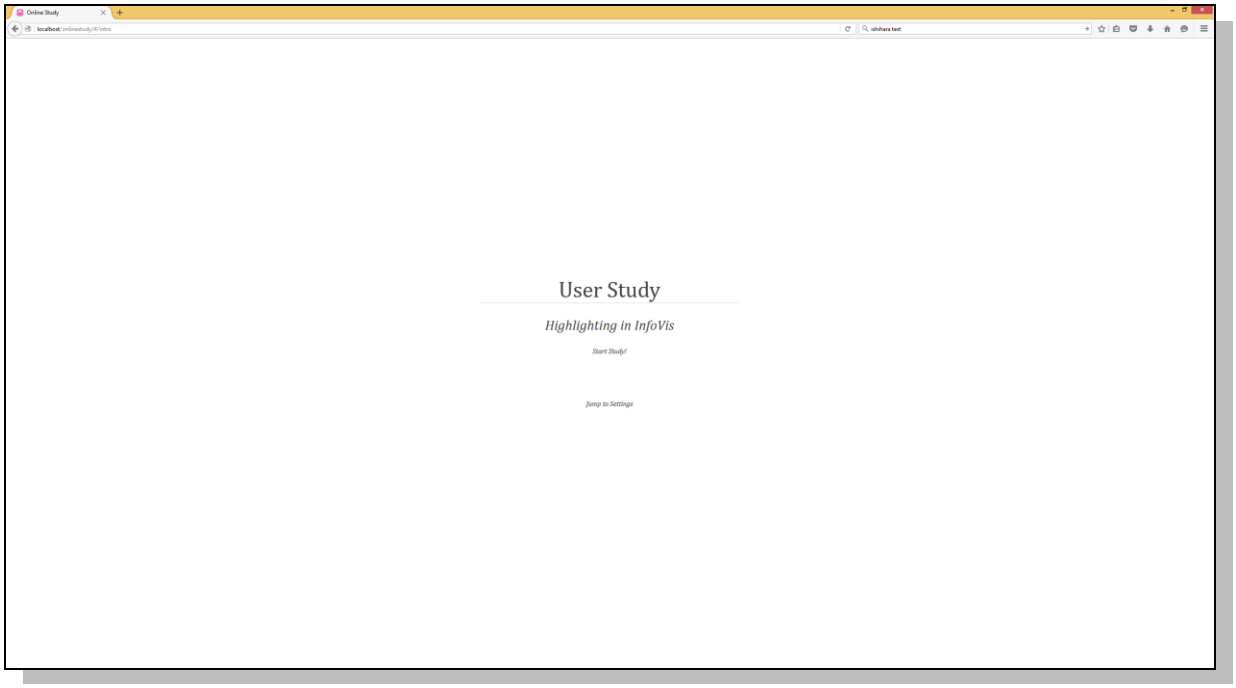
SUPPLEMENTAL MATERIAL FOR „EXPLORING VISUAL PROMINENCE OF MULTI-CHANNEL HIGHLIGHTING IN VISUALIZATIONS“

