

# Supplemental Information on Comparison of Radial and Linear Charts for Visualizing Daily Patterns

This supplemental material contains detailed information about the study procedure, the participants, the analysis process including open coding, statistical analysis of the dependent variables, further exploratory analyses, and all results.

## Contents

Study Procedure .....	3
Participants .....	12
Analysis .....	15
Task 1: Untargeted Analysis .....	15
Open Coding .....	15
Number of Observations.....	16
Correlation between Number of Observations and Demographics .....	17
Correlation between Number of Observations and Trial Order.....	20
Qualitative Analysis of Responses without Observations.....	21
Comparisons, Salient Features, and Time References .....	25
Exploratory Analysis of Time Periods Mentioned .....	28
Task 2: Locate Time .....	34
Accuracy.....	34
Exploratory Analysis of Error Cases.....	34
Completion Time .....	37
Correlation between Completion Time and Demographics.....	38

Correlation between Number of Reported Observations and Task Performance.....	40
Task 3: Read Value .....	41
Accuracy.....	41
Exploratory Analysis of Error Cases.....	43
Completion Time .....	50
Correlation between Number of Reported Observations and Task Performance.....	51
Task 4: Locate Maximum .....	52
Accuracy.....	52
Exploratory Analysis of Error Cases.....	53
Completion Time .....	55
Correlation between Number of Reported Observations and Task Performance.....	56
Task 5: Compare A.M./P.M. Interval Values .....	58
Accuracy.....	58
Qualitative Analysis of Error Cases .....	59
Completion Time .....	61
Correlation between Number of Reported Observations and Task Performance.....	62
Task 6: Subjective Ratings .....	64
Subjective Ratings.....	64
Correlation between Demographics and Preferences.....	66
Correlation between Number of Reported Observations and Subjective Ratings.....	71
Optional Text Comments .....	72

# Study Procedure

Start screen shown on a 27" monitor:

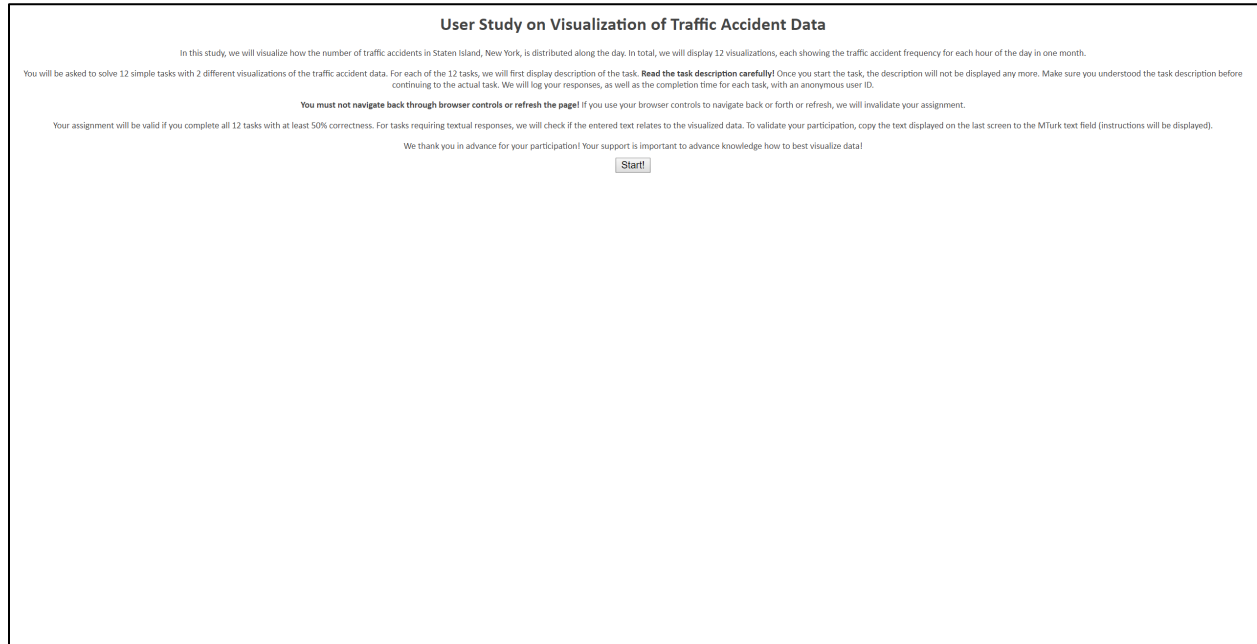


Figure 1: Screenshot of the study start screen on a 27" monitor.

The instruction text was formulated as follows:

## User Study on Visualization of Traffic Accident Data

In this study, we will visualize how the number of traffic accidents in Staten Island, New York, is distributed along the day. In total, we will display 12 visualizations, each showing the traffic accident frequency for each hour of the day in one month.

You will be asked to solve 12 simple tasks with 2 different visualizations of the traffic accident data. For each of the 12 tasks, we will first display description of the task. **Read the task description carefully!** Once you start the task, the description will not be displayed any more. Make sure you understood the task

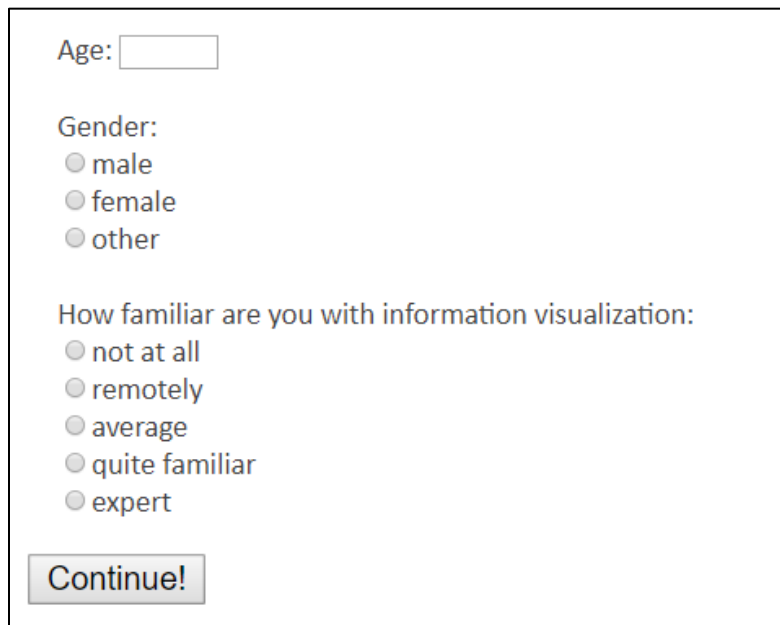
description before continuing to the actual task. We will log your responses, as well as the completion time for each task, with an anonymous user ID.

**You must not navigate back through browser controls or refresh the page!** If you use your browser controls to navigate back or forth or refresh, we will invalidate your assignment.

Your assignment will be valid if you complete all 12 tasks with at least 50% correctness. For tasks requiring textual responses, we will check if the entered text relates to the visualized data. To validate your participation, copy the text displayed on the last screen to the MTurk text field (instructions will be displayed).

We thank you in advance for your participation! Your support is important to advance knowledge how to best visualize data!

After that, users had to fill out a short demographic questionnaire:



A screenshot of a demographic questionnaire form. The form is enclosed in a black rectangular border. It contains the following elements: a text input field for 'Age:'; a 'Gender:' section with three radio button options: 'male', 'female', and 'other'; a 'How familiar are you with information visualization:' section with five radio button options: 'not at all', 'remotely', 'average', 'quite familiar', and 'expert'; and a 'Continue!' button at the bottom left.

Age:

Gender:

- ☐ male
- ☐ female
- ☐ other

How familiar are you with information visualization:

- ☐ not at all
- ☐ remotely
- ☐ average
- ☐ quite familiar
- ☐ expert

Figure 2: Demographic questionnaire.

After that, the first task was started. For each task, we showed a task description before showing the actual visualization:

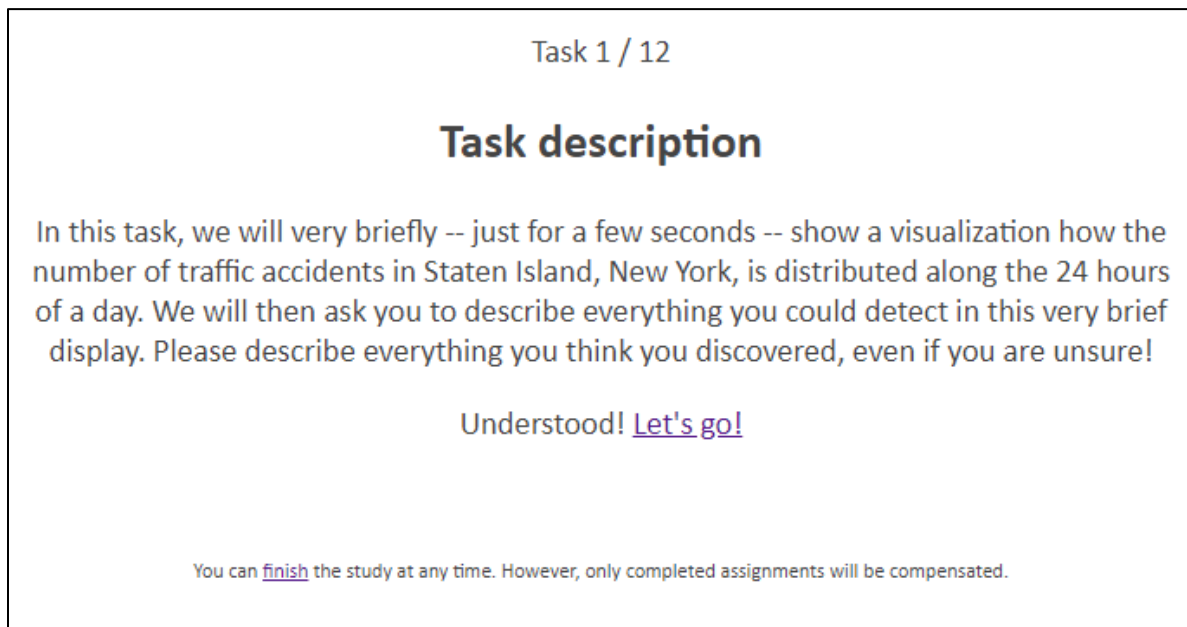


Figure 3: Task description of task 1.

Below, all task descriptions are listed:

1. In this task, we will very briefly -- just for a few seconds -- show a visualization how the number of traffic accidents in Staten Island, New York, is distributed along the 24 hours of a day. We will then ask you to describe everything you could detect in this very brief display. Please describe everything you think you discovered, even if you are unsure!
2. On the top of the next page, the text will specify a one-hour interval. Your task is to click on the bar representing this time period in the visualization. Please make sure to click on the correct bar as quickly as possible.
3. In this task, we will show an arrow marking one hour of the day. Your task is to read the number of traffic accidents at the hour marked by the arrow and enter the number in the text box below the visualization. Please be as precise as possible while performing quickly.
4. Your task is to locate the hour where most traffic accidents happened. Click on the associated bar as quickly as possible. If there are multiple hours with the same amount of accidents, select any of them.

5. In this task, we will mark two hours by a red arrow and a blue arrow. Your task is to decide if in the hour marked by the blue arrow there were more, fewer, or equally as many traffic accidents than in the hour marked by the red arrow. Judge carefully and select your answer as quickly as possible from the radio buttons underneath the visualization.
6. Now we ask you to rate how well you think the visualization is suited for showing hourly traffic accident data on a scale from 1 (= it is not suitable at all) to 5 (= there is no better way to show that).

After clicking “let’s go”, we showed the actual stimulus.

Task 1 (12l):

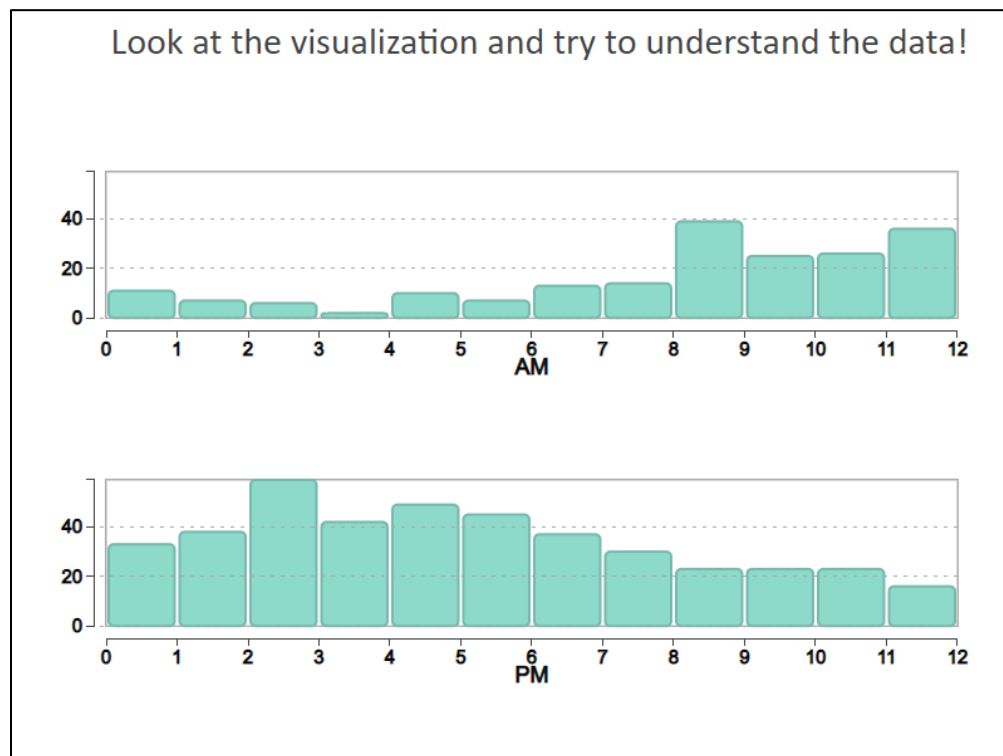


Figure 4: Task 1 (untargeted analysis) for 12l.

After 10 seconds, the answer text box was shown and the visualization was hidden:

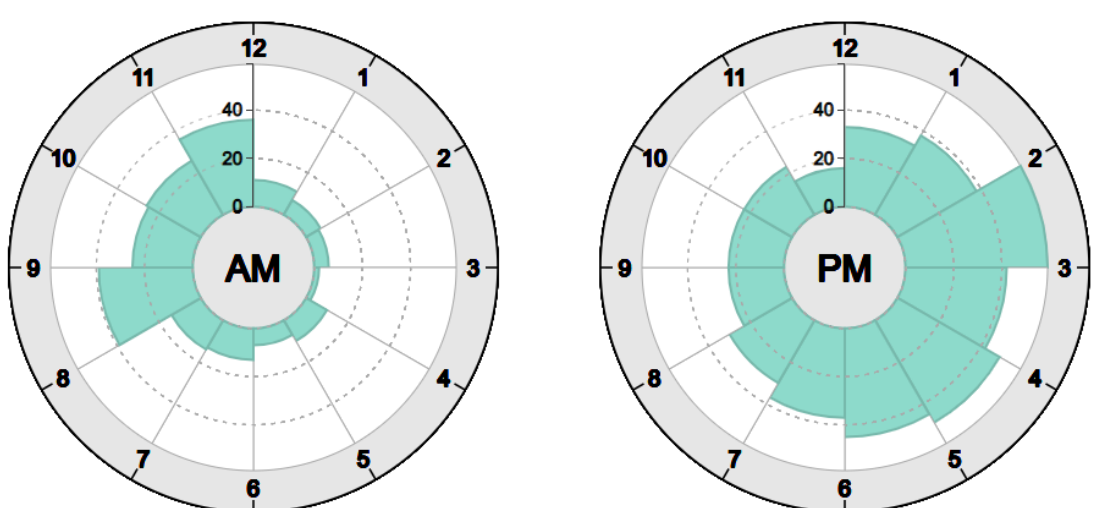
Answer

Continue

Figure 5: Response text field for task 1.

Task 2 (12r):

Please click on on the bar corresponding to the interval: 5 - 6 PM

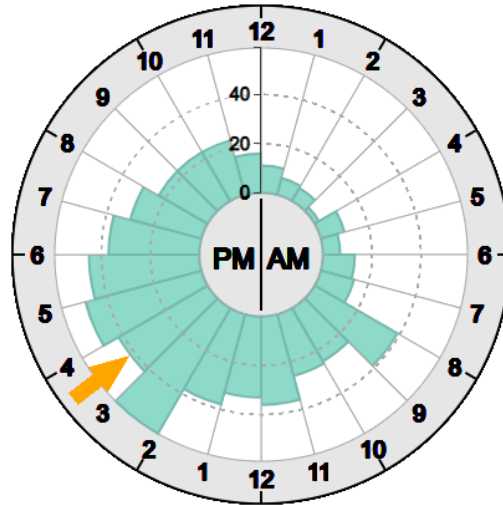


Continue

Figure 6: Task 2 (locate time) for 12r.

Task 3 (24r):

Please estimate the value for the time interval marked by the orange arrow!



Estimated value:

Continue

Figure 7: Task 3 (read value) for 24r.



Task 4 (24l):

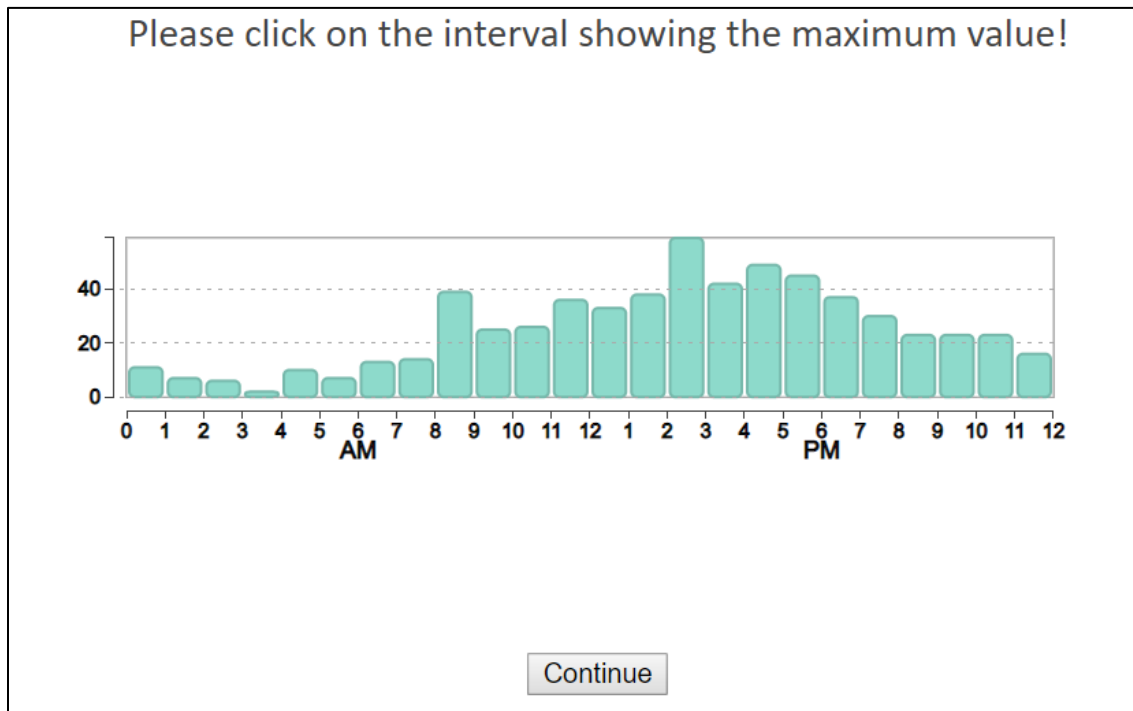
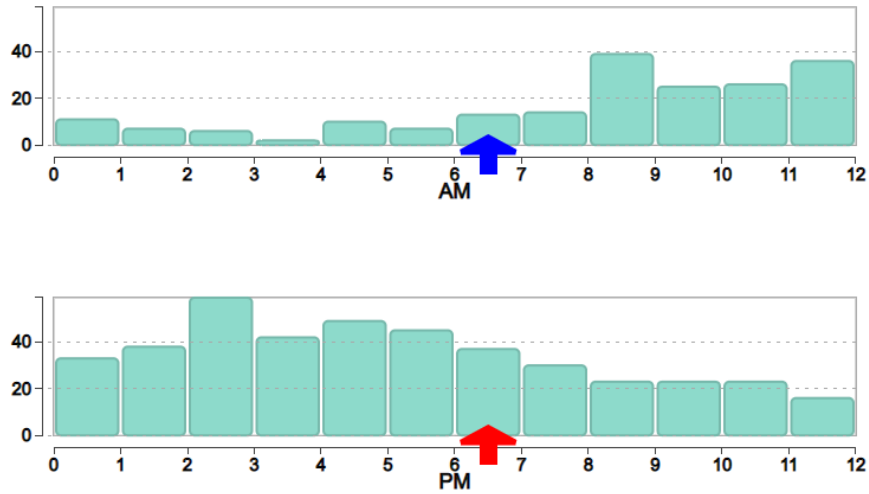


Figure 8: Task 4 (locate maximum) for 24l.

Task 5 (12l):

Does the blue arrow point to fewer, equal, or more accidents than the red arrow?



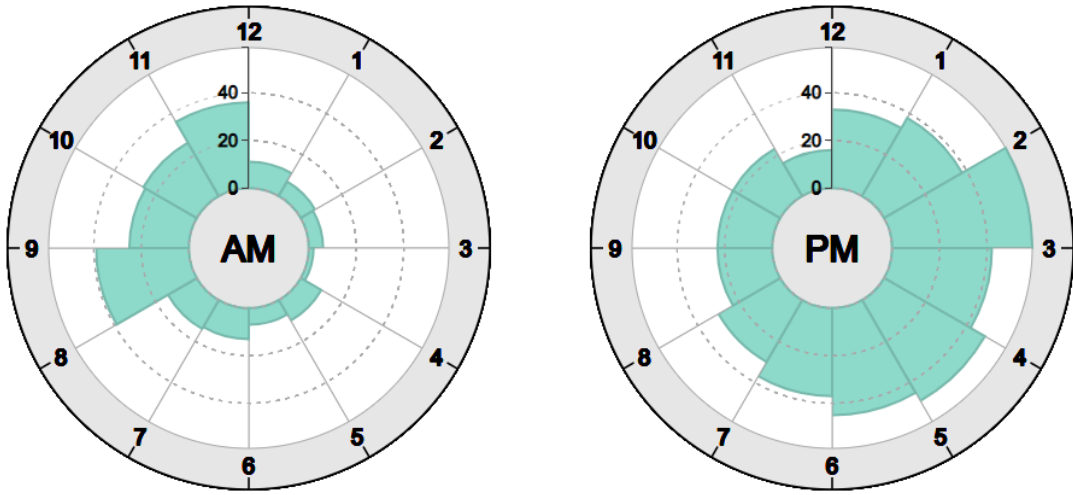
fewer (<)      equal (=)      more (>)

Continue

Figure 9: Task 5 (compare A.M./P.M. intervals) for 12l.

