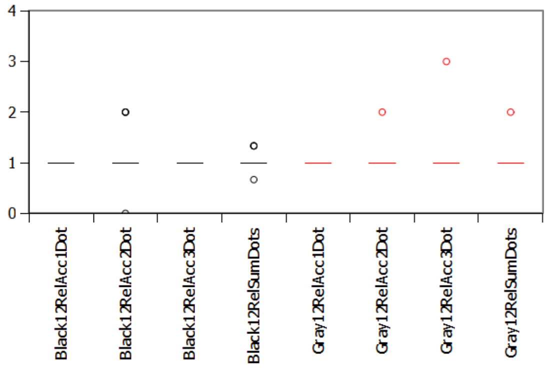


Figure 8. Box plot of the training relative accuracy.

The results show no statistical significant difference ($p > 0.05$, significance 2-tailed = 1.0 (1 dot), 0.596 (2 dots), 0.584 (3 dots), 1.0 (total).

Accuracy: We analyzed the accuracy at two abstraction levels: a detailed accuracy “per dot number, i.e., 1, 2, 3 dots”, and an overview accuracy “per color scheme, i.e., black, gray-opposite”. From the descriptive analysis we conclude:

a. *Tendency*: Both the black and the gray-opposite schemes have led the participants to over-counting, with all $Q1, \tilde{x}, Q3$ equal to 0 (Figure 7). For the tendency, no statistical significance test is needed, because the over-counting is a result of the outliers.

b. *Relative accuracy*: The median of both the black and the gray-opposite schemes were 1, this means the quality of both schemes is good in the training (Figure 8).

Because all relative accuracy data was asymmetric and not normally distributed (N < 50: Shapiro-Wilk normality test ($\alpha = 0.05$): $p < 0.0001$, for all dots and the sum, both black and gray-opposite; for black 1 dot and 3 dots, and gray-opposite 1 dot, all results are constant 1), the sign test was used as significance test between the color schemes to find whether the participants using the black scheme were significantly more accurate than the participants using the gray-opposite scheme. The results show no statistical significant difference ($p > 0.05$, significance 2-tailed = 1 (1 dot), 1 (2 dots), - (3 dots), 1 (total)). We defined the statistical null hypotheses shown in Eq.3. Overall, we conclude that the participants understood the assigned tasks and performed them without problems and that the training was successfully completed by all participants.

$$H_0: \mu_{RelAcc\ Black} = \mu_{RelAcc\ Gray-opposite} \qquad \text{Eq.3}$$

2.2.4 Experimental Treatment I

Time: In general, the participants took more time performing the tasks when using the gray-opposite scheme than when using the black scheme regardless the problem size and dot number (Table 1 and Figure 13). Since the time variables are asymmetric (Figure 14) and cannot be assumed to be normally distributed ($p < 0.0001$, Shapiro-Wilk normality test, $\alpha = 0.05$, $p < 0.001$, all p values are 0, except black, 540, 1 dot, $p = 0.019$, and gray-opposite, 540, 3 dots, $p = 0.021$) and they are asymmetric, we performed the sign test for identifying, if there is a significant difference in the time variables because of the dot number (Table 2).

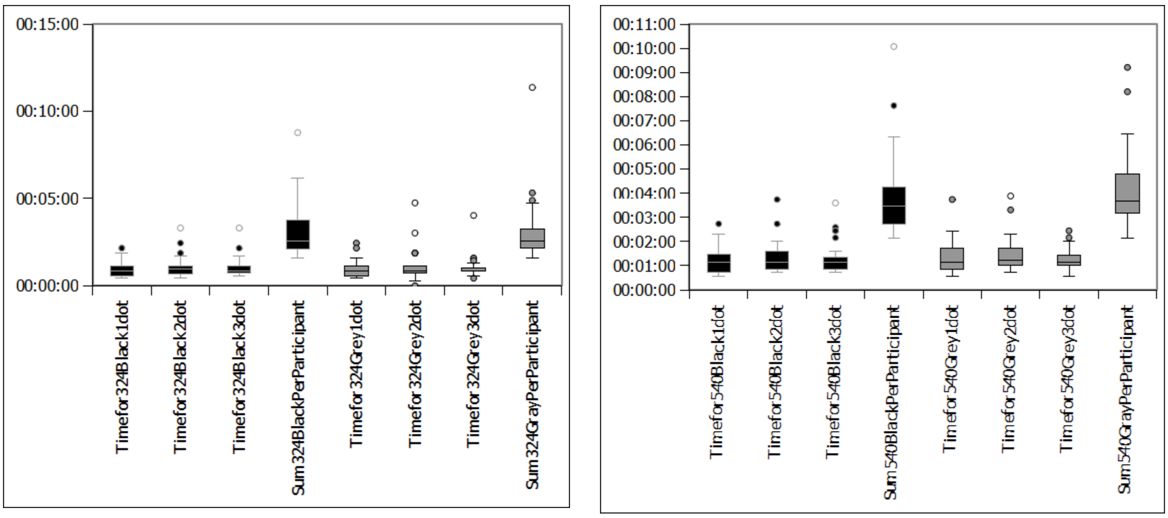
Table 1. Descriptive analysis for the testing time for problem sizes 324 and 540 in different abstractions

Problem size, dot number	Black (Q1, $\tilde{x}$, Q3)	Gray-Opposite (Q1, $\tilde{x}$, Q3)
324 , 1 dot	00:37, 00:48.5, 01:11.7	00:34, 00:50.5, 01:07.5
324 , 2 dot	00:38.7, 00:58.5, 01:11.5	00:39.7, 00:54, 01:06.7
324 , 3 dot	00:41, 00:52.5, 01:11	00:47, 00:54.5, 01:07.7
Median of sum (total)	02:03, **02:33**, 03:50	02:06.5, **02:38**, 03:22
540, 1 dot	00:41, 01:07, 01:34.5	00:54, 01:09.5, 01:40
540, 2 dot	00:52, 01:10, 01:33.5	01:01, 01:13.5, 01:42.7
540, 3 dot	00:55, 01:09.5, 01:25.7	01:01, 01:10.5, 01:23.7
Median of sum (total)	02:44.5, **03:27**, 04:24.5	03:07.7, **03:39.5**, 04:53.7

We define the statistical null hypotheses shown in Eq.4:

$$H_0: \mu_{T\ Black} = \mu_{T\ Gray-opposite} \qquad T \in \{time_{1dot}, time_{2dots}, time_{3dots}\} \qquad \text{Eq.4}$$

The results show that there are no statistical significant differences in the time for both schemes independently of the dot number and of the problem size. However, the “practical significance“ (also called “*effect size*”) $ES = Z/\sqrt{n}$ of the results are large for the total (ES = 0.217) for the problem size 540 in favor of black (taken from [Coh69]: p.142, the effect size is small, when ES **<** 0.05, medium, when $ES \in$ [0.05-0.15[, and large, when ES > 0.15).



a B

Figure 13. Box plots for the testing time of experimental treatment I for the problem sizes: a) 324 b) 540 circles.

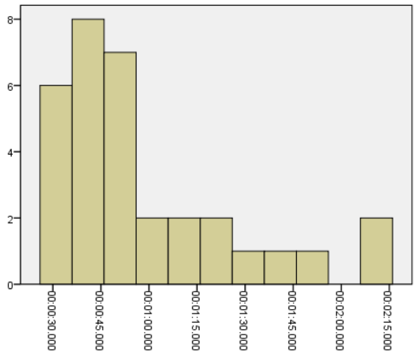


Figure 14. Example of the data distribution of the testing's time for the problem size 324, black scheme, and 1 dot.

Table 2. Sign significance test for the testing time: p value ($\alpha = 0.05$)

	324	540
1 dot	0.28	0.21
2 dot	0.28	0.11
3 dot	0.59	1
Median of sum (total)	0.86	0.21

Accuracy: Two different abstractions were measured to identify special cases: a detailed accuracy "per dot number, i.e., 1, 2, 3 dots", and an overview "per color scheme, i.e., black, gray-opposite". The descriptive analysis of the accuracy shows:

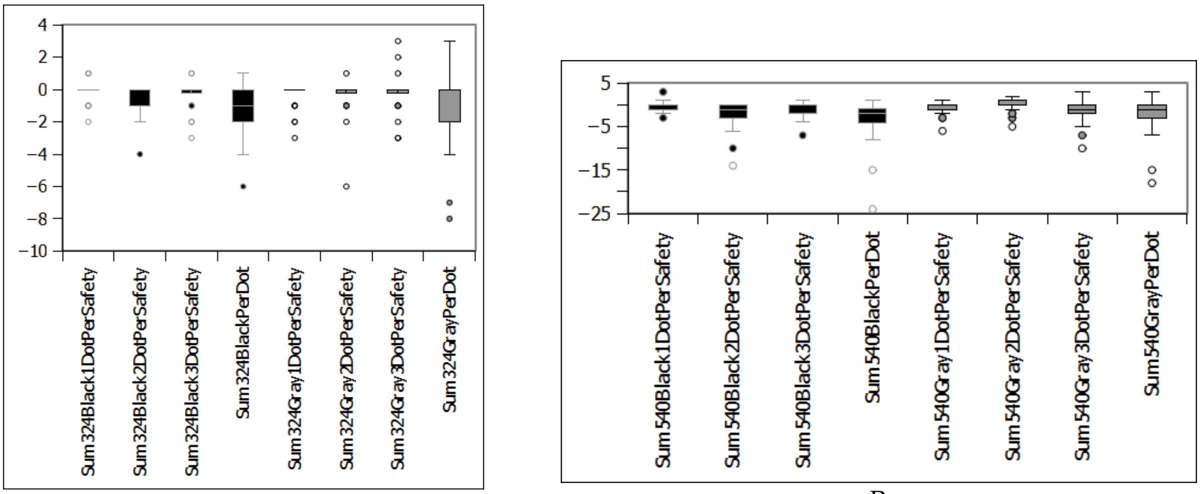
a. *Tendency*: Both the black and the gray-opposite schemes caused the participants to over-look circles (negative values) independently of the problem size (324, 540). Table 3 and Figure 15 show this result. This result was expected, because larger data sets cause more distraction, which leads to less accuracy. The correct number of dots was in the range [1-13] and the dots were distributed randomly over the visualization.
b. *Relative accuracy*: From Table 4, we find that the median of the total relative accuracy for the black was less than for the gray-opposite scheme for both problem sizes. This means that the achieved relative accuracy for the gray-opposite scheme is better than for the black scheme. Figure 16 shows the results for the problem size 324 and 540.