Task 6 (12r):

The visualization is suited for showing hourly traffic accident data:



not suitable at all = 1 2 3 4 5 = there is no better way

● ● ● ● ●

Additional comments (optional)

Continue

Figure 10: Task 6 (subjective rating) for 12r.

## Participants

In total, we had 92 users that were accepted for the HIT (i.e., 184 samples; one for linear, one radial for each user). 44 users performed cardinality 12, 48 users cardinality 24.

41.3% of users were female, 58.7% male.

		gender			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	38	41.3	41.3	41.3
	male	54	58.7	58.7	100.0
	Total	92	100.0	100.0	

The mean age of the workers was 36.2 with a minimum age of 19 and a maximum of 68.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
age	92	19	68	36.18	11.022
Valid N (listwise)	92				

Most users reported that they are not at (19.6%) all to averagely (51%) familiar with information visualization. Eight users stated that they are quite familiar, but no user claimed his- or herself as expert.

		vi			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	18	19.6	19.6	19.6
	1	19	20.7	20.7	40.2
	2	47	51.1	51.1	91.3
	3	8	8.7	8.7	100.0
	Total	92	100.0	100.0	

There are no obvious differences between the male and female workers in terms of age or visualization literacy:

Descriptives				
gender			Statistic	Std. Error
vi	female	Mean	1.34	.143
		95% Confidence Interval for Mean	Lower Bound	1.05
			Upper Bound	1.63
		5% Trimmed Mean	1.32	
		Median	1.50	
		Variance	.772	
		Std. Deviation	.878	
		Minimum	0	
		Maximum	3	
		Range	3	
		Interquartile Range	1	
		Skewness	-.245	.383
		Kurtosis	-.889	.750
	male	Mean	1.59	.125
		95% Confidence Interval for Mean	Lower Bound	1.34
			Upper Bound	1.84
		5% Trimmed Mean	1.60	
		Median	2.00	
		Variance	.850	
		Std. Deviation	.922	
		Minimum	0	
		Maximum	3	
		Range	3	
		Interquartile Range	1	
		Skewness	-.583	.325
		Kurtosis	-.553	.639
age	female	Mean	37.03	1.776
		95% Confidence Interval for Mean	Lower Bound	33.43
			Upper Bound	40.62
		5% Trimmed Mean	36.42	
		Median	35.50	

	Variance	119.864	
	Std. Deviation	10.948	
	Minimum	20	
	Maximum	66	
	Range	46	
	Interquartile Range	17	
	Skewness	.819	.383
	Kurtosis	.186	.750
male	Mean	35.59	1.516
	95% Confidence Interval for Mean	Lower Bound	32.55
		Upper Bound	38.63
	5% Trimmed Mean	34.89	
	Median	33.00	
	Variance	124.057	
	Std. Deviation	11.138	
	Minimum	19	
	Maximum	68	
	Range	49	
	Interquartile Range	10	
	Skewness	1.171	.325
	Kurtosis	.929	.639

# Analysis

We used IBM SPSS Statistics 25 for the analysis. For statistical comparisons in all tasks, we employed an ANOVA (General Linear Model repeated measures technique of SPSS) with layout as within-subjects factor and cardinality as between-subjects factor. Completion times were log-transformed. For analysis of binary codes in task 1, we used a Generalized Linear Mixed Model (GLMM) binary logistic regression and with the user ID as subjects, layout as repeated measures factor and cardinality as between-subjects factor.

## Task 1: Untargeted Analysis

### Open Coding

We established the following code book for the analysis of the text responses:

#### **Observation [numeric]**

Any atomic insight into the data that cannot be further broken down into multiple observations. Example: “There are **a lot of accidents in the morning between 8 am and 10 am**. Then again **in the afternoon between 1 and 5 pm**.” → 2 observations

#### **Comparison [binary]**

Comparing values between two time points or time intervals. Example: “There are far **fewer** accidents in the early AM hours **than** in the afternoon PM hours”

#### **Salient features [binary]**

One or more characteristics that are salient, such as peaks. Example: “there are the **most amount of crashes** from 4 to 5 pm”

#### **Quantitative time reference [binary]**

A point in time or a time interval is specified by a numeric reference or numeric range. Example: “The largest number of accidents took place between **2-6PM**.”

#### **Qualitative time reference [binary]**

A point in time or a time interval is specified by a qualitative time reference with loose semantics. Example: “There are far fewer accidents in the early AM hours than in the **afternoon** PM hours”

For the number of observations, we picked the average number assigned by the three independent coders for each response. For the remaining codes, we assigned per response if the user performed any comparison, reported on salient features, and whether he or she used a quantitative time reference and / or a qualitative time reference the response. The inter-coder reliability score was computed using Krippendorff's alpha (comparison: 0.91, salient features: 0.86, quantitative: 0.94, qualitative: 0.91), where scores above 0.8 are acceptable.

### Number of Observations

For the coded number of observations, we picked the average number reported by the three independent coders and performed an ANOVA.

Cardinality did not have a significant effect on the number of observations:

#### Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	352.820	1	352.820	254.081	.000	.738
<b>cardinality</b>	<b>.259</b>	<b>1</b>	<b>.259</b>	<b>.187</b>	<b>.667</b>	<b>.002</b>
Error	124.975	90	1.389			

Layout had a medium-sized significant effect, but there is no interaction between layout and cardinality:

#### Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
<b>layout</b>	<b>Sphericity Assumed</b>	<b>3.974</b>	<b>1</b>	<b>3.974</b>	<b>8.852</b>	<b>.004</b>	<b>.090</b>
	Greenhouse-Geisser	3.974	1.000	3.974	8.852	.004	.090



	Huynh-Feldt	3.974	1.000	3.974	8.852	.004	.090
	Lower-bound	3.974	1.000	3.974	8.852	.004	.090
layout * cardinality	Sphericity Assumed	.206	1	.206	.458	.500	.005
	Greenhouse- Geisser	.206	1.000	.206	.458	.500	.005
	Huynh-Feldt	.206	1.000	.206	.458	.500	.005
	Lower-bound	.206	1.000	.206	.458	.500	.005
Error(layout)	Sphericity Assumed	40.401	90	.449			
	Greenhouse- Geisser	40.401	90.000	.449			
	Huynh-Feldt	40.401	90.000	.449			
	Lower-bound	40.401	90.000	.449			

## Correlation between Number of Observations and Demographics

We first explored whether the number of reported observations in task 1 correlates with any demographic statistics collected.

Here is a scatterplot of self-reported visualization literacy with the number of observations:

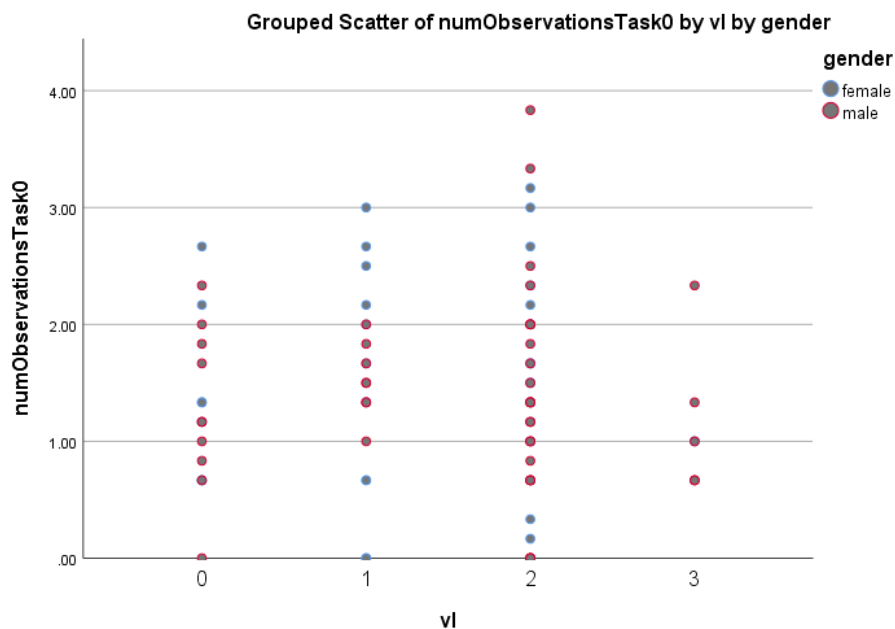


Figure 11: Scatterplot of number of observations reported for task 1 over self-reported visualization literacy. Blue dots are females, red dots are values for male participants.

The correlation between self-reported visualization literacy and number of observations is very low and not significant:

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.024 <sup>a</sup>	.001	-.011	.83388

a. Predictors: (Constant), vl

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.035	1	.035	.050	.824 <sup>b</sup>
	Residual	62.582	90	.695		
	Total	62.617	91			

a. Dependent Variable: numObservationsTask0

b. Predictors: (Constant), vl

There is clearly also no difference between females and males concerning the number of reported observations:

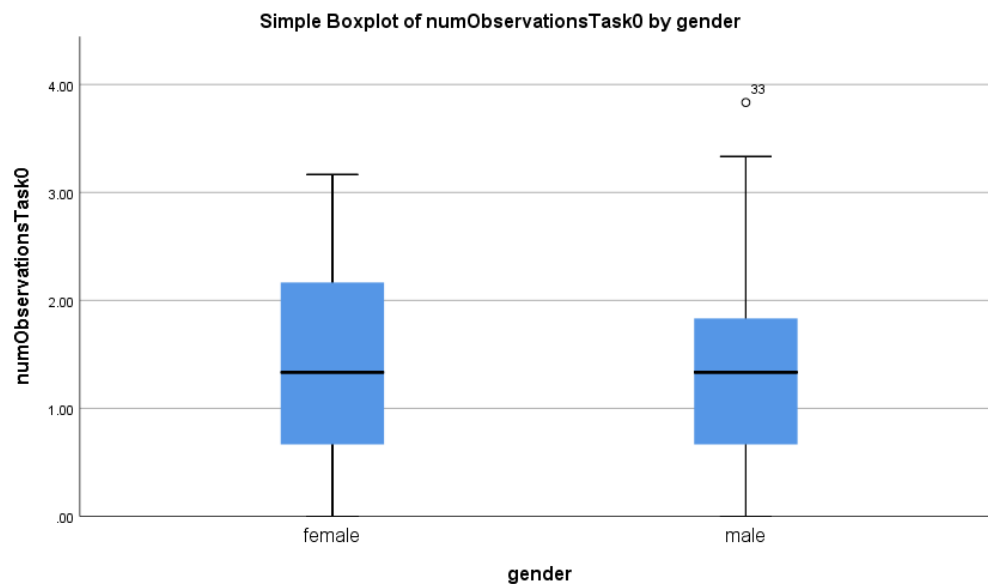


Figure 12: Box plots of the number of reported observations by gender.

We can see a very low effect of age on the number of reported observations, where it looks like the number of reported observations raises with age and then drops again. This behavior can be modeled with a quadratic regression, but the effect is rather low:

### Model Summary and Parameter Estimates

Dependent Variable: numObservationsTask0

Model Summary						Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.038	3.580	1	90	.062	.855	.015		
Logarithmic	.055	5.250	1	90	.024	-1.021	.679		
Inverse	.070	6.761	1	90	.011	2.186	-26.651		
<b>Quadratic</b>	<b>.107</b>	<b>5.308</b>	<b>2</b>	<b>89</b>	<b>.007</b>	<b>-1.526</b>	<b>.140</b>	<b>-.002</b>	
Cubic	.107	3.521	3	88	.018	-.865	.087	.000	-1.063E-5

The independent variable is age.

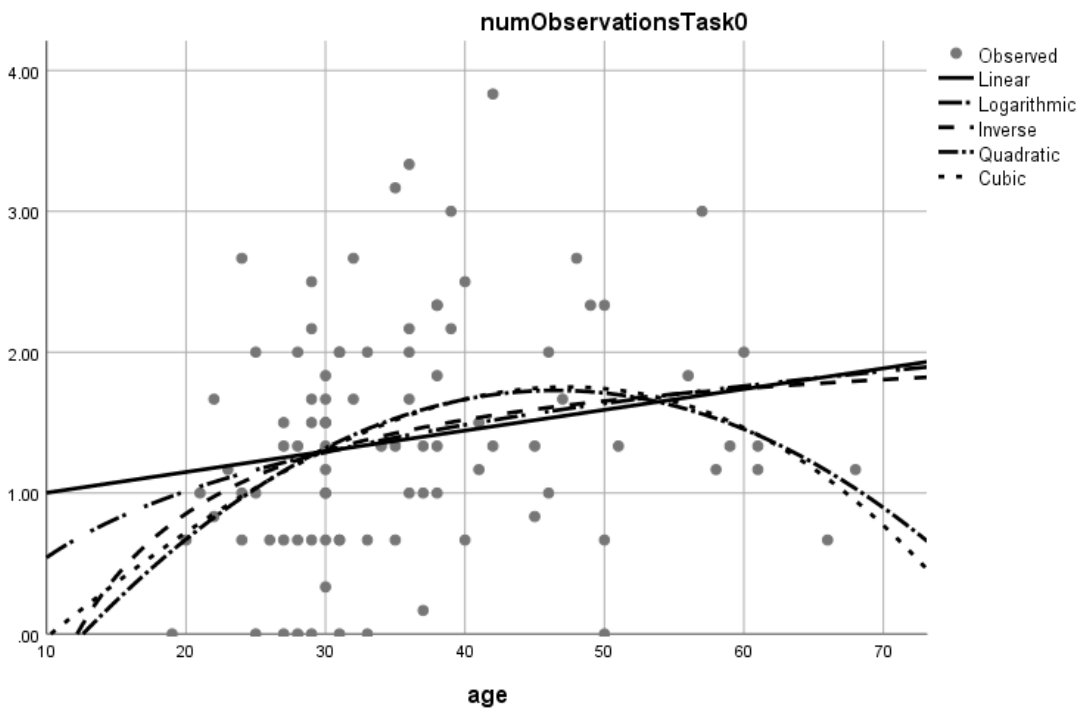


Figure 13: The effect of age (x axis) on the number of reported observations in task 1 (y axis). Different regression curves are overlaid.

We conclude that no specific user characteristic led to a lower observation rate. We only have a slight indication that very low or very high age might be a factor leading to fewer observations.

### Correlation between Number of Observations and Trial Order

We analyzed whether the number of observations was influenced by the order of presentation. It seems that this was not the case:

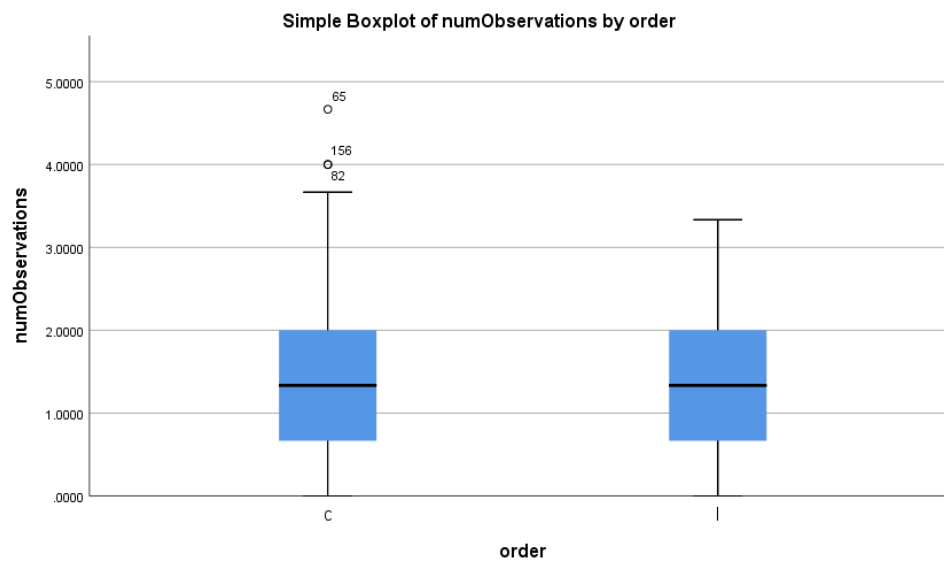


Figure 14: Box plots of the number of observations for both task orders (c: radial first, l: linear first).

Also, separated by condition, we do not see an effect. This means, for instance for 12r, the number of observations were not higher if the linear chart (“l”) was seen first.

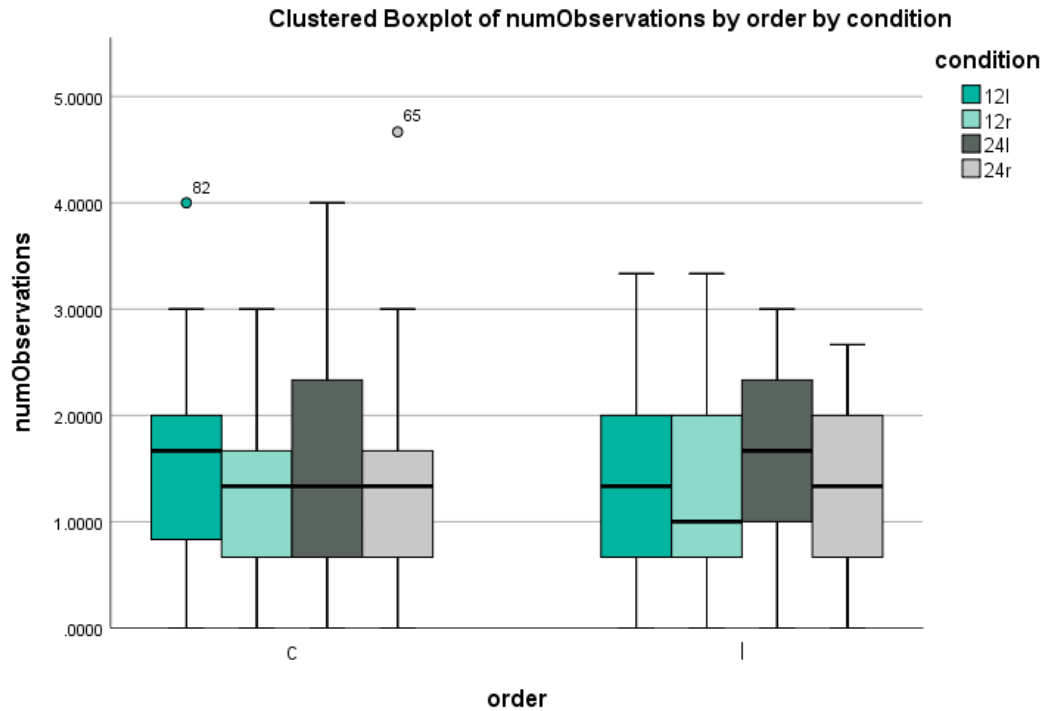


Figure 15: Box plots of number of observations split by layout order (c: radial first, l: linear first) and the four conditions (color).

### Qualitative Analysis of Responses without Observations

From all 184 coded responses, 24 did not contain any observations. The responses without observations were almost equally distributed among the four conditions (first row):

**numObservations \* condition Crosstabulation**

Count

		condition				
		12l	12r	24l	24r	Total
numObservations	.0000	6	6	5	7	24
	.3333	1	1	1	0	3
	.6667	7	10	8	13	38
	1.0000	5	4	5	2	16
	1.3333	5	8	5	9	27
	1.6667	3	3	3	6	15
	2.0000	8	7	6	6	27
	2.3333	3	2	4	2	11
	2.6667	0	1	5	1	7

	3.0000	3	1	3	1	8
	3.3333	2	1	1	0	4
	3.6667	0	0	1	0	1
	4.0000	1	0	1	0	2
	4.6667	0	0	0	1	1
Total		44	44	48	48	184

The 24 responses without data observations were contributed by 16 participants (out of 92). Eight participants did not provide any meaningful reports for both layout conditions, the other eight (marked here in red) had no observations about the data in just one condition. We added the condition in brackets, as well as the trial for the responses marked in red.

1.
  - This looks like a huge mess. You really have to look at it to understand the data and it was hard to read. (24r; 2)
2.
  - The data gives me a timeframe of 24 hours, with the number of accidents (I believe) per hour. (24r)
  - I was able to see the number of accidents per hour, I can tell whether it's for morning hours or past noon. Also you can tell the hours most accidents happen at. (24l)
3.
  - The volume of traffic per hour (12l)
  - The amount of traffic accidents during different hours of the day. (12r)
4.
  - the time and frequency of most accidents (24r)
  - the bars heigh (y-axis) represents the number of accidents and the x-axis the time of the accidents (24l)
5.
  - The number of traffic accidents every hour? (12l; 1)
6.
  - If it goes over 40, you wouldn't know. (12r)

- I chose the one I thought to be the correct answer, but at the last minute, I noticed a different one, which I think was correct on the bottom bar. (12l)
- 7.
- I think the clock showed the times and the worst times for accidents. The blue area represented the heavy times for accidents. (12r)
  - The table displayed the hours where accidents occur and the blue bars showed the number of accidents that occur during those time periods. (12l)
- 8.
- It's a 24 hour chart that goes from the center outward. The closer it gets to the circle's edge, the more accidents that took place during that time frame. It is divided into AM/PM sections. (24r; 2)
- 9.
- This one was very hard to understand and I did not get it at all (12r; 2)
- 10.
- the percent of accident was on the left vertical side and the time of day was on the horizontal side (12l; 2)
- 11.
- The wheel was confusing, I didn't really understand any of it. (24r; 2)
- 12.
- Umm something went down and then back up to twelve over time. No idea what (12l)
  - i have no idea what any of this means. they looked like dart boards (12r)
- 13.
- I'm not sure what I'm answering? (24l; 1)
- 14.
- it's a wheel with 24 hours in 1 hour increments. It shows up to 40 accidents per hour. (24r)
  - it seems self explanatory to me. each bar is one hour. (24l)
- 15.
- Judging from the hours on the graph it looks as if it's measuring the lack of traffic rather than how much traffic there is. (12r; 1)
- 16.

- More accidents occur during certain times of the day. (24r)
- bar graph of the occurrence of accidents and time they happened (24l)

The reports by the users not providing any data insight for both conditions indicate that they misinterpreted the description of task 1 and described the visualization rather than the data. We are particularly interested in those cases where users reported observations in one condition, which implies that they understood the task description, but were not able to decode the visualization.