Table 3. Descriptive analysis for the testing's tendency in different abstractions

	Black (Q1, $\tilde{x}$, Q3)	Gray-opposite (Q1, $\tilde{x}$, Q3)
324, 1 dot	0, 0, 0	0, 0, 0
324, 2 dots	-1, 0, 0	-0.75, 0, 0
324, 3 dots	-0.75, 0, 0	-0.75, 0, 0
Total	-2, **-1**, 0	-2, **0**, 0
540, 1 dot	-1, 0, 0	-1, 0, 0
540, 2 dots	-3, -1, 0	0, 0, 1
540, 3 dots	-2, 0, 0	-2, -1, 0
Total	-4.75, **-2**, -0.25	-3, **-1**, 0

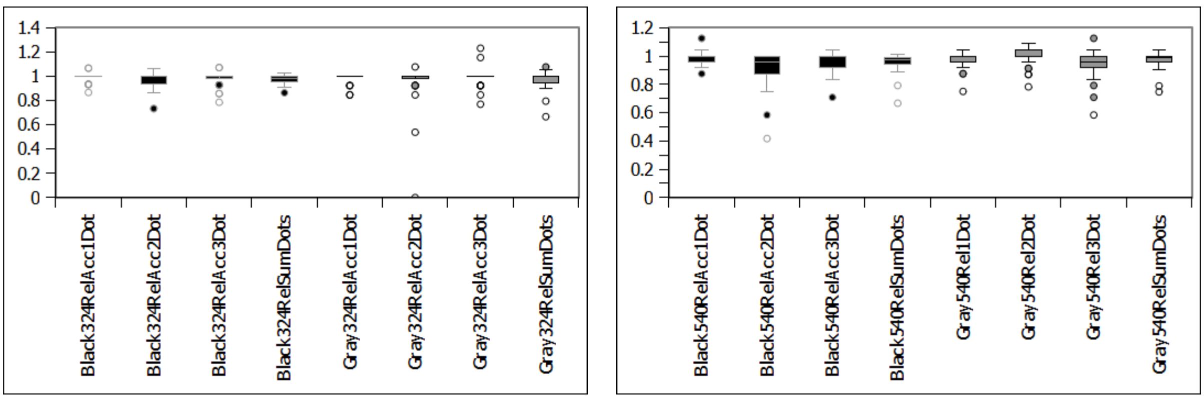
Table 4. Descriptive analysis for the testing’s relative accuracy in different abstractions

	Black (Q1, $\tilde{x}$, Q3)	Gray-opposite (Q1, $\tilde{x}$, Q3)
324, 1 dot	1, 1, 1	1, 1, 1
324, 2 dots	0.93, 1 ,1	0.94, 1, 1
324, 3 dots	0.95, 1, 1	1, 1, 1
Total	0.95, **0.98**, 1	0.95, **1**, 1
540, 1 dot	0.96, 1, 1	0.96, 1, 1
540, 2 dots	0.88, 0.96, 1	1, 1, 1.04
540, 3 dots	0.92, 1, 1	0.92, 0.96, 1
Total	0.93, **0.97**, 1	0.96, **0.99**, 1



a B

Figure 15. Tendency undercounting for testing data in both problem sizes: a) 324 b) 540 circles.



A B

Figure 16. Relative accuracy for the problem sizes: a) 324. b) 540 circles.

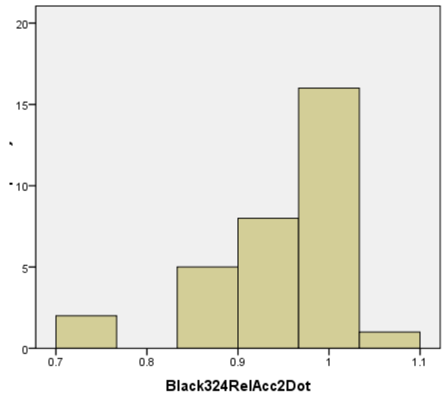


Figure 17: Example of the data distribution of the testing's relative accuracy for black 2 dots problem size 324.

Table 5. Sign significance test for the testing time: p value (α = 0.05)		
	324	540
1 dot	0.77	0.66
2 dot	0.28	0.000
3 dot	1	0.52
Median of sum (total)	1	0.06

In addition, we tested if there is a significant difference for the relative accuracy achieved by participants between the black color scheme and the gray-opposite scheme. We defined the null hypothesis as described by Eq.5. Because both black and gray-opposite data are asymmetric (Figure 17) and are not normally distributed (N < 50: Shapiro-Wilk normality test (α = 0.05): p < 0.0001, all p values are 0, except black, 324, median of sum (total): p = 0.003), the sign test was used. The results (Table 5) show no statistical significant difference for the problem size. Similarly, there was no statistical significant difference found for the problem size 540 for all but one case namely the 2 dots. However, the effect size is large (*ES* = 0.61) for the problem size 324 in favor of the black and large (*ES* = 0.32) for the problem size 540 in favor of the gray-opposite color scheme for the total result.

$$H_0: \mu_{RelativeAccuracy\ Black} = \mu_{RelativeAccuracy\ Gray-opposite} \qquad \text{Eq.5}$$

As the median of the relative accuracy for the gray-opposite scheme is closer to 1 than for the black scheme, we conclude that the participants made less errors for the gray-opposite compared to the black scheme. This is an indication that the gray-opposite leads to more accuracy for the larger problem size. However, from three of the comments written by the participants, we found an indication that the participants had difficulties while using the gray-opposite color scheme. They considered it to be harder and they preferred the black color scheme. The comments were:

- “The dark colors (like black) are better recognizable. Best would be a combination of a dark color (for the dots) and a bright cylinder. Like black over white or dark blue over white. The grey gray was difficult to see, especially over a bright color.” (“Die dunklen Farben (wie Schwwarz) sind besser zu erkennen,Am besten wäre die Kombination von einer dunklen Farbe (für die Punkte) mit einem helligen Zylinder. Wie Schwarz auf Weiß oder Dunkelblau auf Die Grau war schlecht zu erkennen besonders , wenn es auf heller Farbe war.Weiß.”)
- “I liked the color configurations, where dots between neighboring levels had different colors and where the contrast was high. I did not like gray as color for the dots, because the contrast to the cylinders was too low.” (“Mir haben die Farkkonfigurationen gut gefallenbei denen die Punkte zwischen benachbarten Ebenen verschiedene Farben hatten, bei denen der Kontrast der Farben hoch war. Grau fand ich als Farbe für die Punkte nicht so gut da der Kontrast mit den Zylindern zu gering.”)
- “changing dot color made the counting harder for me, especially if the dot color changed within the red/yellow/green area to stay in a single circle. The better the dots distinguish themselves from the gray on the bottom and the color of the cylinder, the better for me.”

From these comments, the accuracy and the time results, we conclude that these participants may have spent more effort (more time) on the gray-opposite scheme to focus on the tasks and in consequence they made fewer errors (higher accuracy).

Influencing factors: To find out whether the results of the relative accuracy were influenced by the demographic variables of our participants, we conducted the following correlation tests. Since the age is not normally distributed and the other variables are of type nominal, we used the Spearman correlation test. The results are shown in Table 6. The results indicate that there is no relation between gender or color deficiency and the relative accuracy. However, the age variable shows a small opposite relation to accuracy for the gray-opposite scheme for the problem size 324. This could be interpreted as weak vision (the higher the age the weaker the vision).

Table 6. Correlation and significance in different abstractions.* → $p < 0.05$, ** → $p < 0.01$, *** → $p < 0.001$

Problem size	Color scheme	Number of dots	Age	Gender	Color deficiency	Profession
324	Black	1 dot	-0.25	0.05	-0.25	0.24
		2 dot	-0.06	-0.04	-0.31	-0.226
		3 dots	-0.07	0.13	0.22	-0.001
		Total	-0.18	0.07	-0.18	-0.141
	Gray-opposite	1 dot	-0.04	0.07	-0.12	0.059
		2 dot	**-0.65****	0.12	-0.01	0.127
		3 dots	**-0.41***	0.18	-0.1	0.008
		Total	**-0.53****	0.06	0.03	-0.19
540	Black	1 dot	0.27	-0.08	-0.02	0.178
		2 dot	-0.07	0.03	-0.29	0.055
		3 dots	-0.25	0.09	0.03	0.045
		Total	0.004	0.00	-0.23	0.074
	Gray-opposite	1 dot	-0.15	-0.20	0.19	**0.395***
		2 dot	-0.06	0.12	-0.02	-0.054
		3 dots	-0.19	-0.04	0.01	0.095
		Total	-0.27	-0.02	0.05	0.088

2.2.5 Experimental Treatment II

In the experimental treatment II, we attempted to find out the user preference regarding the color scheme using a monitor and paper version. Both versions were used, because the colors are perceived differently using different environments.

Monitor-based preference: The illustrations were presented and the participants' answers were collected using PowerPoint slides. We included one slide per problem size (i.e., 12, 324, and 540). The images contained the seven schemes introduced in Section 0. An example is shown in Figure 18. The participants were asked to rank the images shown from "1: you like most" to "7: you like least" in the context of the tasks solved during the first treatment.

$$H_0: \mu_{easy\ to\ use\ \_Scheme} = 4 : \textit{Scheme} \text{ is each one of the seven schemes.} \quad \text{Eq.6}$$

Because our data is ordinal, not normally distributed (Shapiro-Wilk normality test α: 0.05, all $p < 0.01$, Table 7) and asymmetric (Figure 19), we used the sign significance test (statistical null hypotheses: Eq.6) to find out, which scheme was preferred by the participants for performing experimental treatment I. The test value is the median "4" (Table 8). The best color schemes with a statistical significant difference from the test value are:

- 12: the black ($\tilde{x}$ = 1) and the invert (3) color schemes
- 324: the black (2) and the invert (2) color schemes
- 540: the black (1), the invert (3), and the complementary-opposite (3) color schemes.

The three schemes that are best using the Friedman test (Figure 20 d, e, f):

- 12: black (1.28), invert (3.31), complementary-same (3.8). Additionally, the black color scheme is the best with a statistical significant difference from all color schemes (χ^2 (85.861), df (6), $p < 0.000$).
- 324: black (2.16), invert (2.39), complementary-same (3.36). Additionally, the black color scheme is the best with a statistical significant difference from the gray, gray-opposite, gray-same, and complementary-opposite color schemes (χ^2 (82.776), df (6), $p < 0.000$).
- 540: black (2.08), invert (3.05), complementary-opposite scheme (3.13). Additionally, the black colors scheme is the best with a statistical significance difference from the gray, gray-opposite, and gray-same color schemes (χ^2 (86.798), df (6), $p < 0.000$).

To find out, whether the best first two schemes are significantly different, we used the sign test. For the problem sizes 12 and 540: the black had a statistical significant difference from the invert scheme.