First, we checked whether users had a learning effect and just could improve their response for the second condition. However, this was not the case. There were more meaningless reports in the second condition:

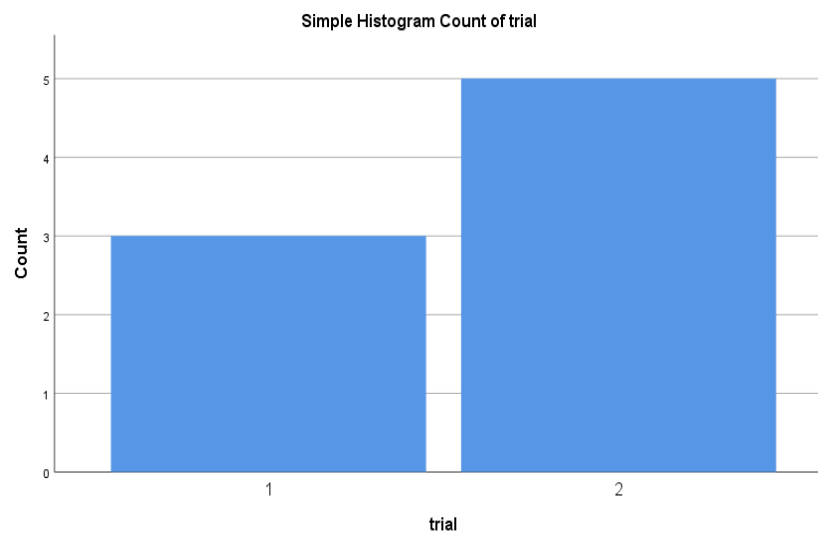


Figure 16: Number of meaningless observation reports per trail (1: in the first round, 2: in the second round).

We then looked into the distribution of meaningless reports for all four conditions. In total, most meaningless reports were delivered for 24r. Two of these reports were very explicitly stating that the user was not able to decode the data – after seeing 24-hours linear bar chart. One user of 12r issued a similar report after seeing 12l.



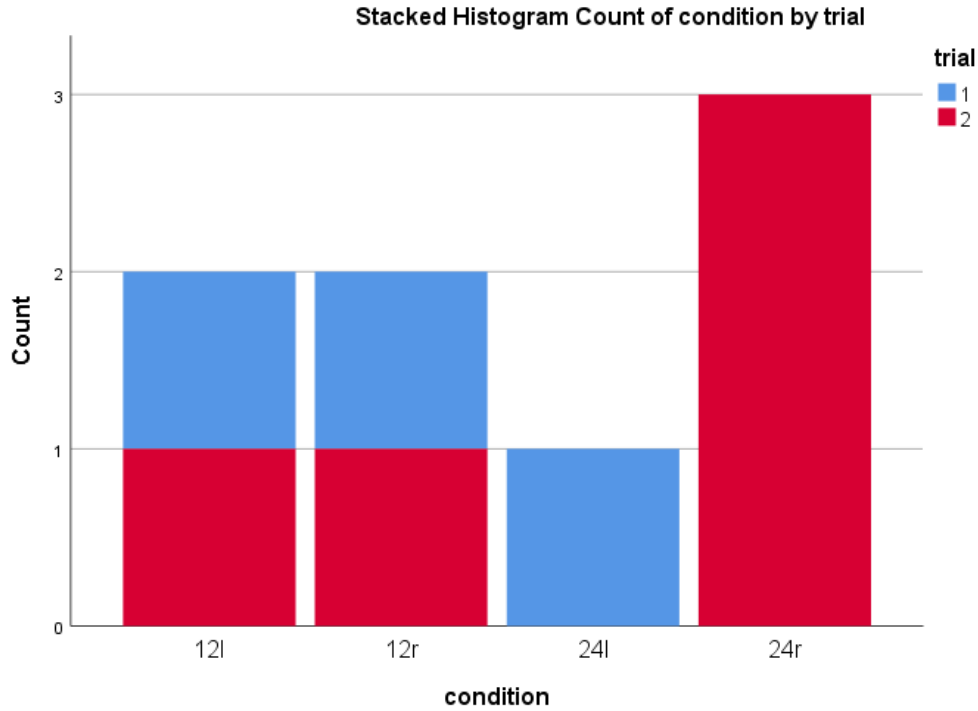


Figure 17: Number of meaningless observation reports for each condition, separated by trial (color; 1: in the first run, 2: in the second run).

Mind that we checked all gathered HIT responses for their overall correctness directly after the experiment. All accepted HITs included in this analysis passed the given quality criteria for the low-level tasks.

### Comparisons, Salient Features, and Time References

The codes for comparisons, salient features, quantitative references, and qualitative references were assigned as binary values. We only picked 1 if at least 2 coders were assigning a 1, otherwise 0. For all remaining codes, we used a GLMM with a binary logistic regression.

The numbers of cases where users compared different time steps or time intervals is neither affected by layout nor by cardinality:

Fixed Effects <sup>a</sup>				
Source	F	df1	df2	Sig.
Corrected Model	.000	3	180	1.000
layout	.000	1	180	.999
cardinality	.000	1	180	1.000

layout * cardinality	.000	1	180	1.000
----------------------	------	---	-----	-------

Probability distribution: Binomial

Link function: Logit

a. Target: numComparisonsAgreed

However, the number of responses containing comparisons was generally low. There are more responses containing comparisons using the radial layout than using the linear one:

#### numComparisonsAgreed \* condition Crosstabulation

Count

		condition				Total
		12l	12r	24l	24r	
numComparisonsAgreed	0	43	38	45	42	168
	1	1	6	3	6	16
Total		44	44	48	48	184

Also, there is no statistically significant effect of layout or cardinality on the number of times users reported salient features in the visualization:

#### Fixed Effects<sup>a</sup>

Source	F	df1	df2	Sig.
Corrected Model	.539	3	180	.656
layout	.042	1	180	.838
cardinality	1.230	1	180	.269
layout * cardinality	.350	1	180	.555

Probability distribution: Binomial

Link function: Logit

a. Target: numSalientFeaturesAgreed

Salient features were observed much more frequently (58%) than comparisons, with a slight tendency to be detected more frequently using 24 hours representations (60% for 24l and 65% for 24r) than 12 hours representations (54% for 12l and 52% for 12r, respectively).

**numSalientFeaturesAgreed \* condition Crosstabulation**

Count

		condition				
		12l	12r	24l	24r	Total
numSalientFeaturesAgreed	0	20	21	19	17	77
	1	24	23	29	31	107
Total		44	44	48	48	184

Similarly, the number of times participants used quantitative expressions to describe points in times or time intervals is not caused by layout or cardinality:

**Fixed Effects<sup>a</sup>**

Source	F	df1	df2	Sig.
Corrected Model	.675	3	180	.569
layout	1.470	1	180	.227
cardinality	.000	1	180	.987
layout * cardinality	.464	1	180	.497

Probability distribution: Binomial

Link function: Logit

a. Target: numQuantitativeTimesAgreed

The number of responses that contain quantitative time identifiers range between 56% (24l) to 64% (24r).

**numQuantitativeTimesAgreed \* condition Crosstabulation**

Count

condition | Total

		12l	12r	24l	24r	
numQuantitativeTimesAgreed	0	18	17	21	17	73
	1	26	27	27	31	111
Total		44	44	48	48	184

However, layout has a significant effect on the number of times participants used qualitative expressions to describe time points or time intervals.

Fixed Effects <sup>a</sup>				
Source	F	df1	df2	Sig.
Corrected Model	2.829	3	180	.040
layout	8.111	1	180	.005
cardinality	.415	1	180	.520
layout * cardinality	.023	1	180	.880

Probability distribution: Binomial

Link function: Logit

a. Target: numQualitativeTimesAgreed

The number of responses containing qualitative time references is highest for 24l (69%) and lowest for 12r (50%).

numQualitativeTimesAgreed * condition Crosstabulation						
Count		condition				
		12l	12r	24l	24r	Total
numQualitativeTimesAgreed	0	16	22	15	21	74
	1	28	22	33	27	110
Total		44	44	48	48	184

### Exploratory Analysis of Time Periods Mentioned

We analyzed the time periods users mentioned in their observation reports to explore whether the visual encoding influences which time periods the users are focusing on. Two independent coders first analyzed the qualitative time periods mentioned by all participants and derived a categorization of time periods to which to assign the mentioned observations.

Time Period	Additional Labels	Time Slot
Night		00-05
Very Early Morning		05-06
Early Morning	Morning Morning rush hour	06-09
Mid Morning	Morning	09-10
Late Morning	Morning	10-12
Noon	Lunchtime Middle of the day	12-13
Early Afternoon	Afternoon	13-15
Mid Afternoon	Afternoon	15-16
Early Evening	Afternoon Late afternoon Afternoon rush hour Evening	16-19
Late Evening	Evening	19-24

For every mention of a time period or quantitative time reference, coders then assigned the observations to one or multiple of these time periods.

Examples:

- It was not very busy at **11 PM** and the busiest time was around **4/5 PM**.  
→ Late morning, early evening
- Most accidents happened within **noon until 6pm**.  
→ Noon, early afternoon, mid afternoon, early evening

As reference, we visualize the value distribution (i.e., the number of traffic accidents) per hour in the twelve data sets used in the experiment:

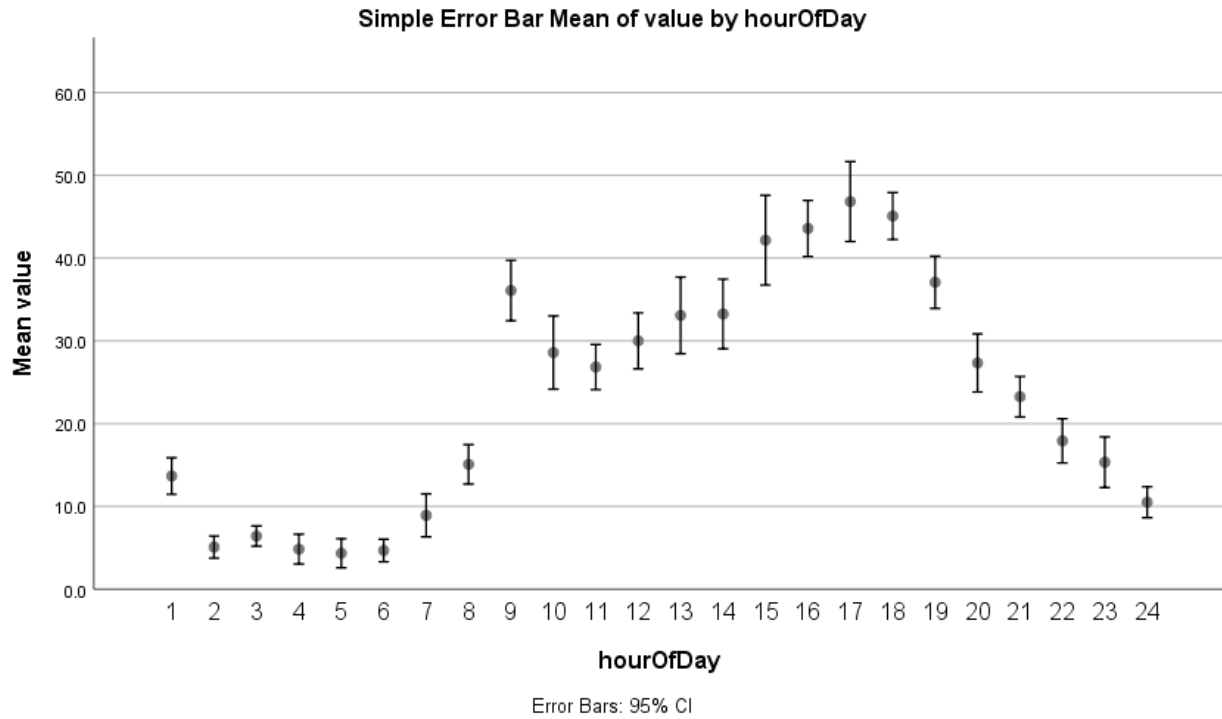


Figure 18: Distribution of the 12 used data sets per bin (i.e., hour of the day).

Below, we visualize the average number of observations containing qualitative or quantitative references to the coded time intervals per condition as superimposed line chart:

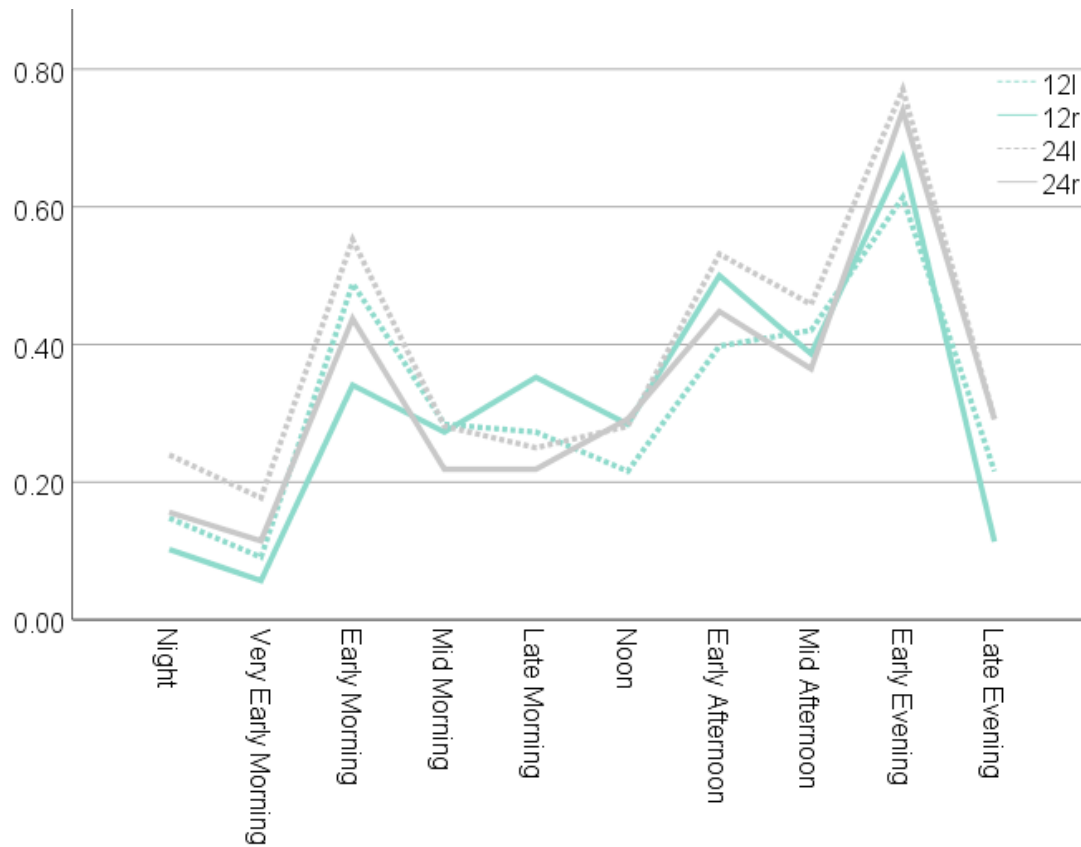


Figure 19: Number of observations associated with the coded day time intervals split by condition.

Separated into the individual conditions, we can observe the following distributions:

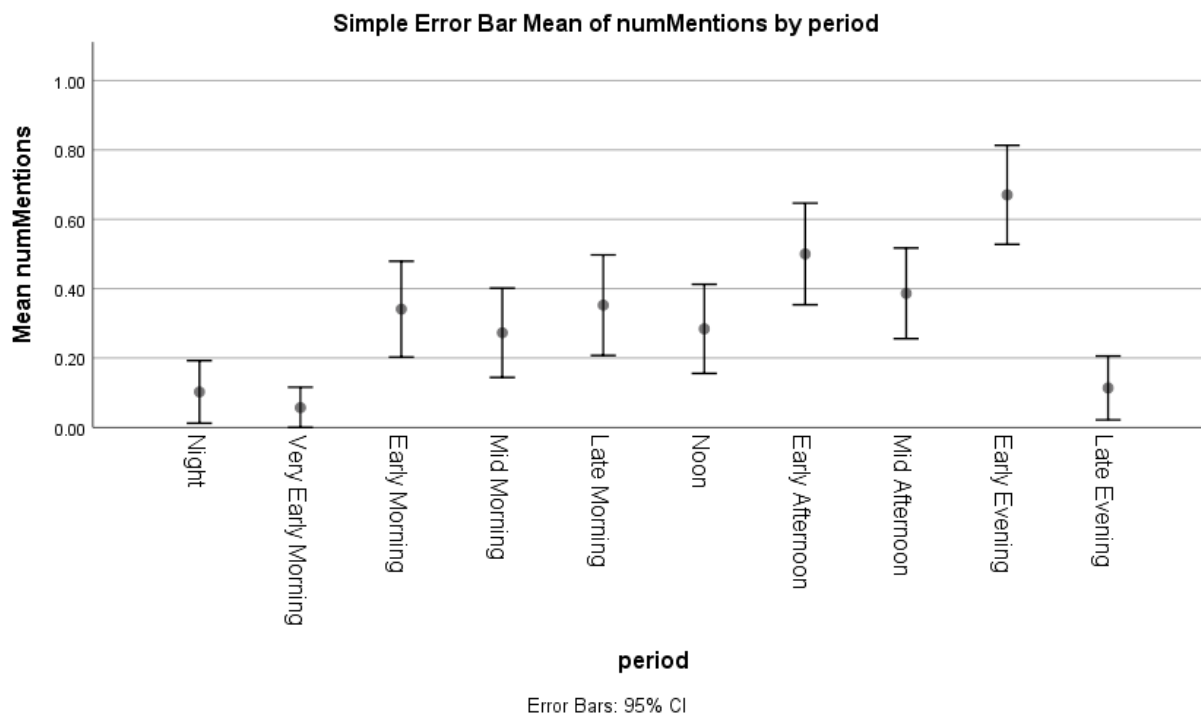


Figure 20: Number of observations per day time period for 12r.

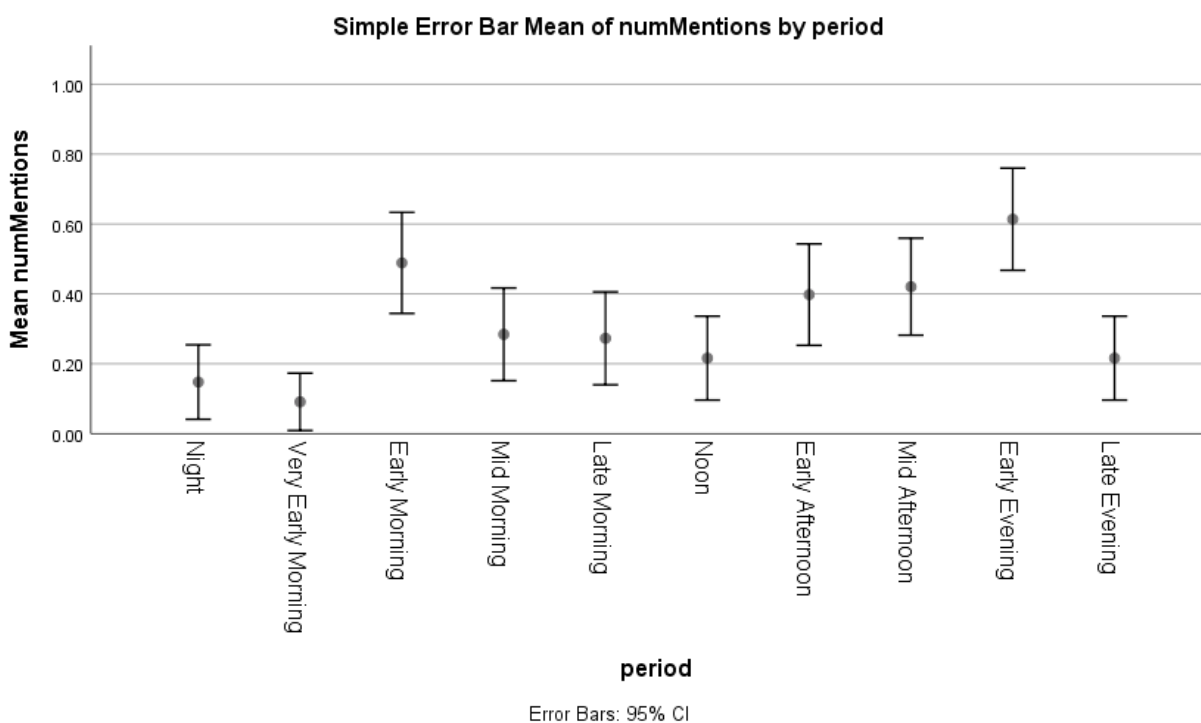


Figure 21: Number of observations per day time period for 12l.



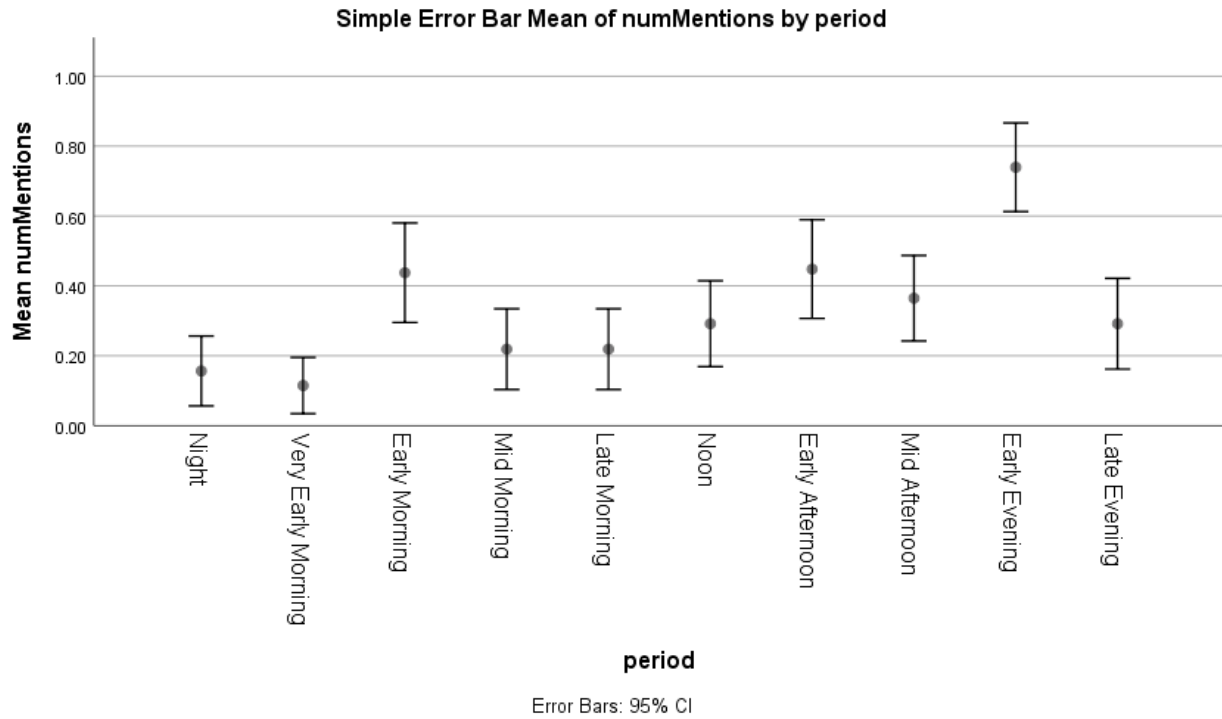


Figure 22: Number of observations per day time period for 24r.

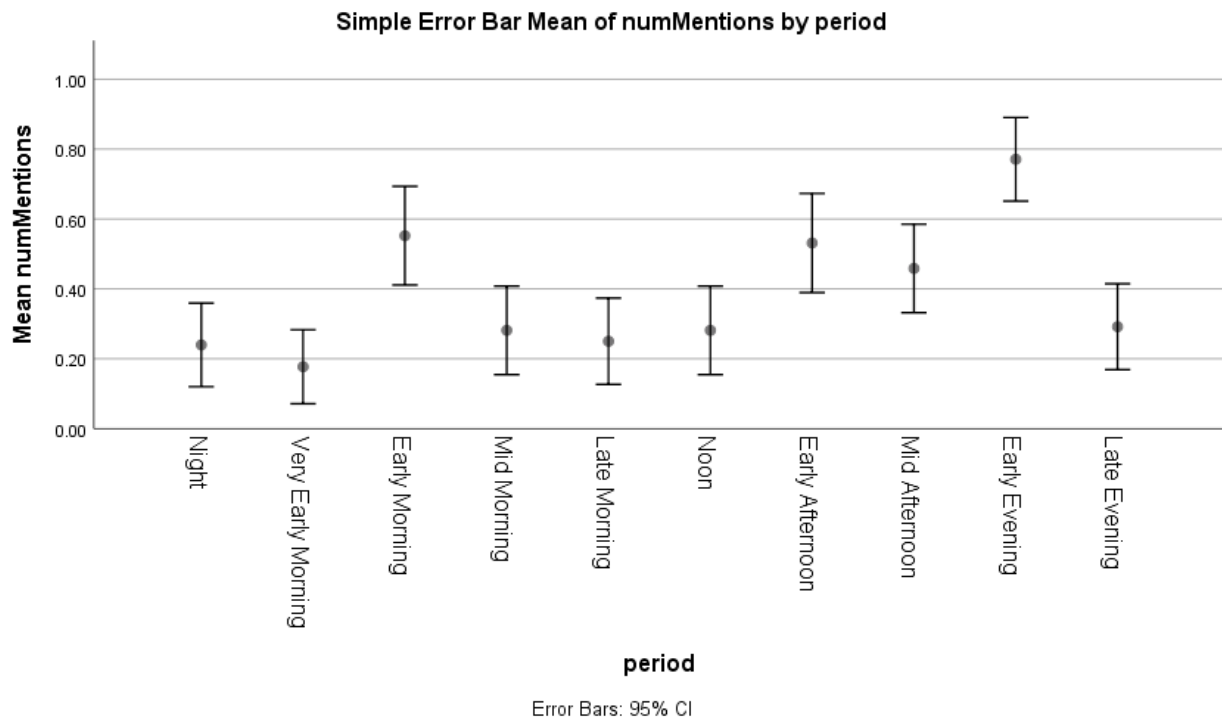


Figure 23: Number of observations per day time period for 24l.

We can make the following qualitative observations from the charts:

- 24l has the highest number of observations due to most mentions of rush hour peaks.
- 12r has the lowest number of mentions for night until early morning.
- 12r has the highest number of observations mentioning late morning and noon.
- 12l has the lowest number of observations mentioning the afternoon peaks, and is the only visualization where users mention the morning peaks almost as frequently as the afternoon peaks.

## Task 2: Locate Time

### Accuracy

We first computed the error by  $\min(\text{abs}(\text{selected bar index} - \text{correct bar index}), 1)$ . 152 samples are correct (.00), which corresponds to 82.6% of all cases. The highest error rate (13 samples) was detected for 12r (30%), the lowest for 12l (10%). 24r has 20% incorrect responses, 24l 17%.

### condition \* keyDiff Crosstabulation

Count		keyDiff		Total
		.00	1.00	
condition	c12	31	13	44
	c24	40	8	48
	l12	40	4	44
	l24	41	7	48
Total		152	32	184

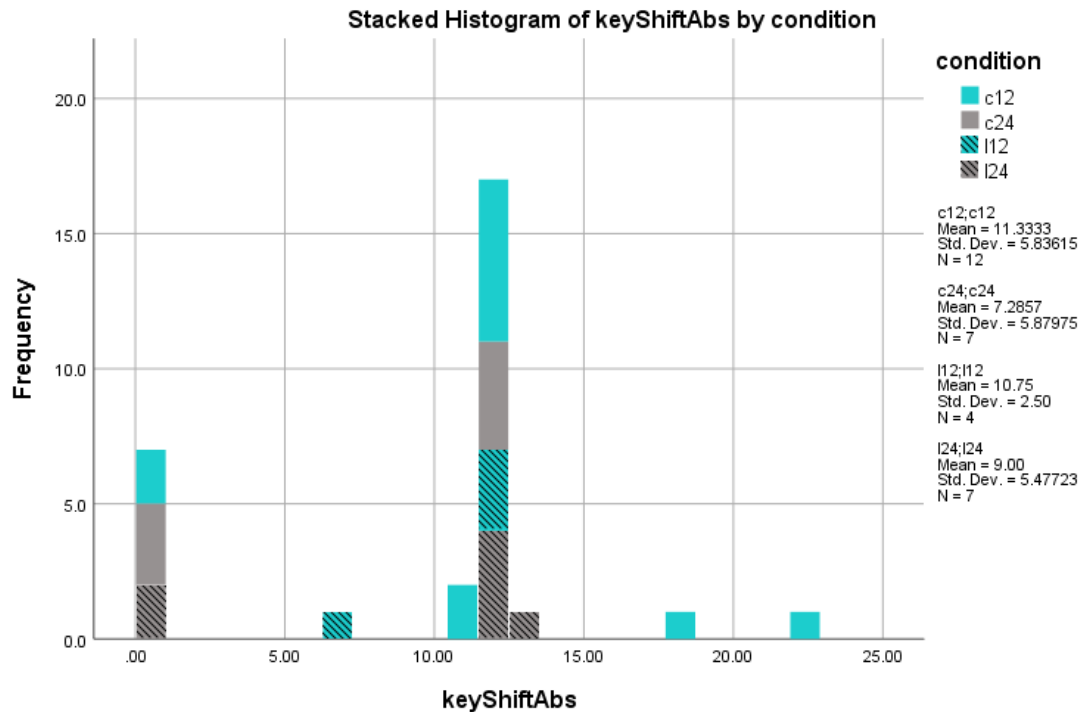
### Exploratory Analysis of Error Cases

Exploring how much the selected bars were shifted from the correct bar, computed by  $\text{abs}(\text{inputKey} - \text{actualKey})$ , we can observe that most incorrect responses were shifted by 12 hours (17 out of 30, 57%). 23% of incorrect selections were shifted by one hour (7 out of 30). Three incorrect cases were shifted 11 and 13 hours, respectively, which might indicate a combination of the two most common errors (switching a.m./p.m. and shifting one hour). Of the remaining three cases, two (7 and 22) were committed by the same participant.

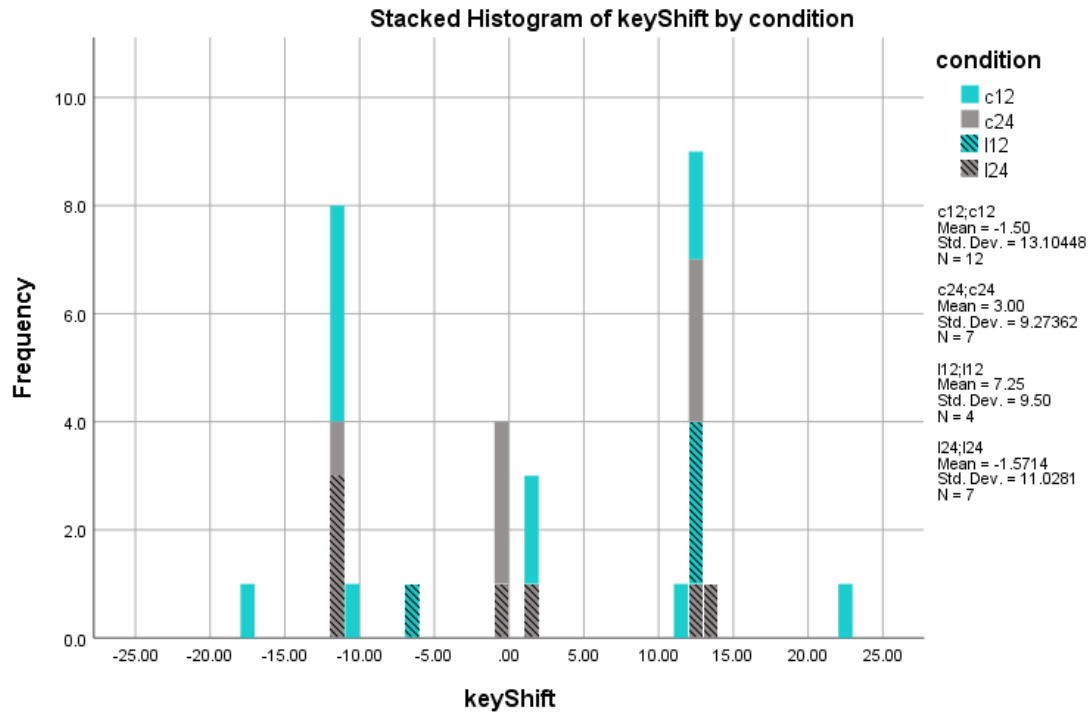
### condition \* keyShiftAbs Crosstabulation

Count

		keyShiftAbs								Total
		.00	1.00	7.00	11.00	12.00	13.00	18.00	22.00	
condition	c12	31	2	0	2	6	0	1	1	43
	c24	40	3	0	0	4	0	0	0	47
	l12	40	0	1	0	3	0	0	0	44
	l24	41	2	0	0	4	1	0	0	48
Total		152	7	1	2	17	1	1	1	182



Here, we show a histogram of the actual shifts per condition. Values < 0 represent incorrect p.m. selections when the actual time interval was a.m. (e.g., selecting 9 p.m. instead of 9 a.m. would result in -12), values > 0 represent incorrect a.m. selections. Correct selections (0) are now shown. We can observe that all 3 12-hours switches by l2l were caused by users incorrectly selecting the given interval on the upper (a.m.) chart. We can also see that with 24r (gray) bars, all 3 one-hour shifts were shifted one hour back (e.g., selecting 3-4 a.m. instead of 2-3 a.m.), and with 12r (green), all 2 one-hour shifts were shifted one hour to the front (e.g., selecting 1-2 a.m. instead of 2-3 a.m.).



**condition \* keyShift Crosstabulation**

Count

		keyShift											Total
		-18.00	-12.00	-11.00	-7.00	-1.00	.00	1.00	11.00	12.00	13.00	22.00	
condition	c12	1	4	1	0	0	31	2	1	2	0	1	43
	c24	0	1	0	0	3	40	0	0	3	0	0	47
	l12	0	0	0	1	0	40	0	0	3	0	0	44
	l24	0	3	0	0	1	41	1	0	1	1	0	48
Total		1	8	1	1	4	152	3	1	9	1	1	182

## Completion Time

We removed all users with an incorrect sample for further analysis, which leaves responses by 63 users (i.e., 126 samples).

We performed an ANOVA layout as within-subjects factor and cardinality as between-subjects factor.

We did not find a main effect for cardinality:

### Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	70.392	1	70.392	1351.407	.000	.957
<b>cardinality</b>	<b>.012</b>	<b>1</b>	<b>.012</b>	<b>.223</b>	<b>.638</b>	<b>.004</b>
Error	3.177	61	.052			

However, we found a large main effect for layout and a small interaction between layout and cardinality:

### Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
<b>layout</b>	<b>Sphericity Assumed</b>	<b>.576</b>	<b>1</b>	<b>.576</b>	<b>21.279</b>	<b>.000</b>	<b>.259</b>
	Greenhouse-Geisser	.576	1.000	.576	21.279	.000	.259
	Huynh-Feldt	.576	1.000	.576	21.279	.000	.259
	Lower-bound	.576	1.000	.576	21.279	.000	.259
<b>layout * cardinality</b>	<b>Sphericity Assumed</b>	<b>.122</b>	<b>1</b>	<b>.122</b>	<b>4.504</b>	<b>.038</b>	<b>.069</b>
	Greenhouse-Geisser	.122	1.000	.122	4.504	.038	.069
	Huynh-Feldt	.122	1.000	.122	4.504	.038	.069

	Lower-bound	.122	1.000	.122	4.504	.038	.069
Error(layout)	Sphericity	1.650	61	.027			
	Assumed						
	Greenhouse-Geisser	1.650	61.000	.027			
	Huynh-Feldt	1.650	61.000	.027			
	Lower-bound	1.650	61.000	.027			

### Correlation between Completion Time and Demographics

In the media, concerns have been expressed about the youth's decreasing ability to read the analog clock<sup>1</sup>. We therefore explored the influence of age on the completion time to verify if this phenomenon could also explain why participants were not able to locate the time more efficiently using 12r. Below, we visualize a scatterplot of the completion time in relation to the participant's age for 12r locate time trials only. Indeed, we can find a correlation. However, this correlation is similar to the correlation between the number of observations and the participants' age: the performance increases until reaching around age 35-40, then it decreases again:

### Model Summary and Parameter Estimates

Dependent Variable: completionTime

Model Summary						Parameter Estimates		
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.324	9.838	2	41	.000	17.794	-.725	.011

The independent variable is age.

<sup>1</sup> <https://www.telegraph.co.uk/education/2018/04/24/schools-removing-analogue-clocks-exam-halls-teenagers-unable/>

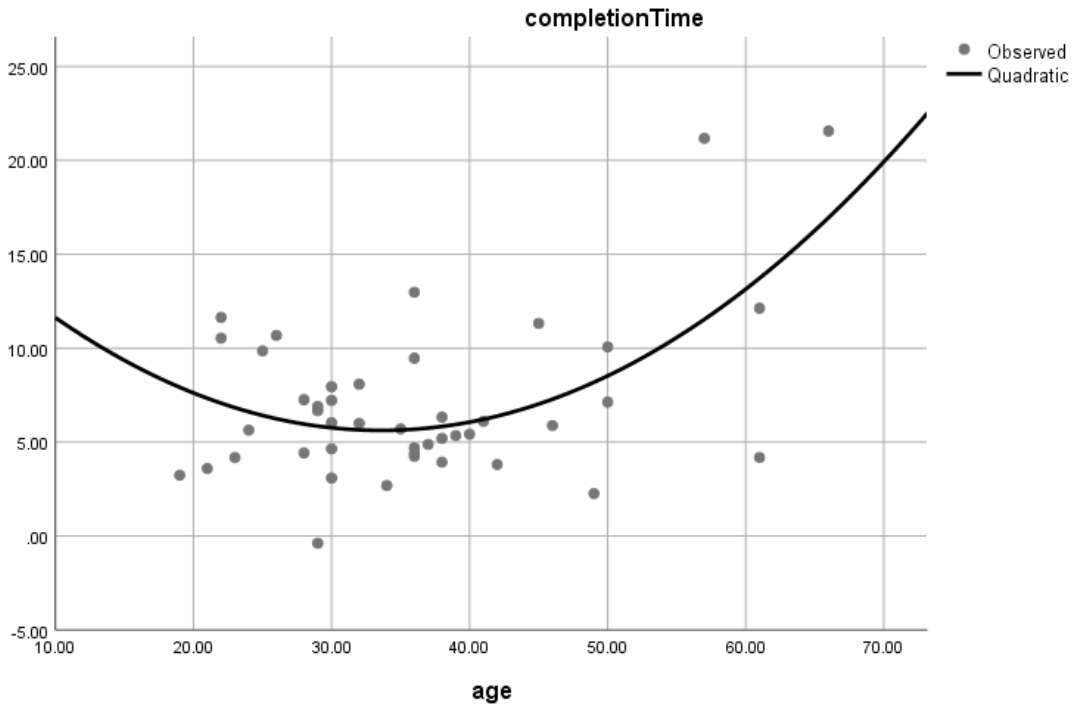


Figure 24: Scatterplot of response time over age for 12r.

On the other hand, we cannot observe such a correlation when looking at the completion time of all conditions.

### Model Summary and Parameter Estimates

Dependent Variable: completionTime

Model Summary						Parameter Estimates		
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.023	2.146	2	181	.120	8.722	-.165	.002

The independent variable is age.

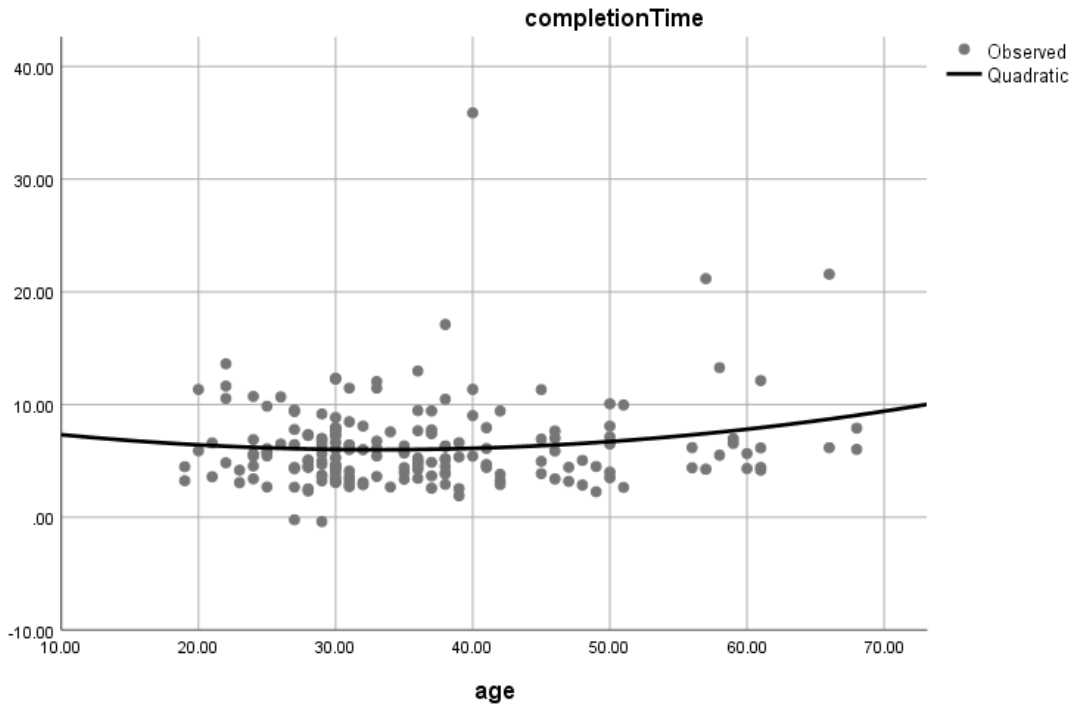


Figure 25: Scatterplot of response time over age for all conditions.

Given the fact that clock drawing is used to test cognitive abilities (see, for instance, Royall et al., 1998), the influence of age on the performance may be an indicator for a task that requires higher cognitive effort.

### Correlation between Number of Reported Observations and Task Performance

In the 24 conditions where users did not report any observations, 3 users committed an error in task 2 (12.5%). The remaining samples had an error rate of 18.1%.

#### anyObservations \* error Crosstabulation

Count

		error		Total
		.00	1.00	
anyObservations	0	21	3	24
	1	131	29	160
Total		152	32	184



Whether or not users had any observations in the first task also did not have any noticeable effect on the task completion time:

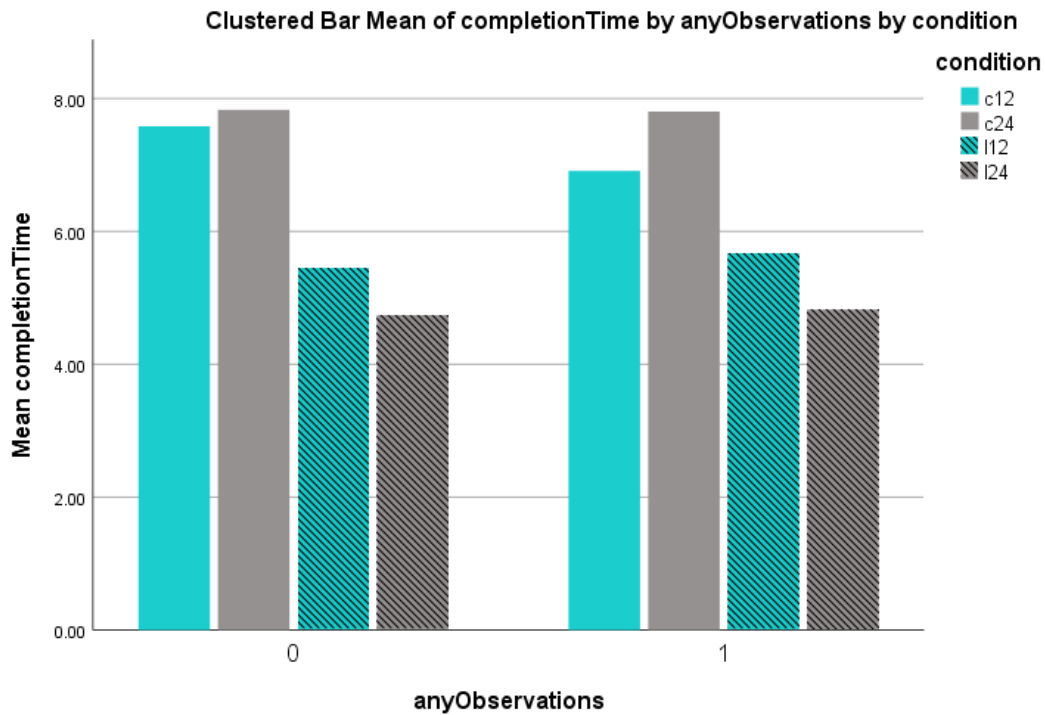


Figure 26: Mean completion times for users not reporting any observations in task 1 (0) and the others (1) per condition.

We can conclude that whether or not users reported observations did not have an influence on the task performance in task 2.

### Task 3: Read Value

#### Accuracy

From the 184 samples, 155 contained a valid input.

Case Processing Summary						
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
valueDiff * condition	155	84.2%	29	15.8%	184	100.0%

From these valid samples, most incorrect input values were  $\leq 10\%$  (the maximum value in our data sets was always around 40, i.e., 4). We marked the cases we counted as incorrect in red below. In total, these are 11 cases, where the user entered a value that deviated from the actual value of more than 4.