Table 7. Normality distribution test for monitor-based: p value of the Shapiro-Wilk normality test (α = 0.05)

	Black	Gray-opposite	Gray	Gray-same	Invert	Complementary-same	Complementary-opposite
12	0.000	0.006	0.006	0.006	0.004	0.027	0.000
324	0.000	0.000	0.004	0.069	0.000	0.025	0.035
540	0.000	0.000	0.000	0.001	0.032	0.039	0.006

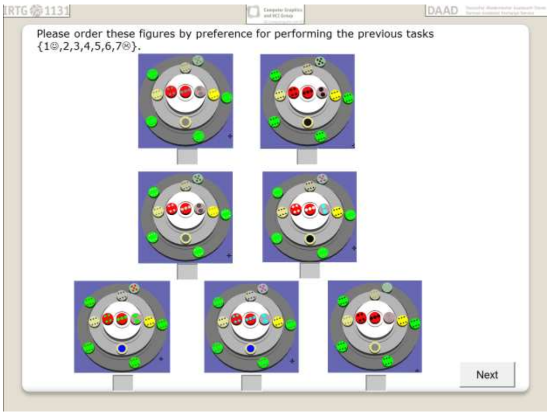


Figure 18. An example of the monitor-based preference slide for the problem size 12.

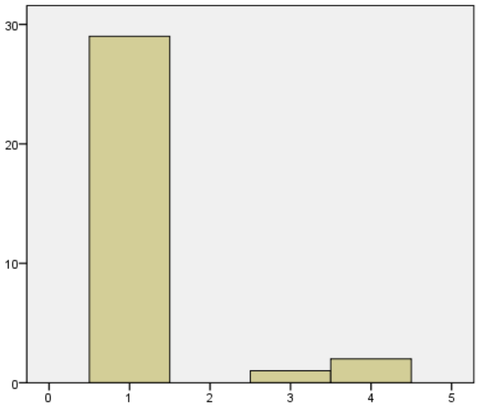


Figure 19. An example of the monitor-based preference result for the problem size 12, black.

Table 8. Descriptive analysis (Q1,$\tilde{x}$, and Q3) for monitor-based test; all schemes, all sizes.

Sign test with test value = 4. * → p < 0.05, ** → p < 0.01, *** → p < 0.001.

	Black	Gray	Gray-Opposite	Gray-Same	Invert	Complementary-Same	Complementary-Opposite
12	1,**1**,1***	2.25, 4, 5.75	4, 5, 6**	4, 5, 7**	2, **3**, 4.75*	2, 3.5, 5.75	4, 6, 7***
324	1, **2**, 3***	4, 6, 6***	5, 6, 7***	3, 4.5, 5	1, **2**, 2.75***	2, 3, 5	4, 4, 6
540	1, **1**, 3***	5, 6, 7***	4, 6, 6***	5, 5, 7***	2, **3**, 4**	2, 3, 5	2, **3**, 4**

Paper-based preference: Seven illustrations were created for the problem size 540 using different color schemes. Each illustration was printed as exemplified in Figure 21. All illustrations were presented in random order to the participants. We asked the participants to rank them from "1: best to 5: worst" with repetition allowed for their ranking, considering the following attributes: "easy to use (count)", "comfortable for the eyes", and "appealing". Because our data is ordinal, not normally distributed (Shapiro-Wilk normality test α: 0.05, all $p \leq 0.01$) and not symmetric (Figure 22), we used the sign significance test (statistical null hypotheses: Eq.7).

$$H_0: \mu_{\text{criteria}\_Scheme} = 3 \qquad \text{Eq.7}$$

The test value is the median "3" (Table 9). The best color schemes with a statistical significant difference from the test value are:

- "Easy to use": black (1) and complement-same (2)
- "Comfortable": black (2) and gray-opposite (2)
- "Appealing": There was no color scheme having a smaller median than the test value with statistical significant difference from the test value.

Using the Friedman test regarding the three criteria, the three best preference ranks were (Figure 23 d, e, f):

- “Easy to count/ useful”: black (rank: 2.02), complementary-same (3.2), and gray-opposite (3.48), $\chi^2(100.253)$, df (6), $p < 0.000$.
- “Comfortable for the eyes”: black (2.91), gray-opposite (3.27), and complementary-same (3.98), $\chi^2(25.680)$, df (6), $p < 0.000$.
- “Appealing”: complementary-opposite (3.47), invert (3.56), and complementary-same (3.61), $\chi^2(19.118)$, df (6), $p < 0.005$.