### valueDiff \* condition Crosstabulation

Count		condition				Total
		12l	12r	24l	24r	
valueDiff	.00	10	9	16	15	50
	1.00	12	9	16	15	52
	2.00	7	5	2	4	18
	3.00	4	1	6	4	15
	4.00	3	2	2	2	9
	5.00	1	0	0	0	1
	6.00	0	0	0	1	1
	7.00	1	0	0	0	1
	8.00	1	0	0	0	1
	9.00	0	1	0	0	1
	10.00	0	0	0	1	1
	16.00	0	1	0	0	1
	18.00	0	2	0	0	2
	30.00	0	1	0	0	1
	32.00	0	1	0	0	1
Total		39	32	42	42	155

12r had most invalid responses (37.5%), while the number of invalid responses was generally low for the other conditions (24r and 24l: 20%, 12l: 12%).

### invalidKey \* condition Crosstabulation

Count		condition				Total
		12l	12r	24l	24r	
invalidKey	.00	39	32	42	42	155
	1.00	5	12	6	6	29

Total	44	44	48	48	184
-------	----	----	----	----	-----

We therefore counted all value differences > 10% and invalid value responses as error.

### error \* condition Crosstabulation

Count		condition				Total
		12l	12r	24l	24r	
error	.00	36	26	42	40	144
	1.00	8	18	6	8	40
Total		44	44	48	48	184

### Exploratory Analysis of Error Cases

It is easily visible that 12r caused the highest deviations from the actual value:

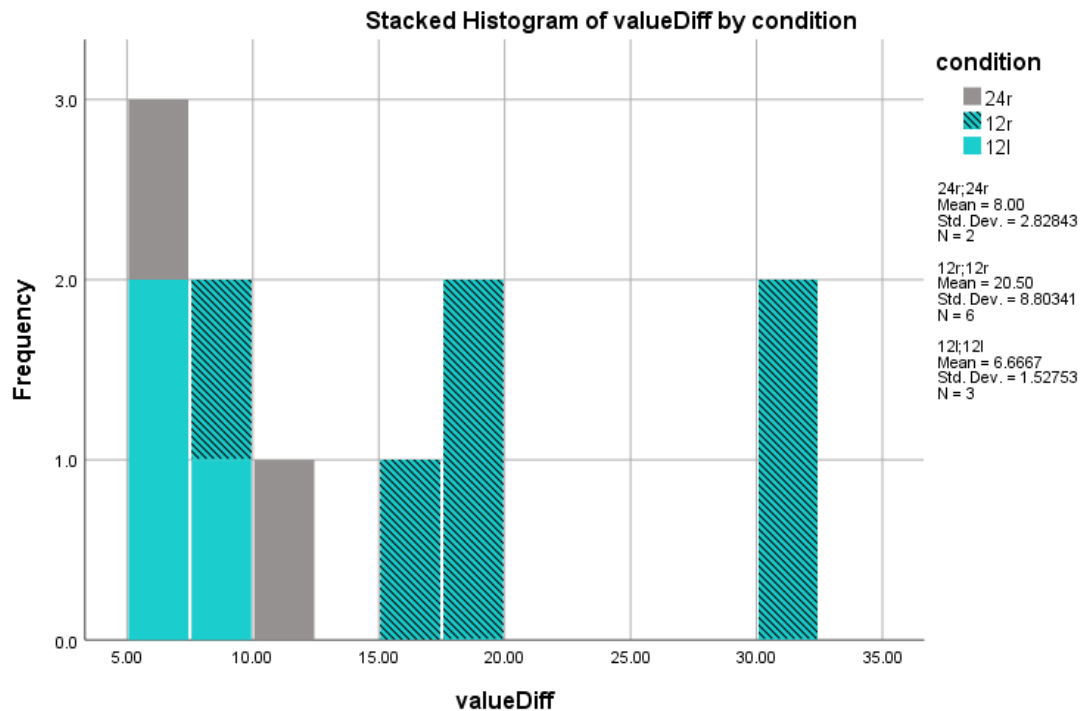


Figure 27: Histogram of deviations of user-entered values of actual values for each condition.

We reconstructed all 11 error cases and tried to infer potential reasons for the error. In a second round, we tested for each error case which reasons could explain the observed error. The potential reasons are:

- **Rounding error (round):** the input value lies within the two correct two grid lines and could be caused by a rounding error. Rounding errors can explain all error cases in 12l, one case for 12r, and one case for 24r.
- **Grid line misinterpretation (grid):** The input value does not lie within the two correct grid lines, but with different grid lines, the input can be explained. For the potential grid line misinterpretation cases, we add the grid line steps that could explain the input. Grid line misinterpretation can be an explanation for two error cases for 12r, and for one using 24r.
- **Closest bar to value axis (axis):** Instead of reading the highlighted bar, the user could have read the value of one of the bars next to the value axis. For the potential cases, we report the interval that could have been read. Reading the wrong bar can be an explanation for three error cases for 12r.

12r:

Input 7 (value = 25, error = 18): grid (5) or axis (12-1 a.m.)

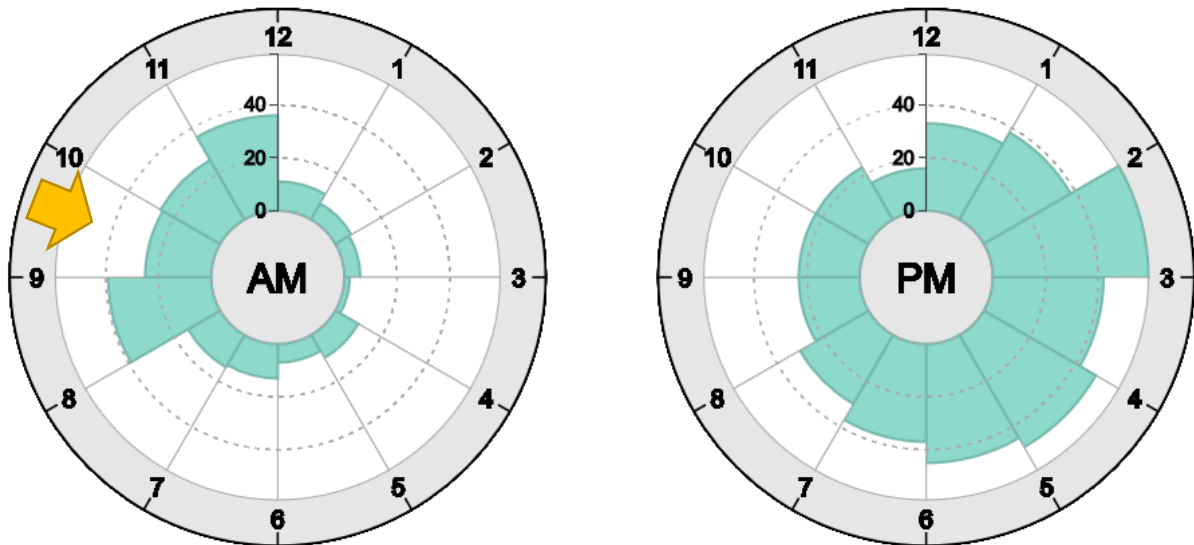


Figure 28: User selected value 7, which can be caused by an incorrect grid interval assumption (5) or by reading the value in the segment 12-1 in the a.m. chart.

Input 35 (value = 19, error = 16): axis (11-12 a.m.)

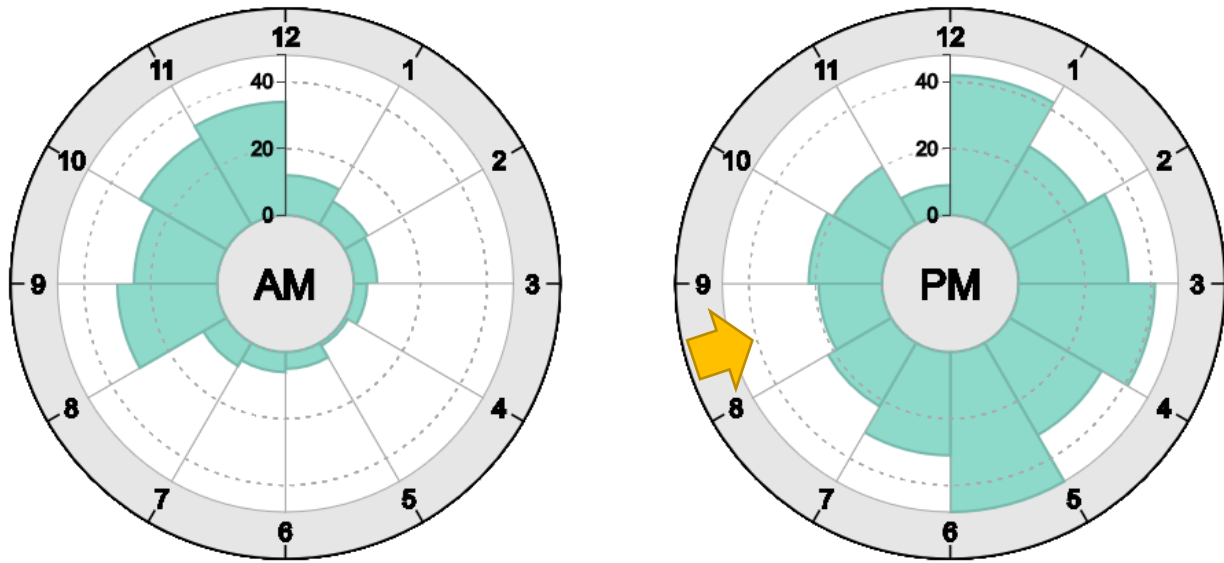


Figure 29: User entered value 35, which can be caused by reading the value of segment 11-12 in the a.m. chart.

Input 20 (value = 38, error = 18): grid (10)

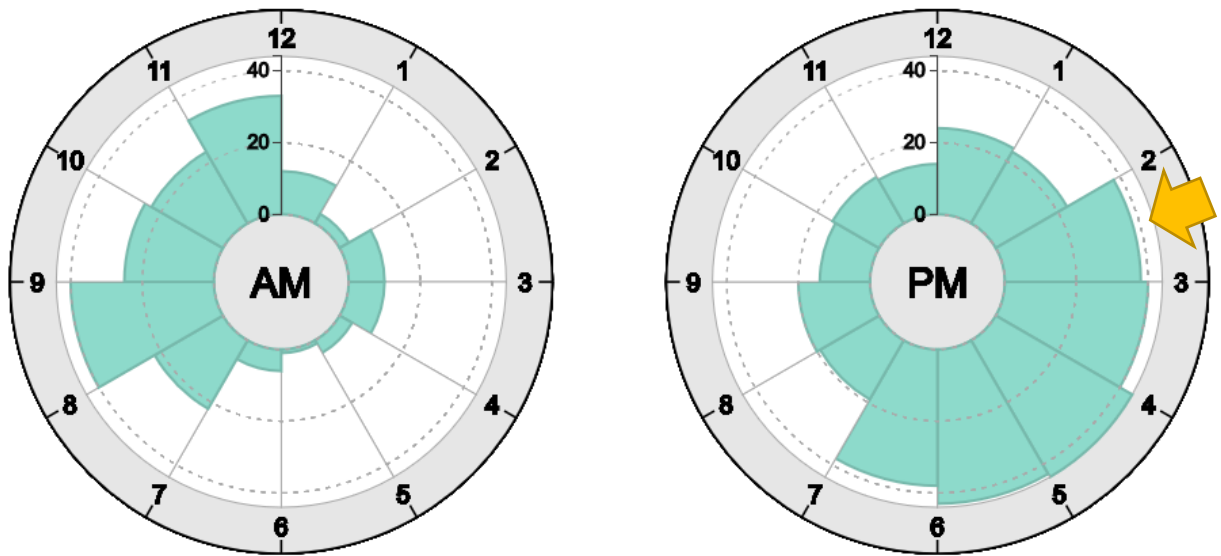


Figure 30: User entered value 20, which can be caused by an incorrect grid interval assumption (10).

Input 37 (value = 5, error = 32): axis (11-12 a.m. or 12-1 p.m.)

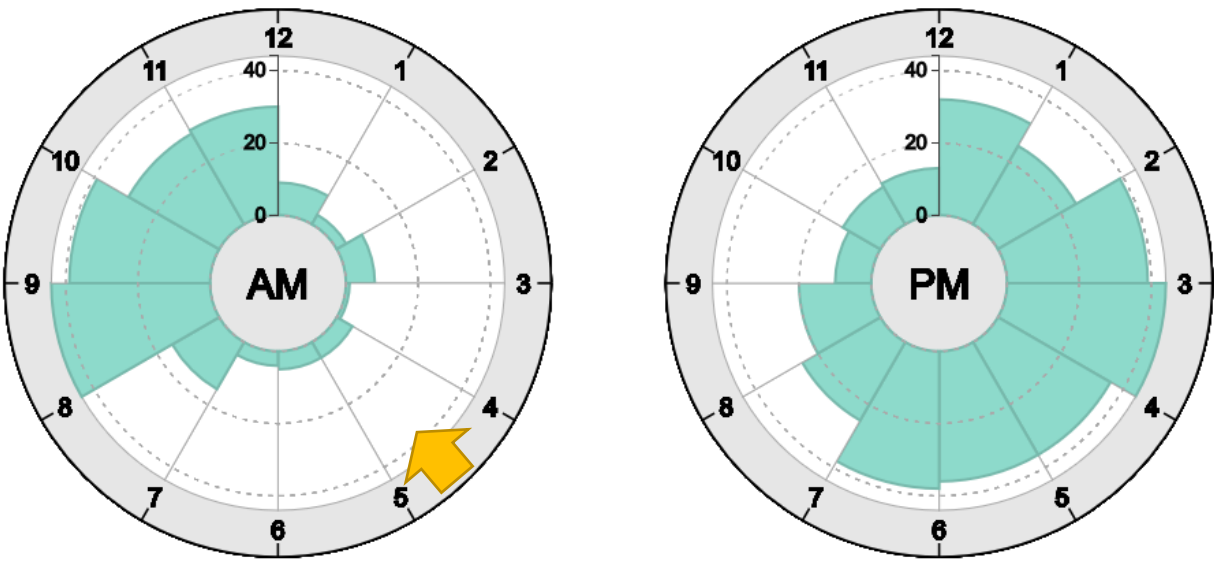


Figure 31: User entered value 37, which can be caused by reading segment 11-12 on the a.m. chart or 12-1 at the p.m. chart.

Input 37 (value = 28, error = 9): rounding

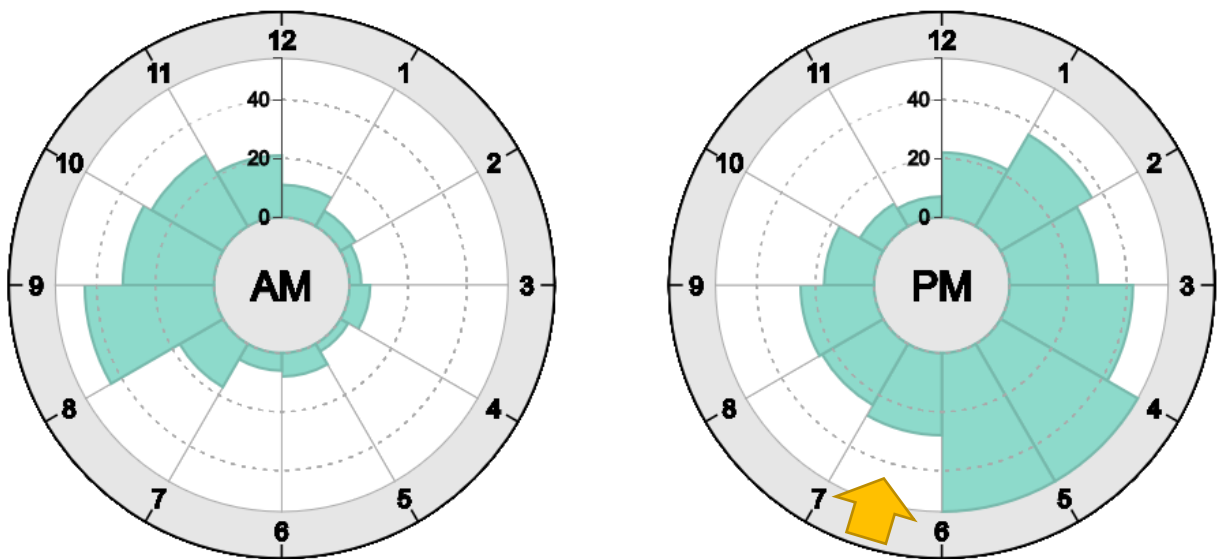


Figure 32: User entered value 37, which can be a rounding error.

Input 2 (value = 32, error = 30): grid (1) ?

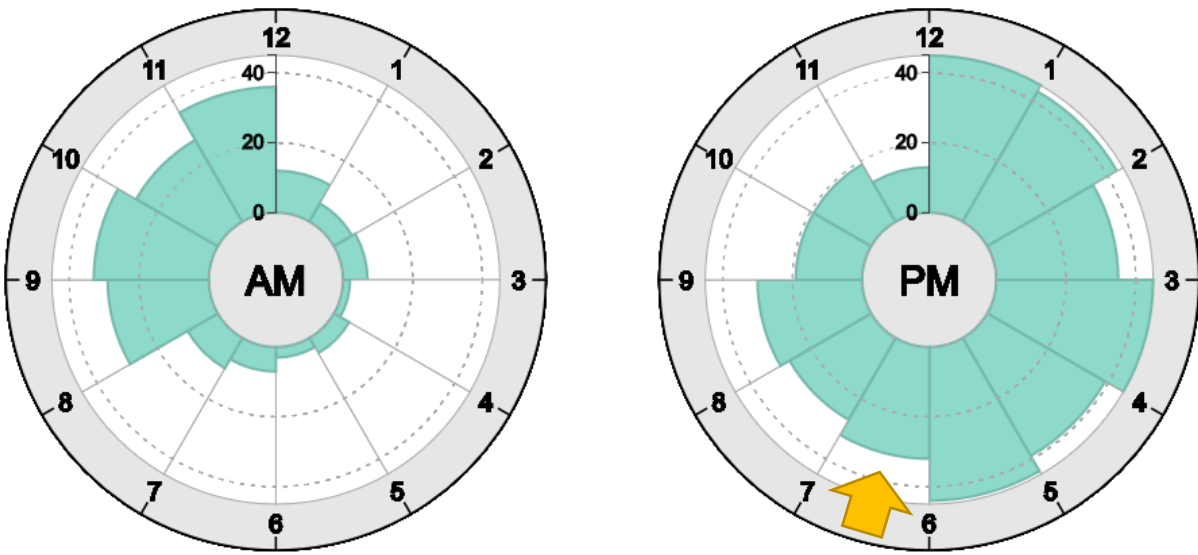


Figure 33: User entered value 2, which could be explained by an incorrectly assumed grid interval of 1.

12l:

Input 10 (value = 5, error = 5): rounding

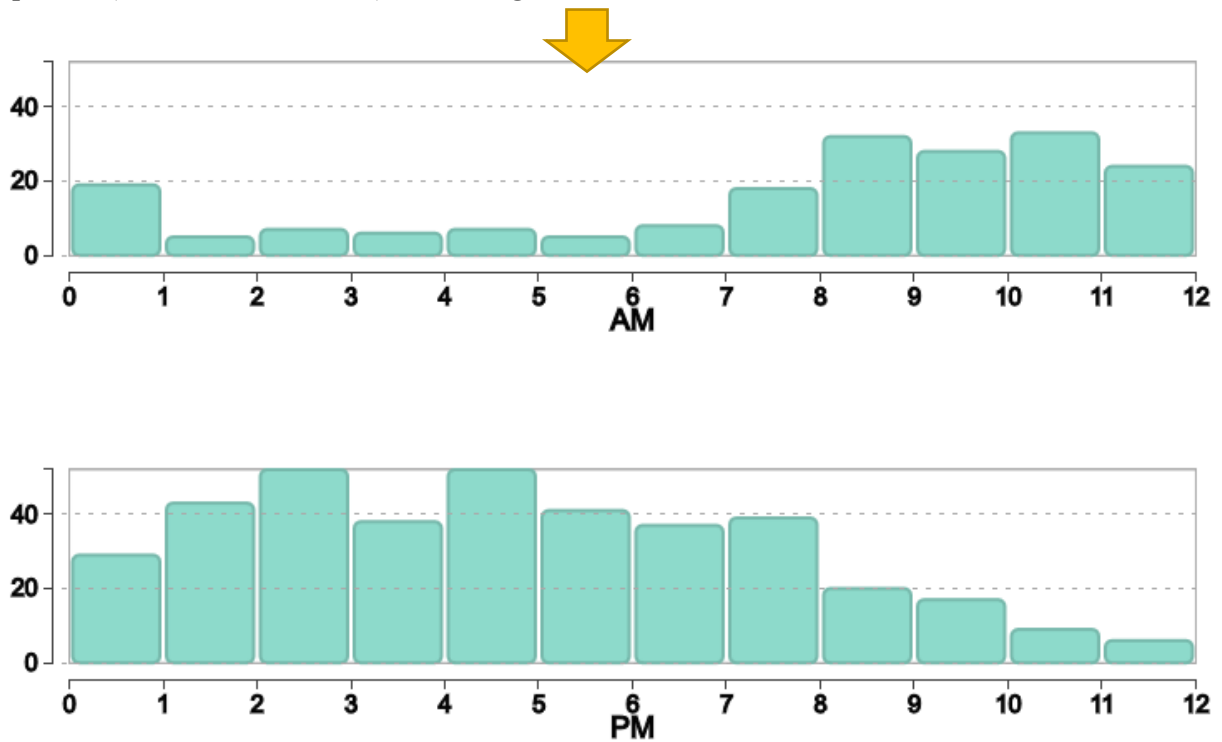


Figure 34: User entered value 10, which can be explained by a rounding error.