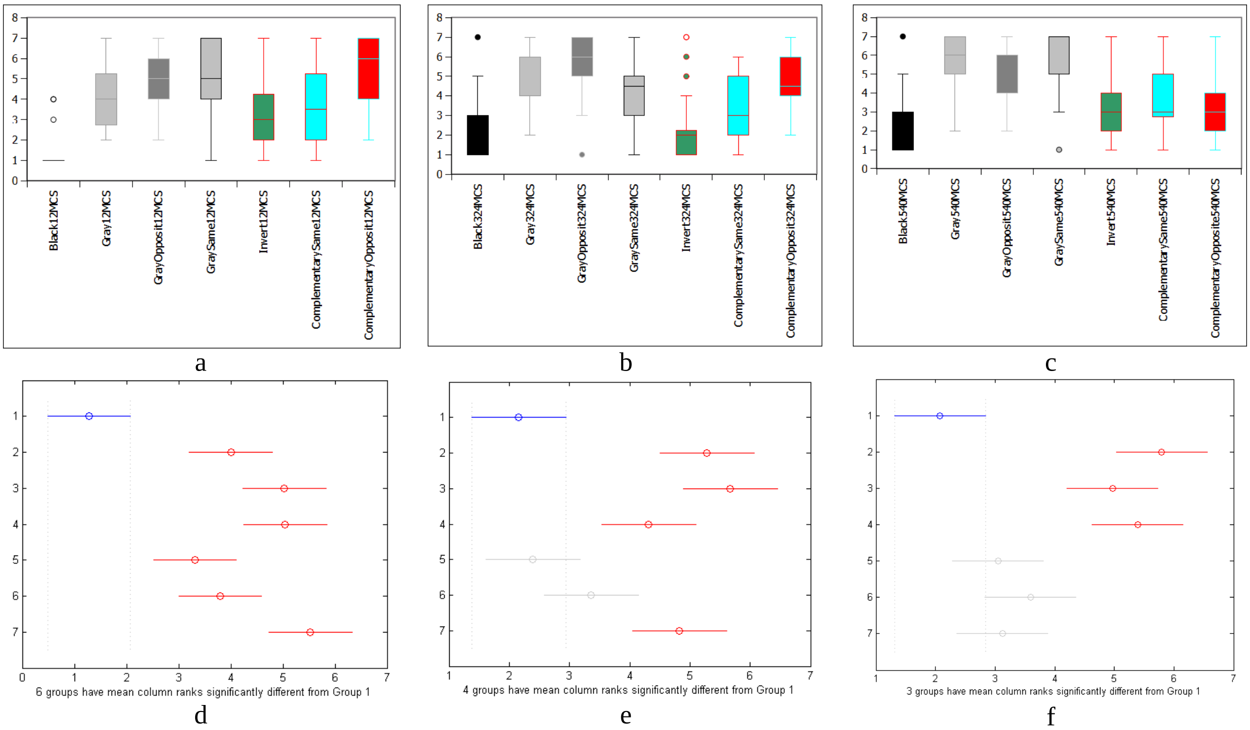


Figure 20. The descriptive results of the monitor-based preference for the problem sizes: a) 12 b) 324 c) 540

Friedman multi-comparison test for the problem sizes: d) 12, e) 324, and f) 540 circles.

The sequence of the color schemes from top to bottom/left to right is: Black, Gray, Gray-opposite, Gray-Same, Invert, Complementary-same, and Complementary-opposite.

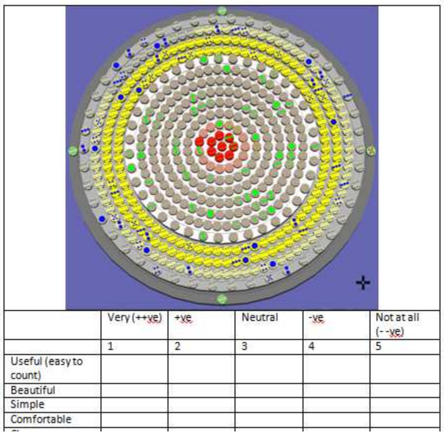


Figure 21. An example of the paper-based preference for the problem size 540 with the invert scheme.

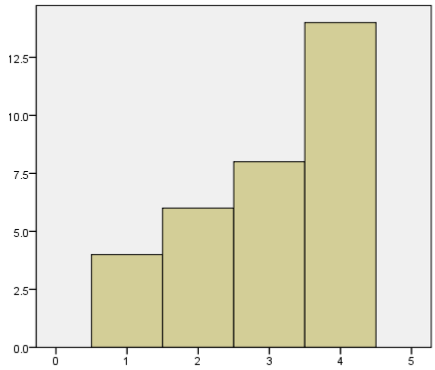


Figure 22. An example of the paper-based preference for the invert scheme regarding “comfortable for the eyes”.

Table 9. Descriptive analysis for the paper-based (Q1, $\tilde{x}$, and Q3) for all schemes and all criteria. Sign test with test value = 3. * → p < 0.05, ** → p < 0.01, *** → p < 0.001.

	Black	Gray	Gray-Opposite	Gray-Same	Invert	Complementary-Same	Complementary-pposite
Easy to use	1, **1**, 2***	4, 4, 5***	2, 2.5, 3	4, 4, 5***	2, 3, 3	2, **2**, 3**	2, 3, 3*
Comfortable	1.25, **2**, 3**	2, 3, 4	2, **2**, 3**	2.25, 4, 4	2, 3, 4	2, 3, 4	2, 3, 4
Appealing	2, **3**, 3	2.25, **3**, 4	3, **3**, 3	3, 4, 4*	2, **3**, 4	2, **3**, 3.75	2, **3**, 4