Input 10 (value = 2, error = 8): rounding

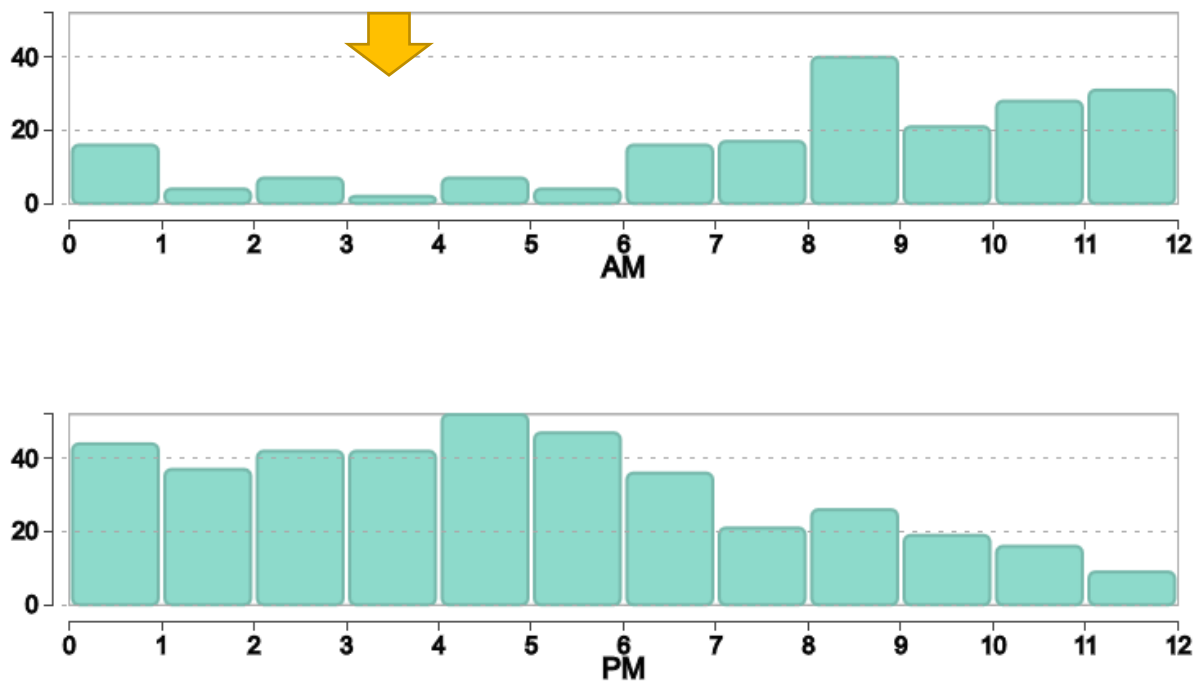


Figure 35: User entered value 10, which can be explained by a rounding error.



Input 50 (value = 57, error = 7): rounding

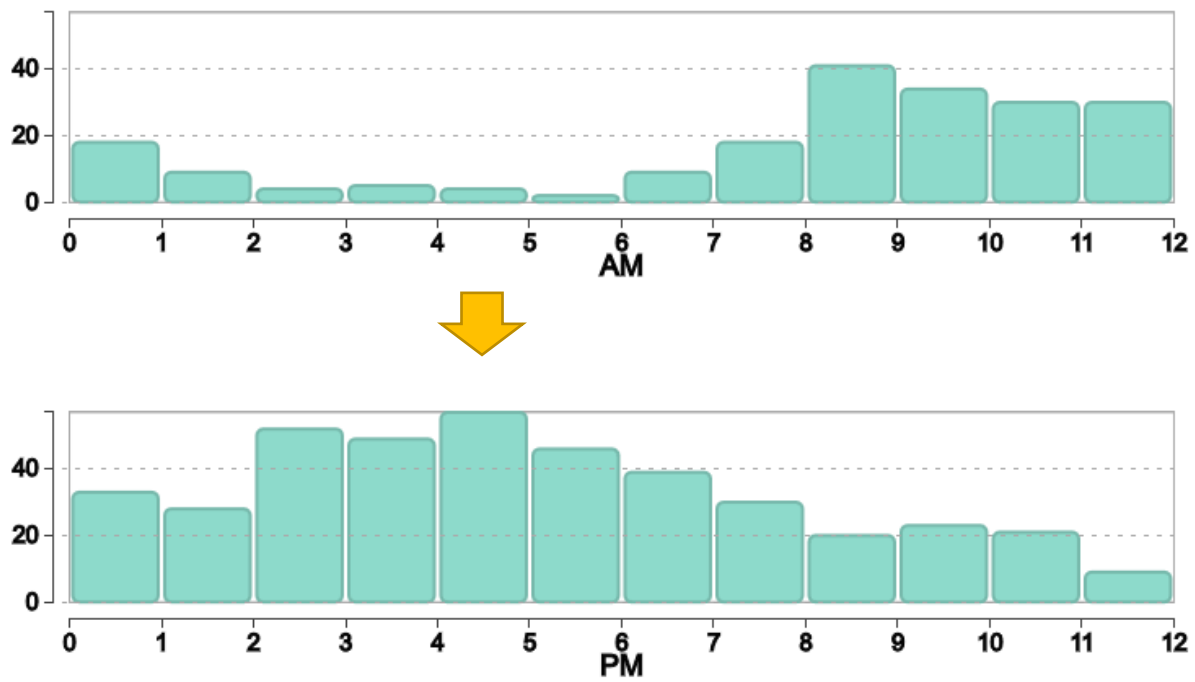


Figure 36: User entered value 50, which can be explained by a rounding error.

24r: Input 20 (value = 30, error = 10): grid (10); axis (12-1 a.m.)

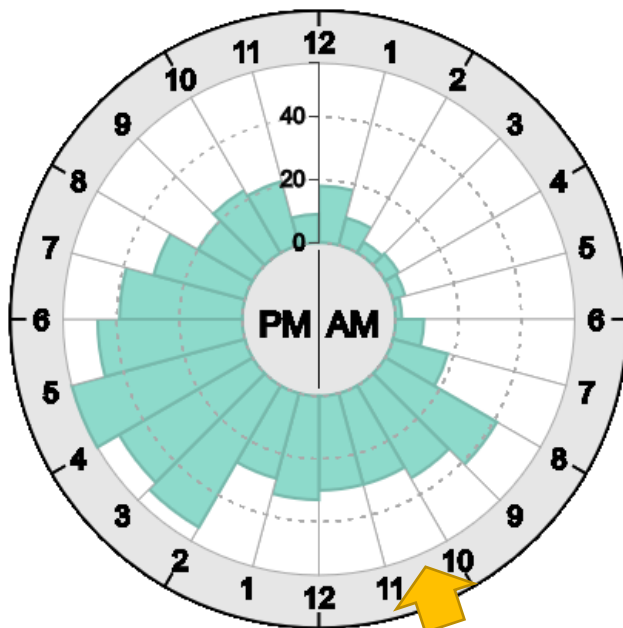


Figure 37: User entered value 20, which can be explained by an incorrectly assumed grid interval of 10 or by reading the segment 12-1 a.m.

Input 60 (value = 54, error = 60): rounding

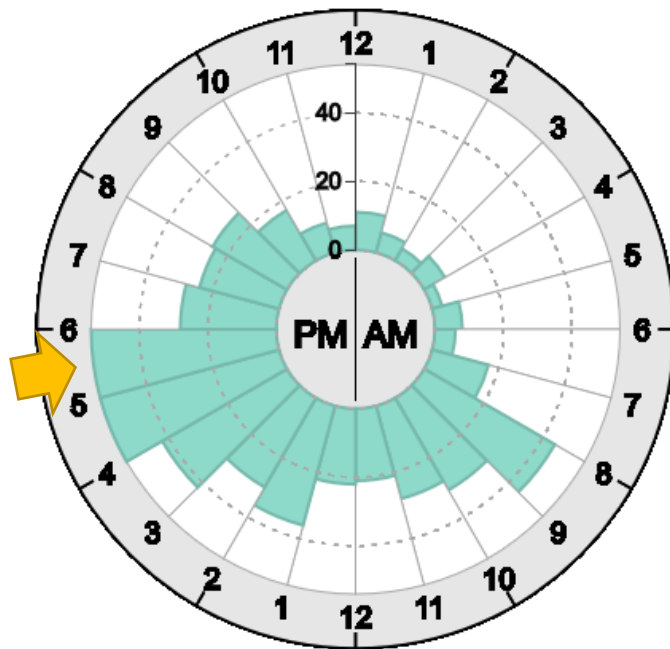


Figure 38: User entered value 60, which can be explained by a rounding error.

## Completion Time

We removed all records of users with at least one error, leaving us with 64 user responses, i.e., 128 samples.

We performed an ANOVA on log-transformed completion times with layout as within-subjects factor and cardinality as between-subjects factor.

We did not find a main effect for cardinality:

## Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	130.758	1	130.758	1969.629	.000	.969
<b>cardinality</b>	<b>.015</b>	<b>1</b>	<b>.015</b>	<b>.220</b>	<b>.641</b>	<b>.004</b>
Error	4.116	62	.066			

We found a large main effect for layout but no interaction between the two factors:

### Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
<b>layout</b>	<b>Sphericity Assumed</b>	<b>.447</b>	<b>1</b>	<b>.447</b>	<b>21.064</b>	<b>.000</b>	<b>.254</b>
	Greenhouse-Geisser	.447	1.000	.447	21.064	.000	.254
	Huynh-Feldt	.447	1.000	.447	21.064	.000	.254
	Lower-bound	.447	1.000	.447	21.064	.000	.254
<b>layout * cardinality</b>	<b>Sphericity Assumed</b>	<b>.009</b>	<b>1</b>	<b>.009</b>	<b>.404</b>	<b>.527</b>	<b>.006</b>
	Greenhouse-Geisser	.009	1.000	.009	.404	.527	.006
	Huynh-Feldt	.009	1.000	.009	.404	.527	.006
	Lower-bound	.009	1.000	.009	.404	.527	.006
Error(layout)	Sphericity Assumed	1.316	62	.021			
	Greenhouse-Geisser	1.316	62.000	.021			
	Huynh-Feldt	1.316	62.000	.021			
	Lower-bound	1.316	62.000	.021			

### Correlation between Number of Reported Observations and Task Performance

The fraction of errors is only marginally higher for conditions without reported observations (25%) than for the remaining conditions (21.2%):

### anyObservationsTask0 \* error Crosstabulation

Count

		error		Total
		.00	1.00	
anyObservationsTask0	0	18	6	24
	1	126	34	160
Total		144	40	184

The task completion times also do not differ considerably between the two groups:

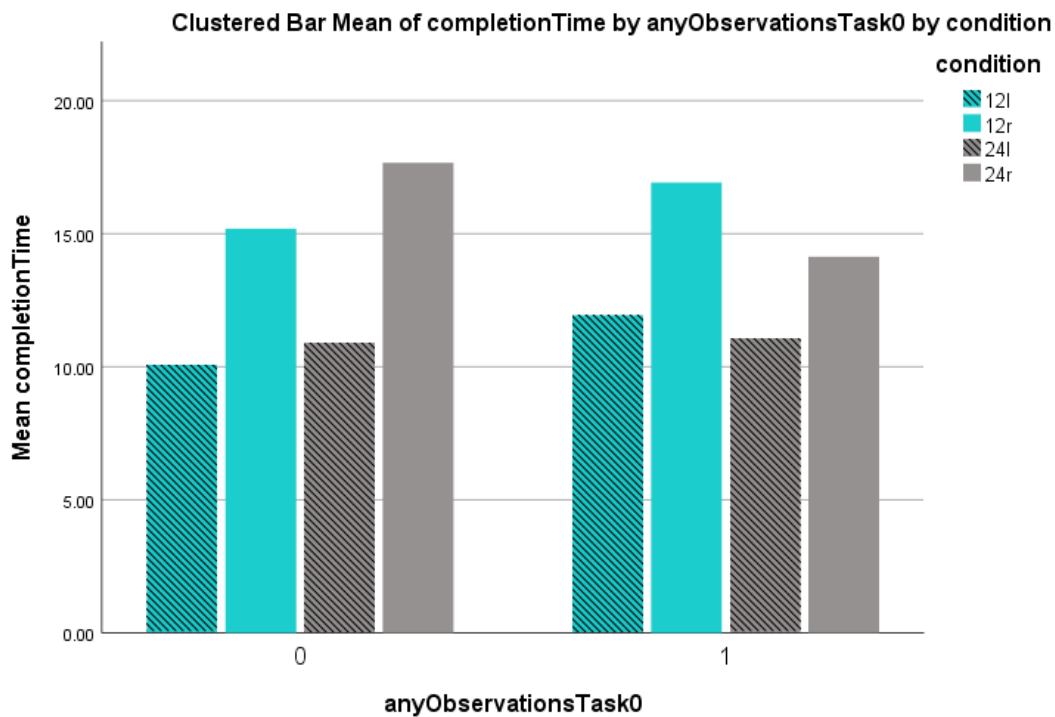


Figure 39: Mean completion times for trials without any observations in task 1 (0) and the others (1).

We can conclude that users that did not report any observations in task 1 did not perform worse in task 3 than the other users.

#### Task 4: Locate Maximum

##### Accuracy

From all 184 cases, 2 responses were invalid and excluded from further analysis:

##### Case Processing Summary

	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
valueDiff * condition	182	98.9%	2	1.1%	184	100.0%

For the remaining 182 cases, there is a total of 170 correct responses (i.e., the difference between the selected maximum value and the actual maximum value is 0). For 24l and 24r, all recorded responses were correct. For 12l, 88% of the responses were correct, for 12r 84%.

#### valueDiff \* condition Crosstabulation

Count		condition				Total
		12l	12r	24l	24r	
valueDiff	.00	38	37	47	48	170
	1.00	0	1	0	0	1
	8.00	0	1	0	0	1
	9.00	1	1	0	0	2
	10.00	2	1	0	0	3
	11.00	1	1	0	0	2
	12.00	1	1	0	0	2
	20.00	0	1	0	0	1
	Total	43	44	47	48	182

#### Exploratory Analysis of Error Cases

We analyzed all 12 incorrect responses in task 4 by looking again at the visualizations with the corresponding data sets. We then classified the reasons for incorrect responses into three categories:

- **AM/PM:** the maximum value was selected from the incorrect chart. In all cases, this was the AM chart, where the maximum value was lower than in the PM chart. This was the most common case with 5 occurrences for 12l and 4 for 12r.
- **2<sup>nd</sup>:** the second-highest value was selected. This was the case twice for 12r: once, the difference between the two maximum values was just 2%, but the second case it was 14%.
- **Noon:** the highest value next to noon was selected. We found one case for 12r.

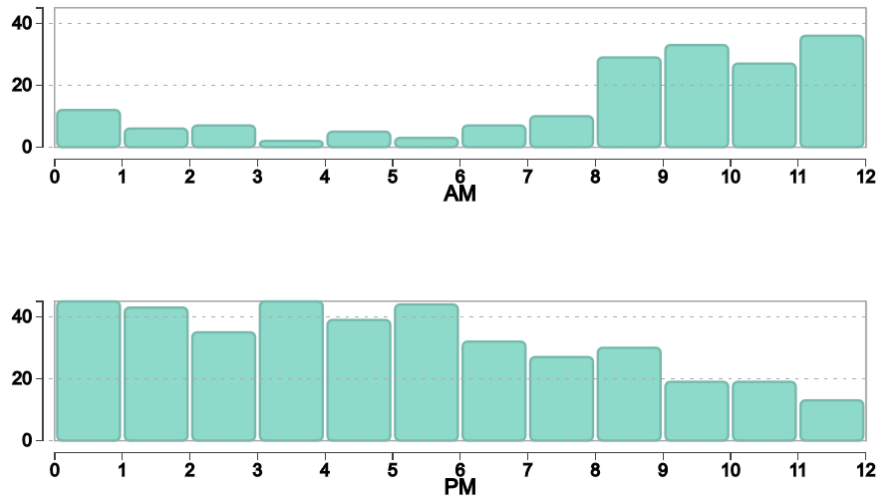


Figure 40: AM/PM: The user selected 11-12 a.m. instead of 12-1 p.m. as maximum.

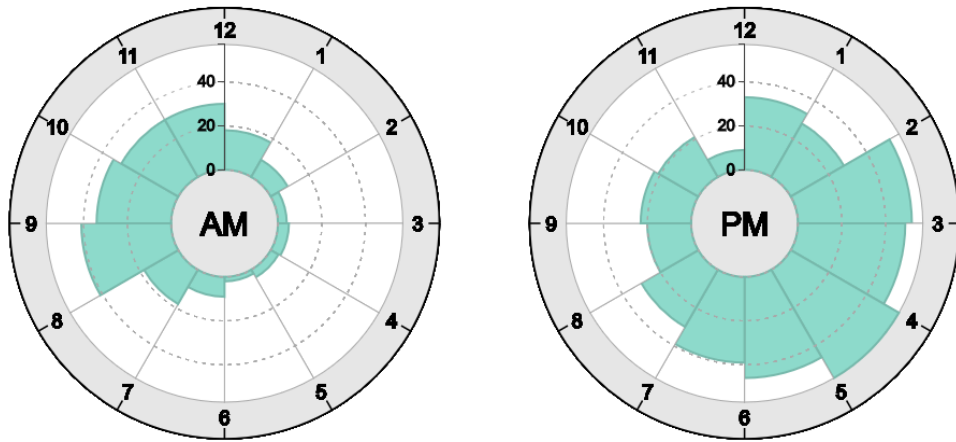


Figure 41: 2<sup>nd</sup>: The user selected 2-3 p.m. instead of 4-5 p.m. as maximum.

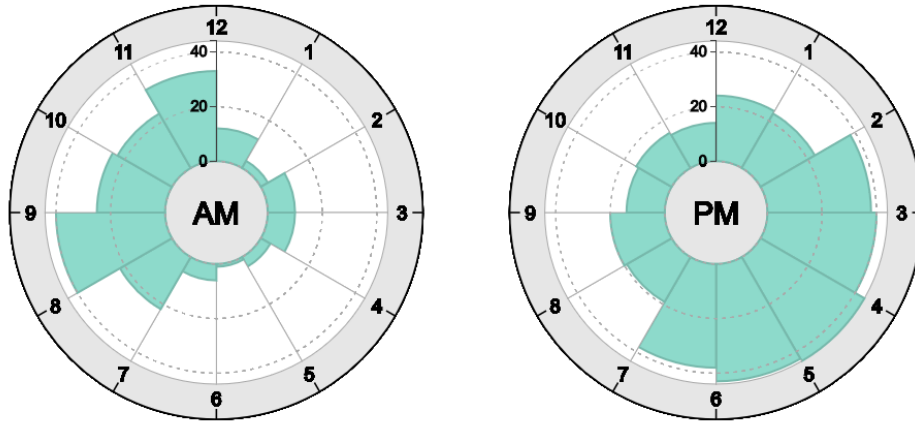


Figure 42: Noon: The user selected 11-12 a.m. instead of 4-5 p.m. as maximum.

## Completion Time

We removed all user records with at least one incorrect response, leaving 81 user records, i.e., 162 samples.

We performed an ANOVA, which showed a small main effect for cardinality:

### Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	52.816	1	52.816	1235.674	.000	.940
<b>cardinality</b>	<b>.245</b>	<b>1</b>	<b>.245</b>	<b>5.721</b>	<b>.019</b>	<b>.068</b>
Error	3.377	79	.043			

We also found a small main effect for layout, but no interaction between layout and cardinality:

### Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
<b>layout</b>	<b>Sphericity Assumed</b>	<b>.113</b>	<b>1</b>	<b>.113</b>	<b>7.507</b>	<b>.008</b>	<b>.087</b>
	Greenhouse-Geisser	.113	1.000	.113	7.507	.008	.087
	Huynh-Feldt	.113	1.000	.113	7.507	.008	.087
	Lower-bound	.113	1.000	.113	7.507	.008	.087
<b>layout * cardinality</b>	<b>Sphericity Assumed</b>	<b>.041</b>	<b>1</b>	<b>.041</b>	<b>2.742</b>	<b>.102</b>	<b>.034</b>
	Greenhouse-Geisser	.041	1.000	.041	2.742	.102	.034
	Huynh-Feldt	.041	1.000	.041	2.742	.102	.034
	Lower-bound	.041	1.000	.041	2.742	.102	.034
Error(layout)	Sphericity Assumed	1.193	79	.015			
	Greenhouse-Geisser	1.193	79.000	.015			
	Huynh-Feldt	1.193	79.000	.015			
	Lower-bound	1.193	79.000	.015			

### Correlation between Number of Reported Observations and Task Performance

Users not reporting any observations in the first task did not commit any errors in task 4:

#### anyObservationTask0 \* error Crosstabulation

Count

		error		Total
		.00	1.00	
anyObservationTask0	0	24	0	24
	1	146	14	160
Total		170	14	184

We can observe a slightly shorter task completion time for the users reporting observations in the condition 24l compared to those 5 users without any observations:



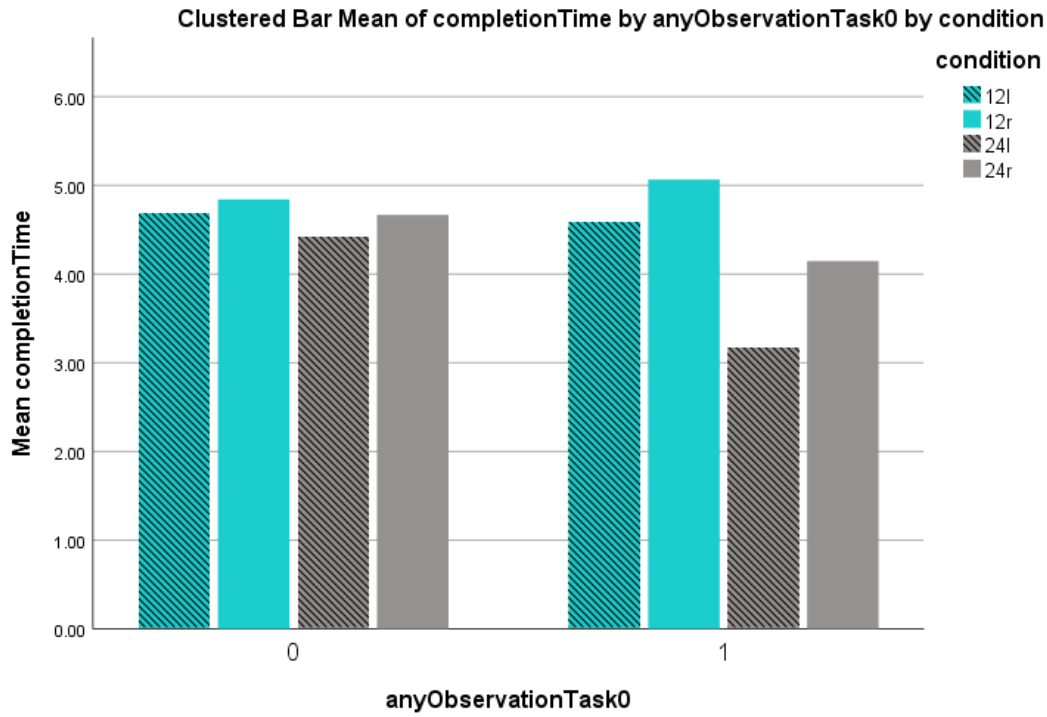


Figure 43: Mean task completion time for users not reporting any observations in task 1 (0) and the others (1).

Looking at the distribution of completion times for 24l only, we can see two outlier samples between 6 and 8 seconds:

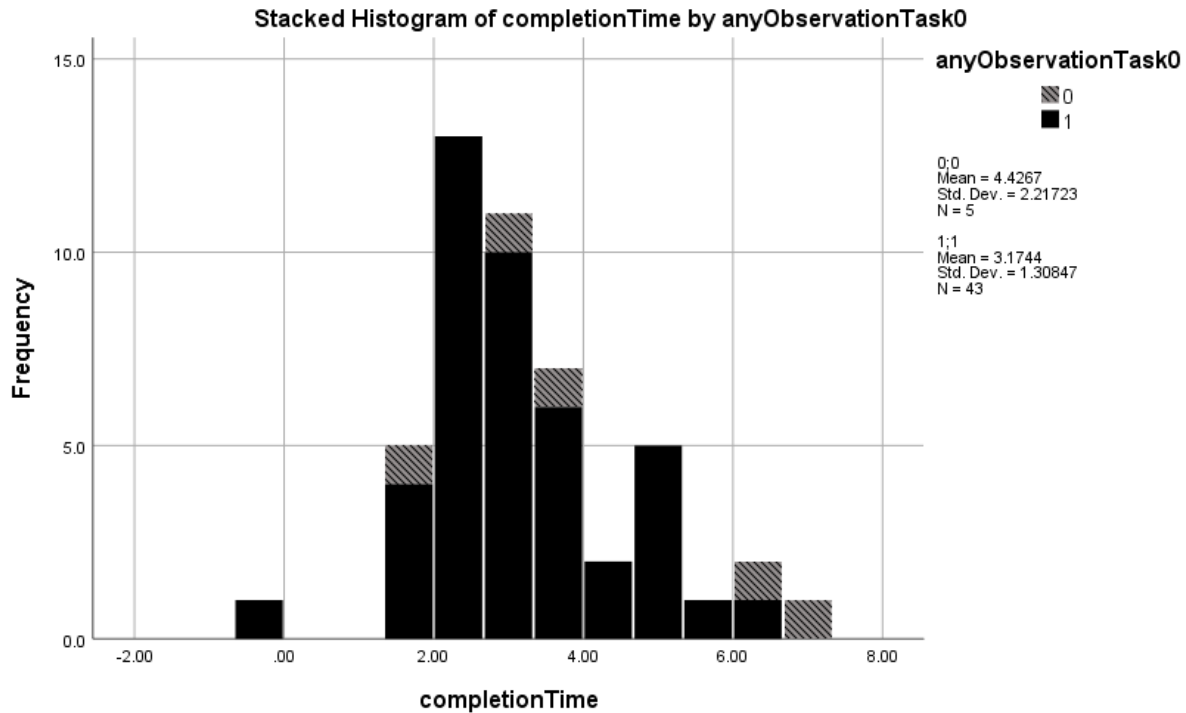


Figure 44: Histogram of task completion times (in seconds) for 24l, separated into users reporting observations in task 1 (1, black bars) and those not reporting any observations (0, gray bars with stripes).

We conclude that users not reporting any observations could solve the task equally accurate as the other users. Two of these users were quite slow to find the maximum for 24l.

## Task 5: Compare A.M./P.M. Interval Values

### Accuracy

We first computed the error by comparing the sign of the actual difference with the sign expressed by the user through the radio buttons. We found the highest error for 12r (9%) and the lowest for 24l (2%).

### error \* condition Crosstabulation

Count		condition				Total
		12l	12r	24l	24r	
error	.00	41	40	47	46	174
	1.00	3	4	1	2	10
Total		44	44	48	48	184

### Qualitative Analysis of Error Cases

Over all conditions, we first looked at the incorrect user responses (-1 = fewer, 0 = equal, 1 = more) over the absolute value differences between the AM- and PM-value. We can observe that for all absolute differences between 1 and 7, the user incorrectly selected “equal”. In one case, the user incorrectly selected “more” for equal values.

### inputDiff \* absoluteActualDiff Crosstabulation

Count		absoluteActualDiff								Total
		.00	1.00	3.00	7.00	8.00	22.00	28.00	31.00	
inputDiff	-1	0	0	0	0	1	0	0	0	2
	0	0	2	1	1	0	0	0	0	4
	1	1	0	0	0	0	1	1	1	4
Total		1	2	1	1	1	1	1	1	10

First, we looked at those cases, where users selected the wrong radio button while having a significant value difference (from 8 to 49), which are 5 cases in total. These five cases were observed for 12l (1), 12r (2), and 24r (2):

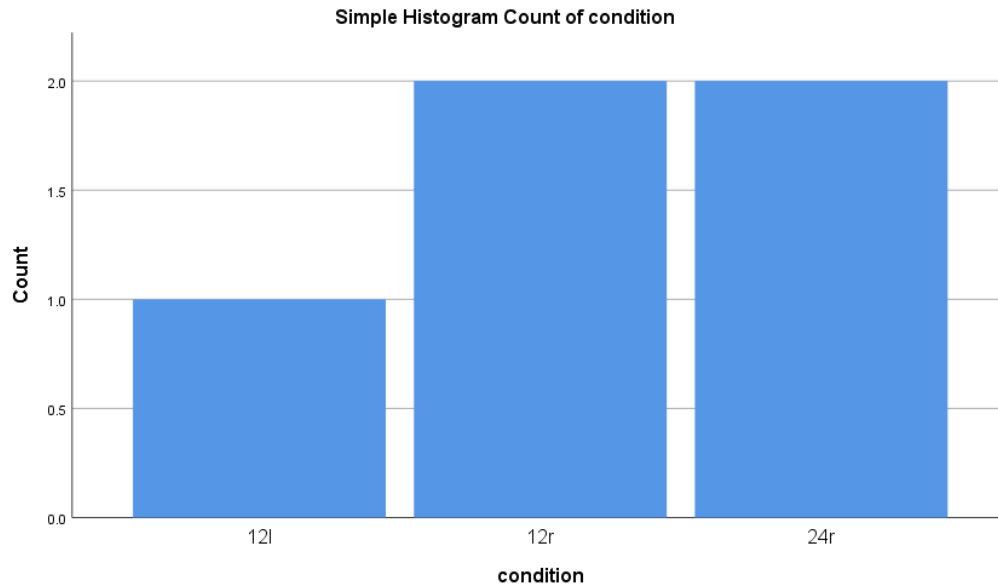


Figure 45: Number of incorrect user responses when the value differences between a.m. and p.m. were high ( $> 8$ ) per condition.

Second, we looked at those cases, where users thought the bars had the same height despite some difference (1-7). These cases are caused by 12-hours variants only:

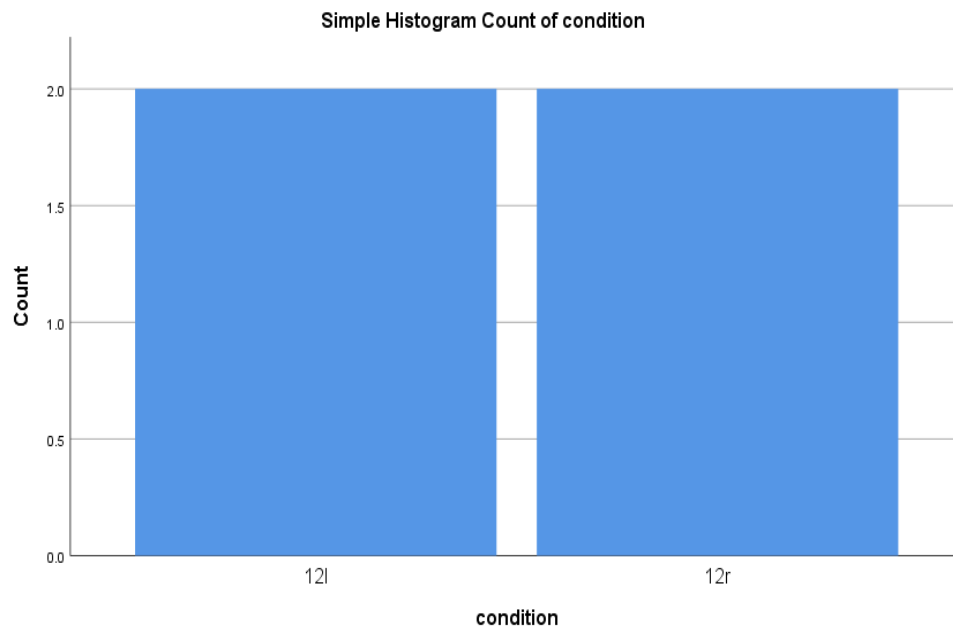


Figure 46: Number of incorrect "equal" responses per condition.

Below, there is one example of 12l with incorrectly assumed equal values (10-11 a.m. and p.m.), where the actual difference is 3:

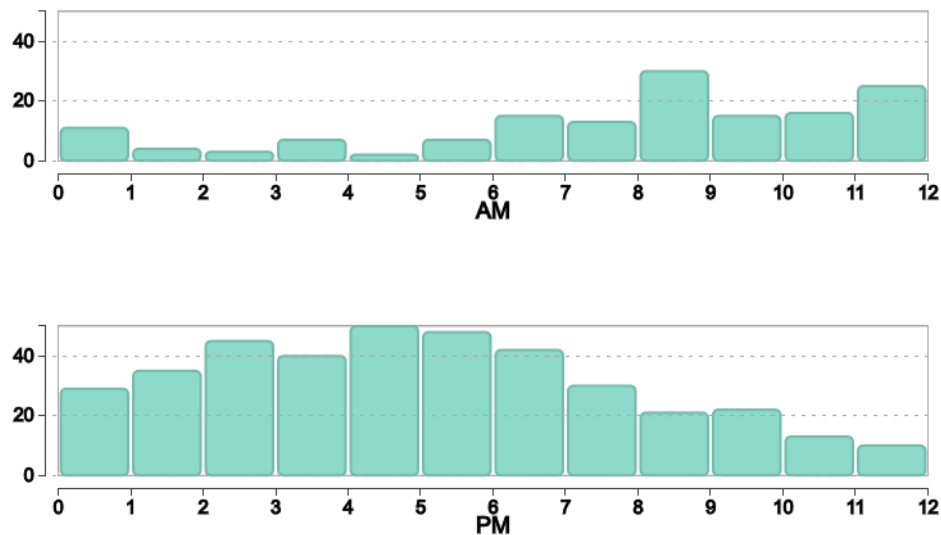


Figure 47: Incorrect "equal" response for interval 10-11 (actual difference is 3).

## Completion Time

We removed all user responses with incorrect responses in at least one of the layout conditions, leaving us with 82 user responses, i.e., 164 samples.

We did not find a main effect of cardinality:

### Tests of Between-Subjects Effects

Measure: logCompletionTime

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	122.806	1	122.806	3111.187	.000	.975
cardinality	.009	1	.009	.216	.644	.003
Error	3.158	80	.039			

However, we found a medium-sized main effect for layout and a medium-sized interaction between layout and cardinality:

### Tests of Within-Subjects Effects

Measure: logCompletionTime

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
layout	<b>Sphericity Assumed</b>	<b>.148</b>	<b>1</b>	<b>.148</b>	<b>11.516</b>	<b>.001</b>	<b>.126</b>
	Greenhouse-Geisser	.148	1.000	.148	11.516	.001	.126
	Huynh-Feldt	.148	1.000	.148	11.516	.001	.126
	Lower-bound	.148	1.000	.148	11.516	.001	.126
layout * cardinality	<b>Sphericity Assumed</b>	<b>.108</b>	<b>1</b>	<b>.108</b>	<b>8.393</b>	<b>.005</b>	<b>.095</b>
	Greenhouse-Geisser	.108	1.000	.108	8.393	.005	.095
	Huynh-Feldt	.108	1.000	.108	8.393	.005	.095
	Lower-bound	.108	1.000	.108	8.393	.005	.095
Error(layout)	Sphericity Assumed	1.025	80	.013			
	Greenhouse-Geisser	1.025	80.000	.013			
	Huynh-Feldt	1.025	80.000	.013			
	Lower-bound	1.025	80.000	.013			

### Correlation between Number of Reported Observations and Task Performance

The number of errors was 8.3% for those users not reporting any observations in task 1 and 5% for those reporting observations:

#### anyObservationTask0 \* error Crosstabulation

Count

		error		Total
		.00	1.00	
anyObservationTask0	0	22	2	24
	1	152	8	160
Total		174	10	184

Task completion times were comparable between the two groups. Only for 12r, we can observe a tendency that users performed slower when they had not reported any observations:

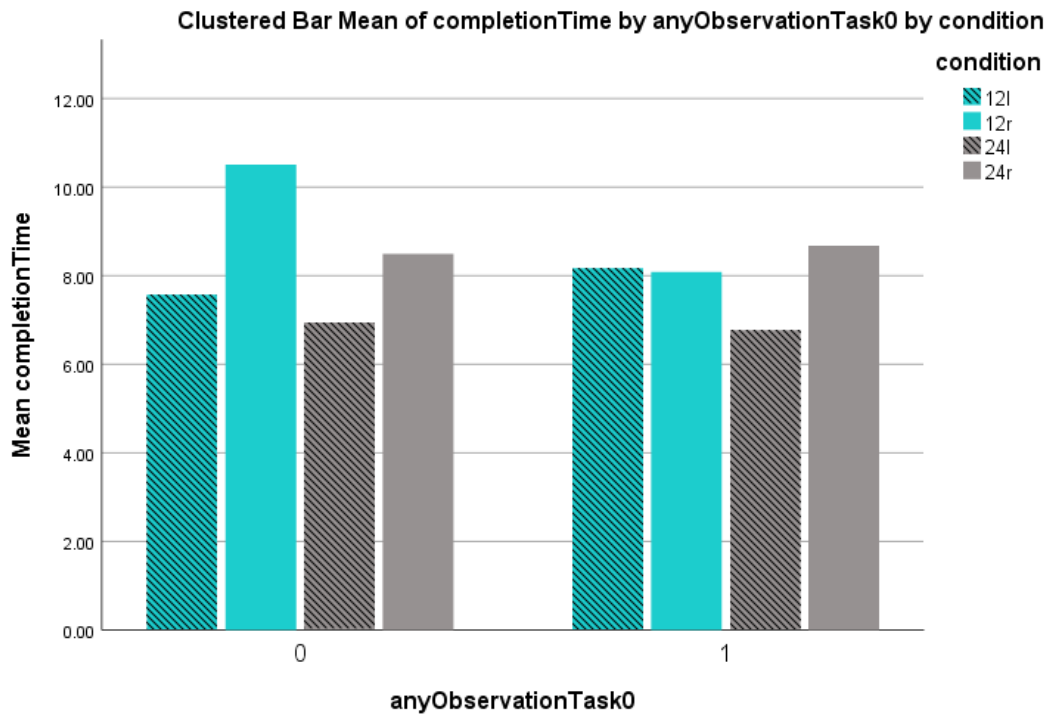


Figure 48: Mean task completion times for users not reporting any observations in task 1 (0) and the others (1) per condition.

Observing the completion time distribution between the two groups of users for 12r, no obvious performance difference between the two groups becomes visible:

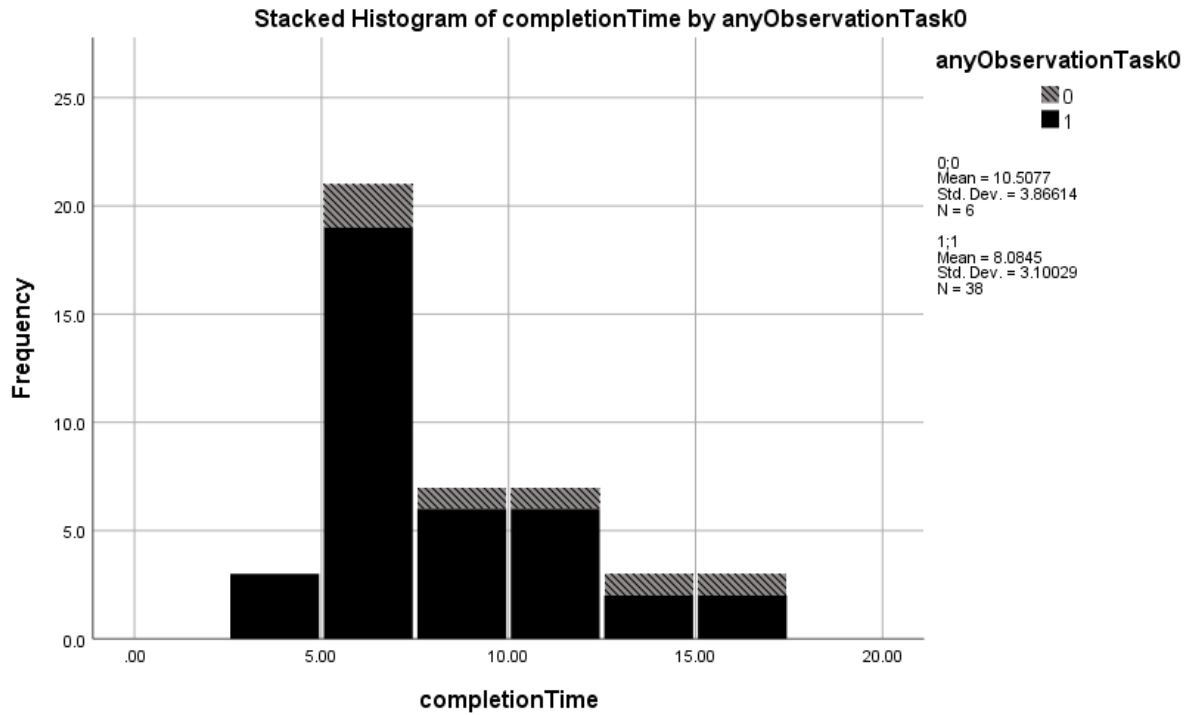


Figure 49: Histogram of completion times (in seconds) for 12r, split into users reporting observations in task 1 (black bars) and users not reporting any observations (gray bars with stripes).

We conclude that users not reporting any observations were performing similarly accurate and efficient as the other users in the comparison task.

## Task 6: Subjective Ratings

### Subjective Ratings

To compare the subjective ratings, we performed an ANOVA on Likert scale responses.

There is no main effect for cardinality:

### Tests of Between-Subjects Effects

Measure: rating

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	2390.135	1	2390.135	2448.128	.000	.965



<b>cardinality</b>	<b>.535</b>	<b>1</b>	<b>.535</b>	<b>.548</b>	<b>.461</b>	<b>.006</b>
Error	85.915	88	.976			

However, there is a large effect for layout, and no interaction between layout and cardinality:

### Tests of Within-Subjects Effects

Measure: rating

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
<b>layout</b>	<b>Sphericity Assumed</b>	<b>58.598</b>	<b>1</b>	<b>58.598</b>	<b>83.912</b>	<b>.000</b>	<b>.488</b>
	Greenhouse-Geisser	58.598	1.000	58.598	83.912	.000	.488
	Huynh-Feldt	58.598	1.000	58.598	83.912	.000	.488
	Lower-bound	58.598	1.000	58.598	83.912	.000	.488
<b>layout * cardinality</b>	<b>Sphericity Assumed</b>	<b>.109</b>	<b>1</b>	<b>.109</b>	<b>.156</b>	<b>.694</b>	<b>.002</b>
	Greenhouse-Geisser	.109	1.000	.109	.156	.694	.002
	Huynh-Feldt	.109	1.000	.109	.156	.694	.002
	Lower-bound	.109	1.000	.109	.156	.694	.002
Error(layout)	Sphericity Assumed	61.452	88	.698			
	Greenhouse-Geisser	61.452	88.000	.698			
	Huynh-Feldt	61.452	88.000	.698			
	Lower-bound	61.452	88.000	.698			

The stacked histogram shows the distribution of subjective ratings by condition:

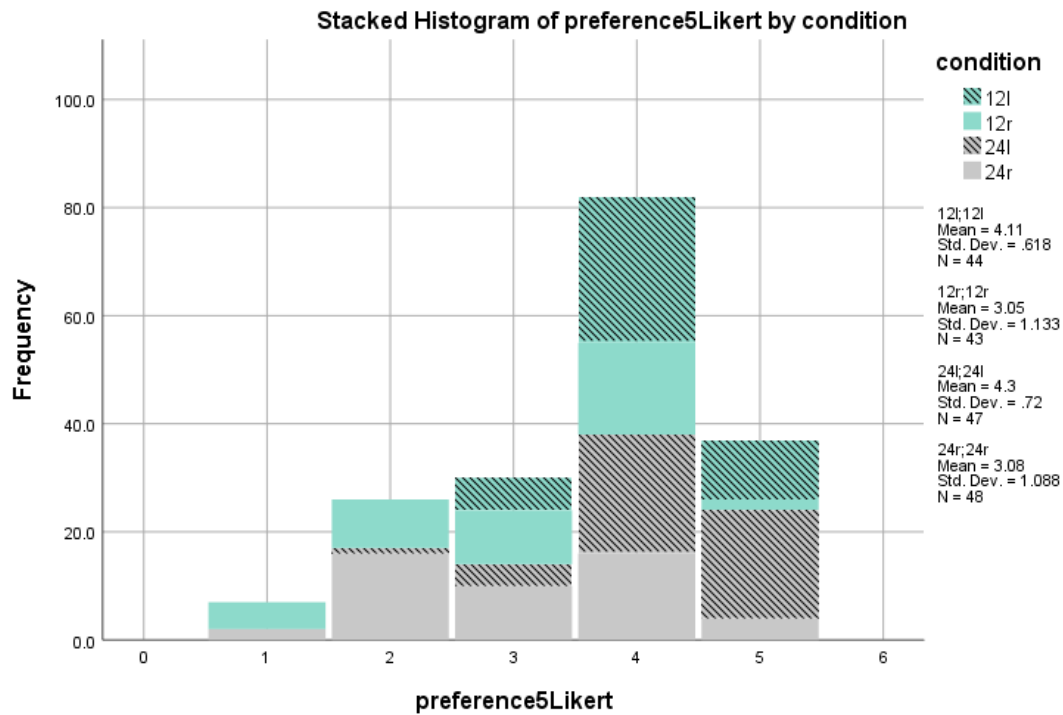


Figure 50: Histogram of subjective ratings per condition.