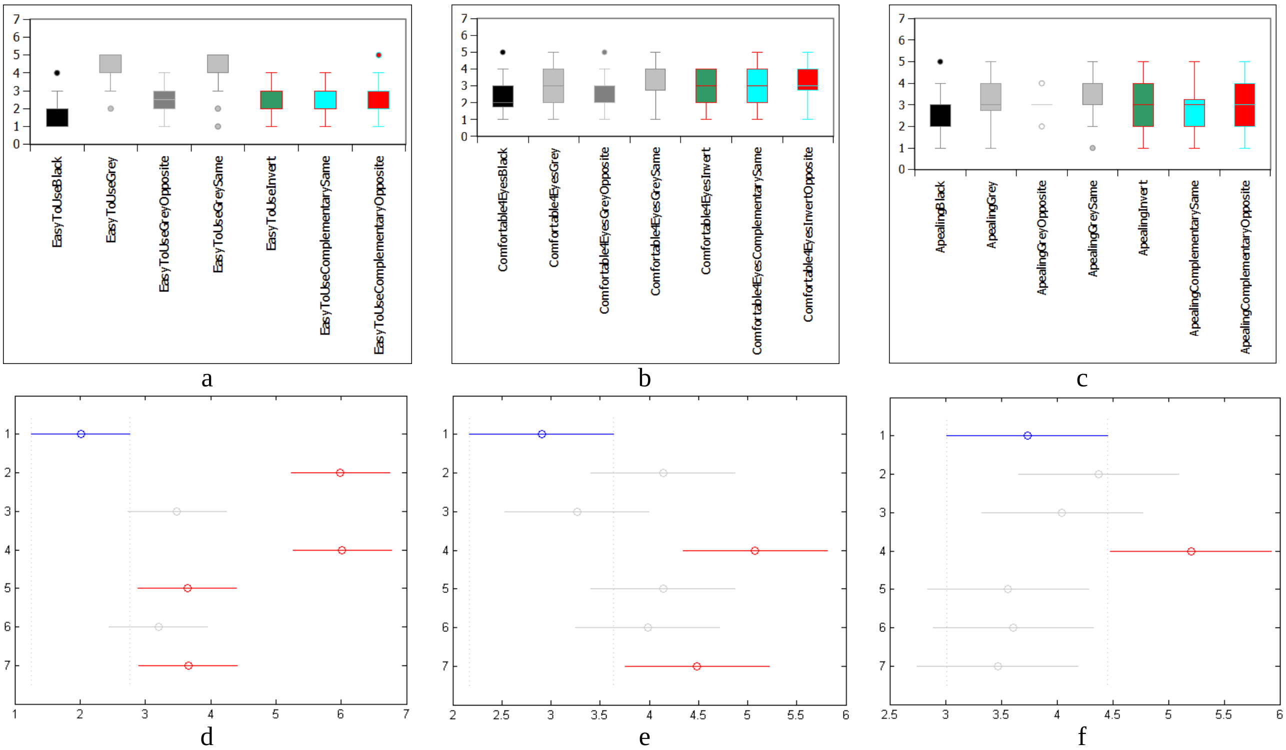


Figure 23. The descriptive results of the paper-based preference regarding for the problem sizes:
a) easy to use b) comfortable c) appealing.
Friedman test multi-comparison for: d) easy to use, e) comfortable, and f) appealing.
The sequence of the color schemes from top to bottom/left to right is: Black, Gray, Gray-opposite, Gray-Same, Invert, Complementary-same, and Complementary-opposite.

To find out whether the best two schemes regarding “easy to count” and “comfortable for the eyes” are significantly different, we used the sign test. For “easy to count”, the black scheme was significantly different from the invert scheme. For “comfortable for the eyes”, there was no significant difference between the black and the gray-opposite schemes. This indicates that the black scheme seems to be the best choice for coloring the dots.

2.2.6 Participants Feedback

We received 14 comments from the participants. In summary, all participants except one preferred the less colorful color schemes with high contrast to the circles color: e.g., a participant stated: “I prefer the dots to be black for contrast.”

Two participants commented on the design of the safety-domino. The first comment supports our choice of the safety-domino: “If I remember correctly, icons with a specific number were always oriented in the same way. I liked that, it made finding identical numbers much easier.” The second comment supports our overall design, namely the safety-domino on saturated colors: “Color and value (lightness) are pre-attentive features. This makes it very comp(l)icated to count the dots for example in the big yellow field (the eye follows more the change of value than the number of dots)." While formulated in a negative way, it supports our design: we show as well the failure probability (main focus) through color and saturation as the size of the circles through the safety-domino (secondary focus).

2.2.7 Threats to Validity

It is important to remark that the experiment was conducted without deviations from the design. Moreover, the statistical analysis shows that the observed effect in the dependent variables (e.g., time and accuracy) can be attributed to the experimental treatment. This supports the conclusion validity and the internal validity. Nevertheless, we observed that the participants became tired during experimental treatment I. This can be avoided in the design of further experiments by splitting the assigned tasks into several laboratory sessions.

Construct and external validity are closely related in this experiment. Choosing a fundamental counting task instead of the safety engineering tasks is a threat to the construct validity. Moreover, we used a convenient sample instead of safety experts, which reduces the generalizability and thus is a threat to external validity. However, the observed demographic variables did not influence the dependent variables. Thus, the results seem to be generalizable to a large group of users. Further, due to the lack of safety experts available for this experiment, the tasks and the sample allow an assessment of the influence of the independent variables on the dependent variables. Several replications including larger and different samples are needed to come to more general and more conclusive empirical results.

3. DISCUSSION OF THE EXPERIMENTAL RESULTS

The differences between the black and the gray-opposite color scheme in time and in accuracy of the experimental treatment I are not statistically significant (Section 2.2.4). However, there is a strong indication of the advantages of the black color scheme. Although the results in accuracy were slightly better for gray-opposite, the participants took more time using this scheme (the effect size is large in the execution time for black for the problem size 540 whereas it is large in the accuracy for the gray-opposite). Based on the participants' remarks, we assume that they perceived the gray-opposite color scheme as being more demanding and therefore the participants took more time during the treatment and this caused the better accuracy.

The results of the experimental treatment are generalizable, as the demographic variables show no correlation to the color schemes. Especially, the fact that the results are independent of the profession allows generalizing the results to a large public.

The black color scheme was the best color scheme for the monitor based test of the experimental treatment II (Section 2.2.5). It was significantly better than the test value, and it was significantly better rated than the gray-opposite color scheme. For the paper based test of this treatment, black was ranked best for two out of three categories. Only for "Appealing", the invert and complementary color schemes were preferred. In the category "easy to count/useful" the black color scheme was significantly better than the gray-opposite one. Altogether, this indicates that the participants were most comfortable with the black color scheme and that it appeared most suitable for them for the task used in the experimental treatment I.

Regarding the visualization, our choices were confirmed. The operationalization and the system size should be increased in forthcoming evaluations to see, whether or not there is a significant difference between black and gray-opposite.

4. SUMMARY AND FUTURE WORK

Contrary to the current state of the art, safety tools need not only to represent the minimal cut sets' (MCS) failure probability in a way to ease prioritizing during analyses, but also to represent their order (the number of basic events they contain). In this paper, we presented the design decisions of the safety-domino and the detailed evaluation results, extending our previous report [AZZD*12]. The safety-domino is a visualization representing the MCSs order in addition of the minimal cut sets' failure probability, This design is also applicable to many other representations.

After having reviewed the literature and the techniques related to representing numbers, the domino representation for a minimal cut sets' order was selected and adapted to fit the safety requirements (small order: more important). The safety-domino adapts the size of the standard domino dots to directly provide the importance impression. Further, it reorients the "2" dots such that it is easily distinguishable from the "3" dots. Additionally, we designed several color schemes based on current research about perception. We evaluated the color schemes with 32 participants to find the best one that enables the users to get an overview over the safety of a system. The results of this evaluation show a tendency that the gray-opposite color scheme leads to higher accuracy but the participants also spent more time compared to the black color scheme. Moreover, while observing the participants and from their feedback, we noticed that they needed to concentrate more on the gray-opposite dots to get the correct answers, which explains these results. We propose to interpret the results of any evaluation considering the feedback of the participants. For this reason, we propose to use the largest contrast (black dots) for this visualization.

In the future, additional experiments could be used to find if there is a significant difference between the black and the gray-opposite color scheme. However, the difference between them might be too small leading to similar results compared to this evaluation.

ACKNOWLEDGEMENT

This work is partially supported by the DAAD, the IRTG 1131 sponsored by the DFG, and the BMBF VierforES I project. We thank all our colleges for participating in the evaluation and providing feedback.

REFERENCES

[AZKS*11] AL-ZOKARI Y. I., KHAN T., SCHNEIDER D., ZECKZER D., HAGEN H.: CakES: Cake Metaphor for Analyzing Safety Issues of Embedded Systems. In Scientific Visualization: Interactions, Features, Metaphors, Hagen H., (Ed.), vol. 2 of Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2011, pp. 1–16.

[AZSZ*11] AL-ZOKARI Y. I., SCHNEIDER D., ZECKZER D., GUZMAN L., LIVNAT Y., HAGEN H.: Enhanced CakES representing Safety Analysis results of Embedded Systems. In Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS) (Szczecin, Poland, 2011), pp. 793–800.

[AZYZ*11] AL-ZOKARI Y. I., YANG Y., ZECKZER D., DANNENMANN P., HAGEN H.: Towards Advanced Visualization, and Interaction Techniques for Fault Tree Analyses, Comparing existing methods and tools, 2011. Abstract submitted to PSAM/ESSAREL 2012.

[AZZD*12] AL-ZOKARI Y. I., ZECKZER D., DANNENMANN P., GUZMAN L., HAGEN H.: Safety-domino: A graphical metaphor for supporting minimal cut sets analysis. In IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing (2012), pp. 63–74.

[AZZH11] AL-ZOKARI Y. I., ZECKZER D., HAGEN H.: Safety- Domino Representing Criticality of Embedded Systems. In EuroVis11, Eurographics /IEEE Symposium on Visualization 2011, poster proceedings (2011), p. 21.

[Bec66] BECK J.: Effect of orientation and of shape similarity on perceptual grouping, Perception & Psychophysics, 1966.

[BV07] BOZZANO M., VILLAFIORITA A.: The FSAP/NuSMVSA Safety Analysis Platform. International Journal on Software Tools for Technology Transfer (STTT) 9, 1 (2007), 5–24.

[BV10] BOZZANO M., VILLAFIORITA A.: Design and Safety Assessment of Critical Systems. CRC Press (Taylor and Francis), an Auerbach Book, 2010.

[Coh69] COHEN, J.: Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1969.

[ER11] EXPERIMENT-RESOURCES: Convenience sampling applied to research, 2011. http://www.experiment-resources.com/convenience-sampling.html; Online; accessed 7-Feb-2011.

[Hun00] HUNTER N.: Art of Floral Design 2e. ART OF FLORAL DESIGN. Delmar, 2000.

[JI60] ITTEN, J.: The Art of Color. New York: Van Nostrand Reinhold Company, 1960.

[KZ96] KARA-ZAITRI C.: An improved minimal cut set algorithm, vol. 13. MCB UP Ltd, 1996.

[PRL12] PENTAK S., ROTH R., and LAUER D.: Design Basics: 2D and 3D. Cengage Learning, 2012 .

[Sav] SAVIVE PTY LTD: Safety-Critical. http://www.savive.com/inform/safetycritical.html; Online; accessed 31-Mar-2011.

[VGRH81] VESELY W. E., GOLDBERG F. F., ROBERTS N. H., HAASL D. F.: NUREG0492. 1981.

[War04] WARE C.: Information Visualization: Perception for Design / 2nd Edition. Morgan Kaufmann, 2004.

[wiki12] http://en.wikipedia.org/wiki/Color_wheel