## Correlation between Demographics and Preferences

We analyzed the correlation between the preference ratings and the self-reported visualization literacy. There might be a small tendency that the acceptance for radial charts increases with increasing visualization literacy:

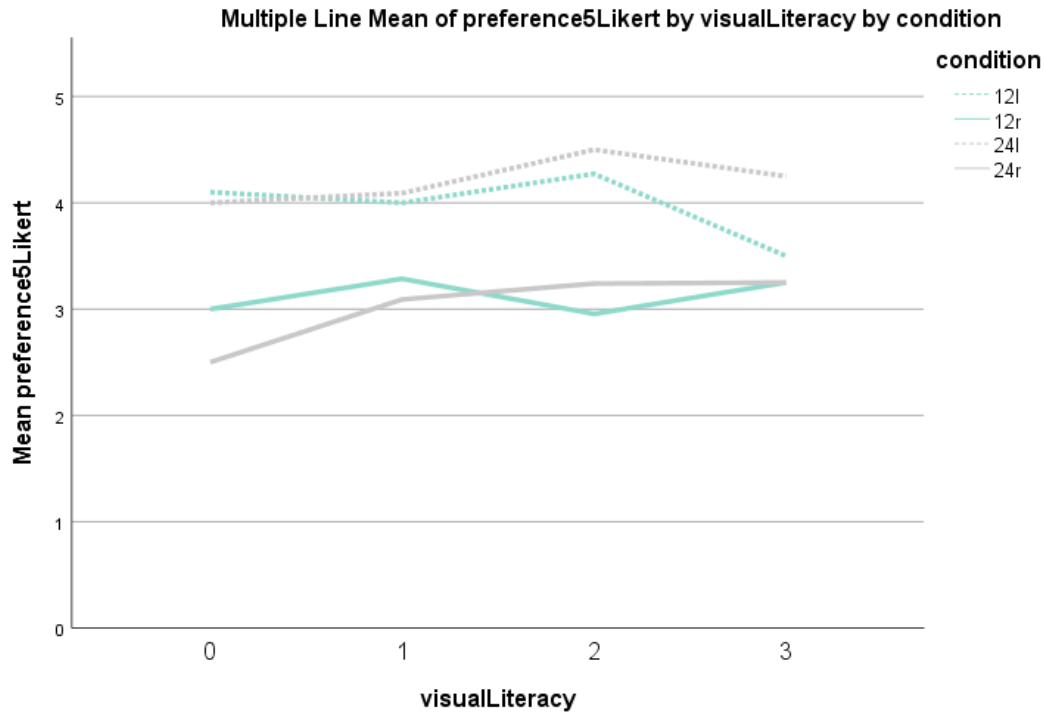


Figure 51: Subjective ratings of the four conditions by self-reported visualization literacy.

However, this interaction effect is not significant (Sig =  $p > .05$ ).

### Tests of Between-Subjects Effects

Dependent Variable: preference5Likert

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	68.791 <sup>a</sup>	15	4.586	5.389	.000
Intercept	1583.567	1	1583.567	1860.713	.000
visualLiteracy	3.117	3	1.039	1.221	.304
cardinality	.151	1	.151	.177	.674
layout	31.994	1	31.994	37.593	.000
visualLiteracy * cardinality	2.496	3	.832	.978	.405
visualLiteracy * layout	2.513	3	.838	.984	.402
cardinality * layout	.916	1	.916	1.076	.301
visualLiteracy * cardinality * layout	.785	3	.262	.308	.820
Error	141.275	166	.851		

Total	2618.000	182			
Corrected Total	210.066	181			

a. R Squared = .327 (Adjusted R Squared = .267)

There are no preference differences between genders:

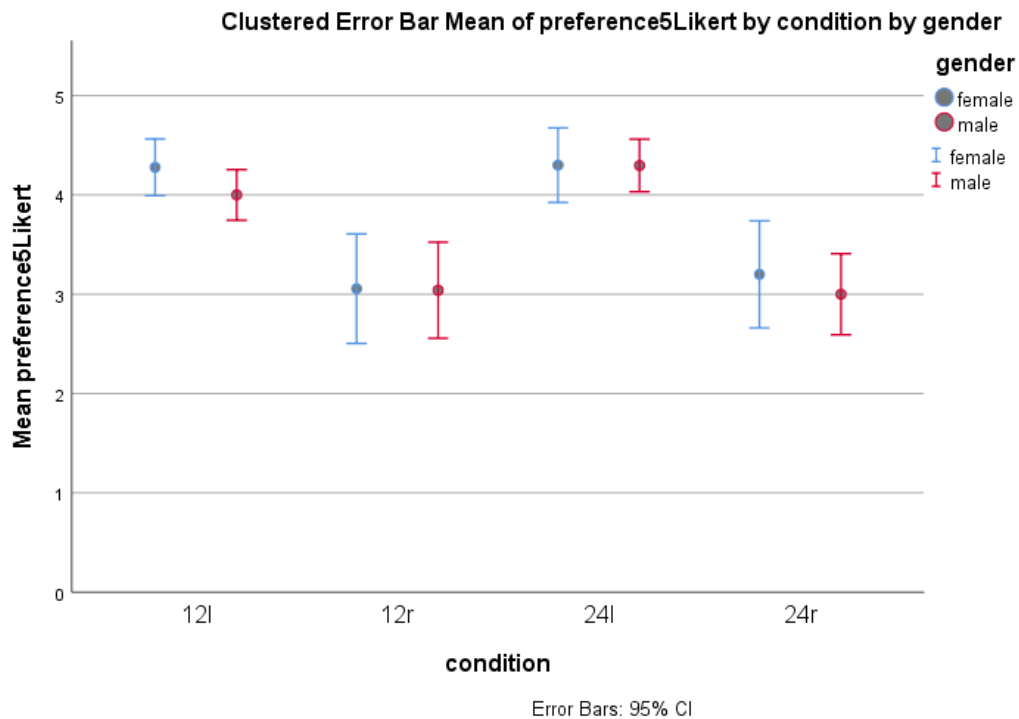


Figure 52: Subjective ratings of the four conditions separated by gender.

The distribution of age for each of the five subjective ratings for each condition. There is no obvious interaction between age and preference:

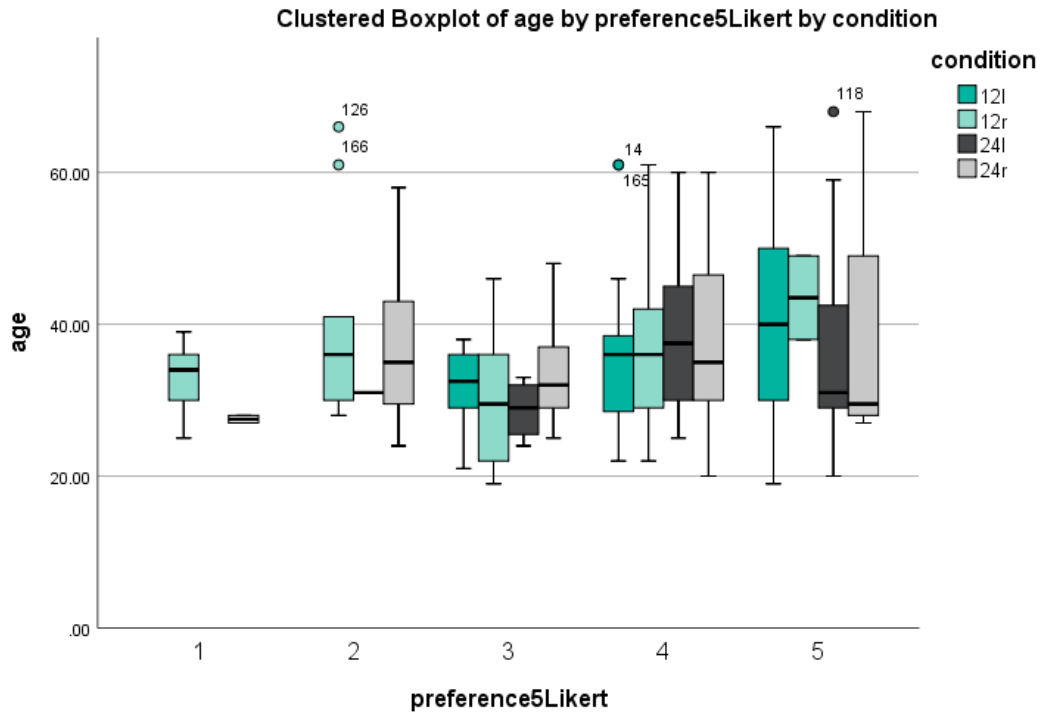


Figure 53: Box plots of age groups per each Likert-scale subjective rating (1-5) per condition.

There seems to be an interaction between the presentation order of the conditions and the finally assigned subjective ratings. When the linear condition was seen first (“l”), the subjective rating of the radial charts was lower than if the radial condition was seen first (“c”).

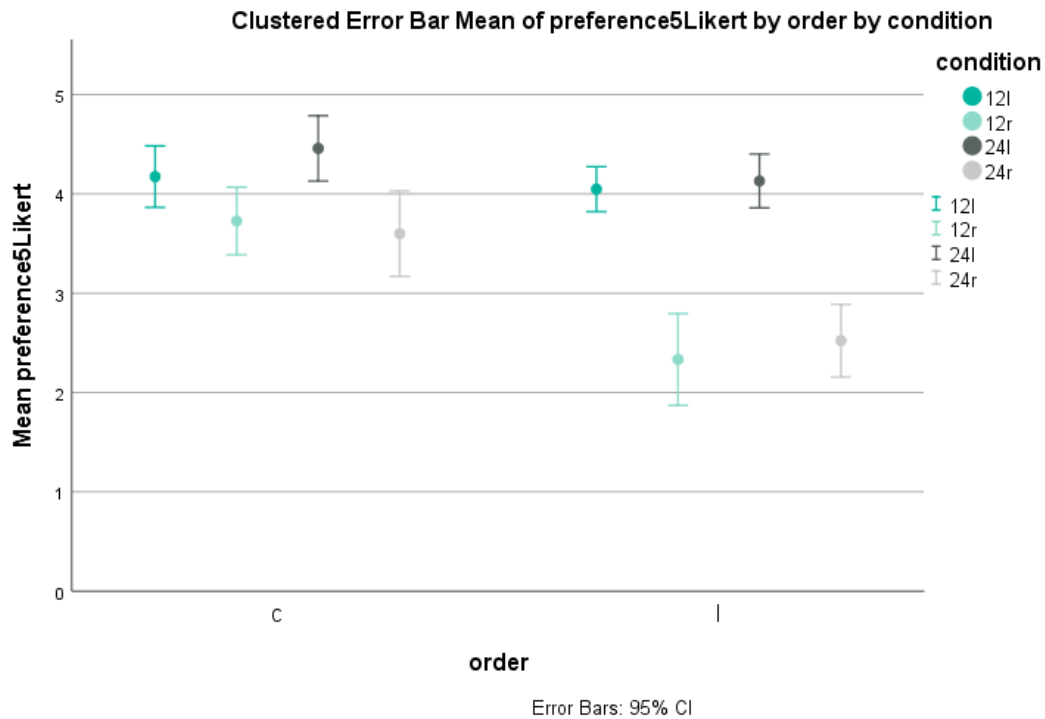


Figure 54: Subjective ratings per condition, split up by order (c=radial first, l=linear first).

Indeed, this interaction is significant:

### Tests of Between-Subjects Effects

Dependent Variable: preference5Likert

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	96.473 <sup>a</sup>	7	13.782	21.111	.000
Intercept	2382.618	1	2382.618	3649.651	.000
cardinality	.520	1	.520	.797	.373
layout	60.710	1	60.710	92.994	.000
order	24.274	1	24.274	37.183	.000
cardinality * layout	.266	1	.266	.407	.524
cardinality * order	.037	1	.037	.057	.812
layout * order	11.543	1	11.543	17.682	.000
cardinality * layout * order	.758	1	.758	1.162	.283

Error	113.593	174	.653		
Total	2618.000	182			
Corrected Total	210.066	181			

a. R Squared = .459 (Adjusted R Squared = .437)

## Correlation between Number of Reported Observations and Subjective Ratings

We plotted the subjective ratings separately for users reporting no observations in task 1 and all others.

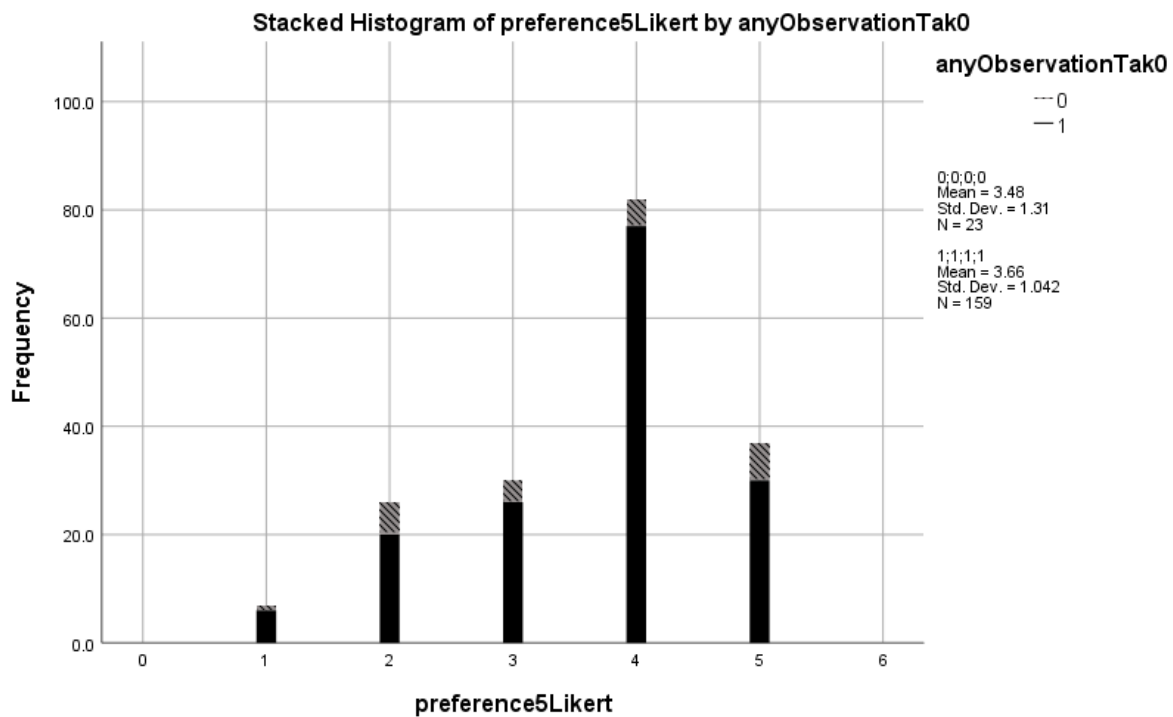


Figure 55: Histogram of subjective ratings (Likert scale 1-5) for users reporting at least one observation in task 1 (black bars) and users not reporting any observation (gray bars with stripes).

It is clearly visible that users not reporting any observations did not issue lower or higher subjective ratings than the other users.

## Optional Text Comments

We then analyzed the optional text comments. We received 66 text comments – most for 24r (21, 43% comment rate). Fewest comments were received for 12l (27%).

**textAvailable \* condition Crosstabulation**

Count		condition				Total
		12l	12r	24l	24r	
textAvailable	.00	32	26	32	27	117
	1.00	12	18	15	21	66
Total		44	44	47	48	183

Here, we list all feedback received for the four conditions:

12r:

- It is very **complicated**. A bar graph would be much easier to comprehend.
- It is very **informative** and **clear**. Also it's **intuitive**.
- **Takes a few seconds to process**.
- There isn't enough information based on the chart to determine if there isn't a better way or not.
- It's **difficult to accurately determine** how many accidents occurred at any given hour.
- I found this **very hard to read at a glance** to get cohesive data.
- I think I would **prefer a table to be honest**, but this is okay.
- This is really **hard to look at** until you realize that it is showing like a **clock**. But the data is still **difficult to decipher**.
- I understand the graph now after looking at it and interacting with it but I **do not prefer this graph**
- a **rectangle graph might be better** for showing information and understanding it faster
- It's **pretty decent** but I JUST noticed that each bar represents 20 accidents. Before I thought 1 bar = 1 accident because **I didn't look closely enough**
- The **clock** picture is quite understandable. I would prefer a graph, but see the merit in displaying it this way. One big **disadvantage is the need to have two clocks**.
- I think a **bar graph would be more easily understood**
- **I can understand the graph presented**
- I **don't feel this visualization is suitable at all** to show traffic accidents. I found it **very hard to read initially** as I was trying to figure out if the 12pm on the left was for PM or AM, based on it stating AM in the center. This is very **confusing**.
- this is **more difficult to look at, takes more time to figure it out**



- I like this visualization because I can **easily understand it based on intuition**. It looks like a **clock**. More accidents mean more blue. The 0, 20, 40 bar is good too so I can see what the blue means.
- I **can't think of any better graphs** offhand

24r:

- Circular **clock** graphs are **hard to read and unnecessarily complex**.
- It is **challenging to read**.
- This graph was **harder to comprehend**.
- This one is **not as easy** as the bar graph
- It is **much more difficult to read** than a bar graph.
- **It takes a second**, but it is **effective** to evaluate daily traffic patterns and subsequent accidents
- It's a bit **confusing to look at initially**. Finding the data **isn't as intuitive as a standard bar graph**, for me anyway. **Once familiar with it, it's not so bad**, but still, I feel is **unnecessarily cluttered and confusing** to look at.
- This visual is **much harder and less intuitive to read** than the previous one with linear bars.
- This is **a lot more confusing** than the bar graph. **After looking at it for a few seconds I could understand it**, but it's **not as clear** (and I also think I understood this version because I was able to become familiar with the bar graph version first).
- It's **neat in that it parallels an analog clock face**, but it **takes a moment to understand** what's going on.
- It is **really good**, but I think I can understand it better than most people
- It's actually a **useful way to show the info**, but is **unusual** and **could be confusing at first**.
- It is **not easy to interpret**. I know it follows a **clock** format but it **doesn't translate well**.
- **A bit confusing** in my opinion
- It's **not bad** pro its shaped like a **clock** but **upon first glance its not easily understood**.
- it is **harder to read at a first glance**
- I think this gives a **clear and concise overview** of the distribution of traffic accidents throughout the day.
- This type of graph is **confusing when you first look at it**. It takes a minute or so before you can really grasp the data being shown.
- This visualization is **hard to read**.
- It's **pretty good** and **intuitive** to follow.
- It's **really difficult to decipher** this way, the standard graph is much better.

12l:

- This way looks much better and is way easier to take in and understand.
- I think this one is more efficient than the circular one.
- Now that I've seen both, this method is easier and more intuitive than the clock type visualization. It's easier to look at the bar charts and digest the information.
- not enough info to determine
- I prefer seeing a bar graph but it could be because I am more familiar with this.
- i think this is by far the best graph and easy to understand
- It's not quite a 5 because I'm sure there is some better way, but it's better than the circles.
- This makes more sense to me than the clock graph, but I can see the merits of both. I think reading particular numbers of this graph is easier, but the clock graph gives a better feel for time of day. I think I could grow to like the clock graph better as I saw it more. [12r]
- It is quickly understandable in bar format
- I feel that it's quite suitable, but one improvement that should be made is there should be more intervals or numbers on the left and more horizontal lines correlating with those for more accuracy.
- I think this was good, and a more traditional way. Also, for the last visualization, I believe I was asked to rank the visualization from 1 to 5, but I didn't. I'd give the last one (the circular visualization) a 4 due to better fitting the intuitive clock model [12r], whereas I'd give this one a 3 for just being a regular bar chart that doesn't stand out in the mind.
- OK, it's a little easier to read than the circles

24l:

- This chart seems much easier to understand at first glance.
- The bar graph seems much easier than the circle graph and indicates the trends per day for traffic accidents
- It's very good at letting you know the specific time of day accidents are occurring.
- It seems clear and simple, very easy to comprehend at a glance. I'm not sure how it could be improved while keeping the simplicity.
- Can't really say there's no better way, but this is a form that's more familiar -- no time really needed to understand the format of the data.
- This is simpler, less pretty or fancy but more effective than the previous one
- This is a great way to visualize it. It can be improved though.
- I like this one much better than the previous
- Not as original as the clock shape [24r] but easier to interpret at first.
- Honestly it's easier to see things quickly on the this graph, but I just think the other one is much more fun [24r].
- This method, while it works, is not optimal for traffic accident visualization. First of all, it's hard to estimate the exact numbers, and the row of bars makes it annoying, and it takes more time to glance over the entire day. I would not choose this method.
- Breaking it down by hour makes it easy to understand the data.

- This is **much easier to ready** than the first chart
- Most people are **familiar** with bar graphs so this is probably the **best way to convey the information to the most people**.

For each text response, we coded if the response contained positive and / or negative utterances. The overall sentiment was then computed as  $\text{sentiment} = (-1) \cdot \text{negative} + \text{positive}$ . This means that each text response received either a negative (-1), neutral (0), or positive (1) overall sentiment. The histogram shows the distribution of the three sentiment scores over the four conditions:

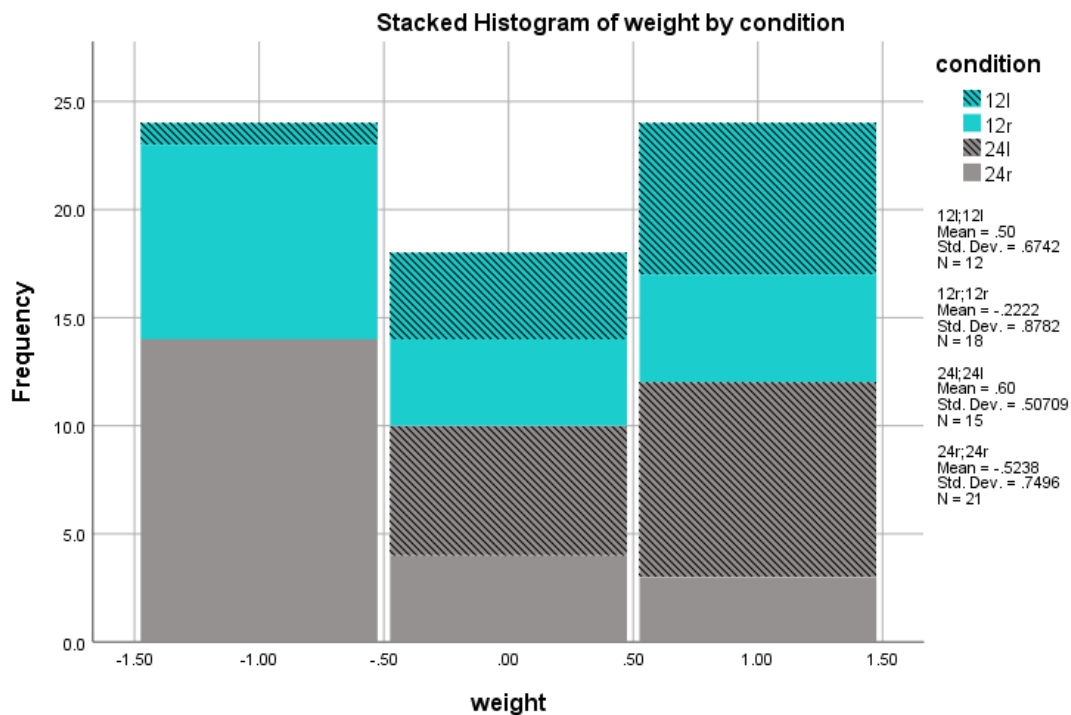


Figure 56: Histogram of overall sentiment scores (-1 = negative, 0 = neutral, 1 = positive) of all optional text responses for all four conditions.

weight \* condition Crosstabulation

Count

		condition				
		12l	12r	24l	24r	Total
weight	-1.00	1	9	0	14	24
	.00	4	4	6	4	18
	1.00	7	5	9	3	24
Total		12	18	15	21	66