1 CONTENTS

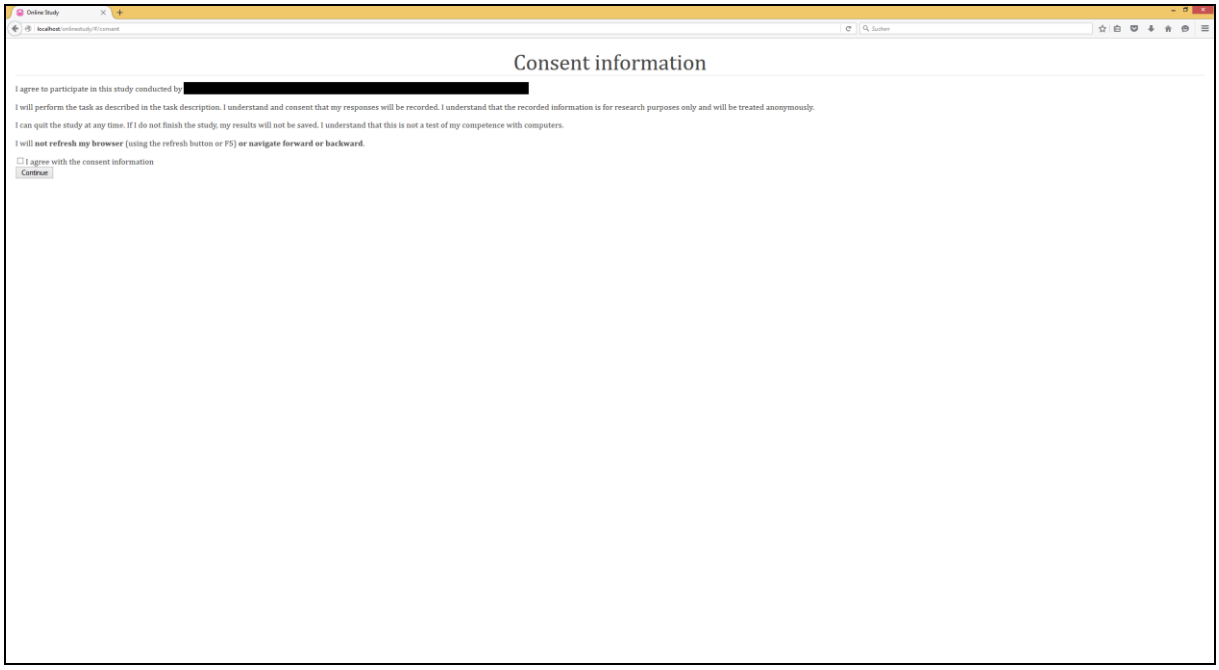
2 APPARATUS OF EXPERIMENTS

The experiments were implemented in a Firefox web browser, with the following procedure:

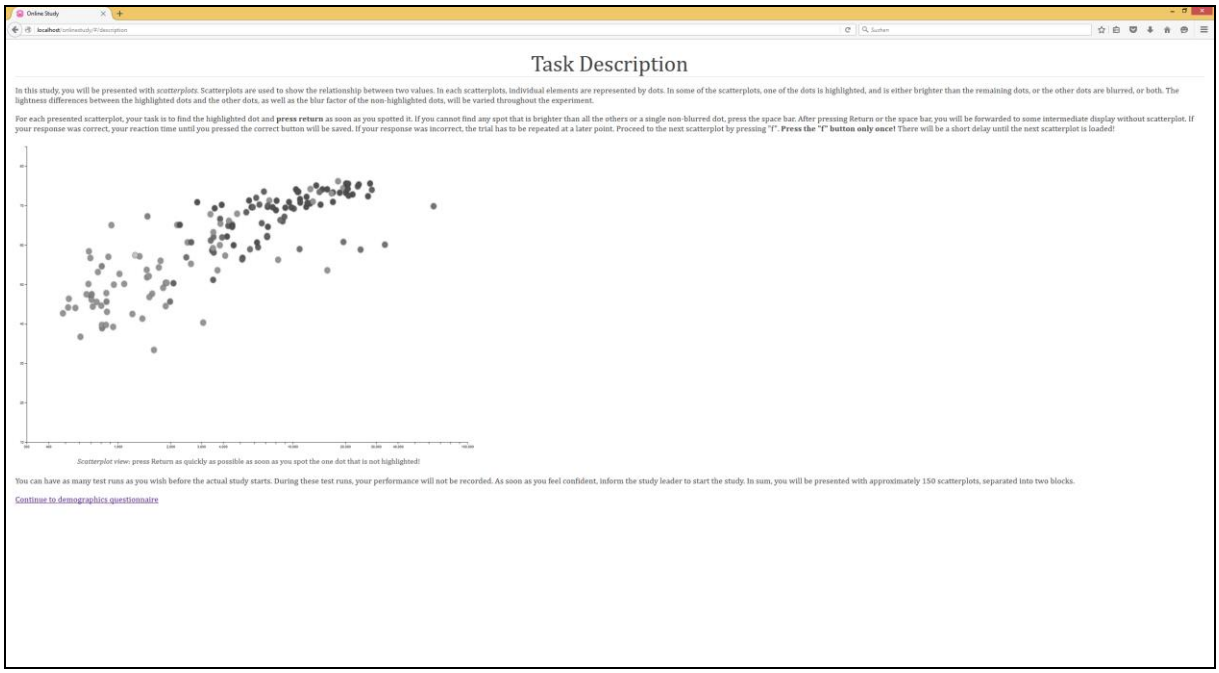
During welcome, the following screen was shown:



Users were asked to press the "Start Study!" button. After that, the consent form was presented:

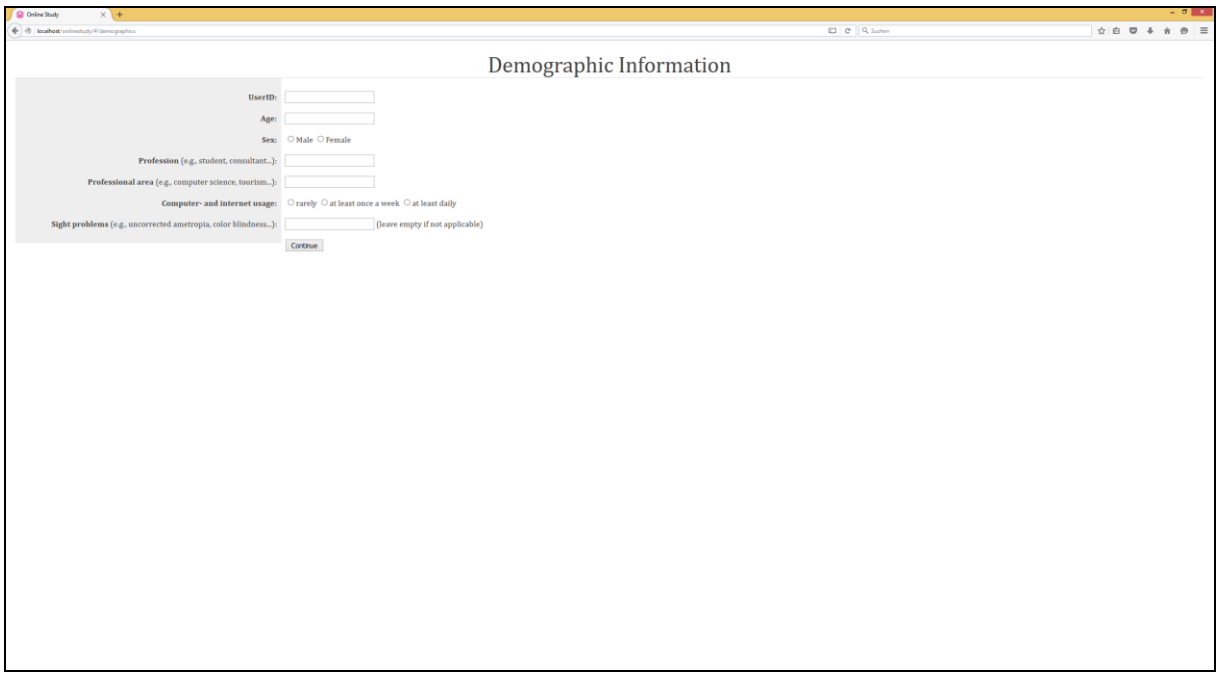


After checking the consent box, users continued to the task description. The following screenshot shows a task description for the single- vs. multi-channel experiment:



The task description texts and example images are presented further below.

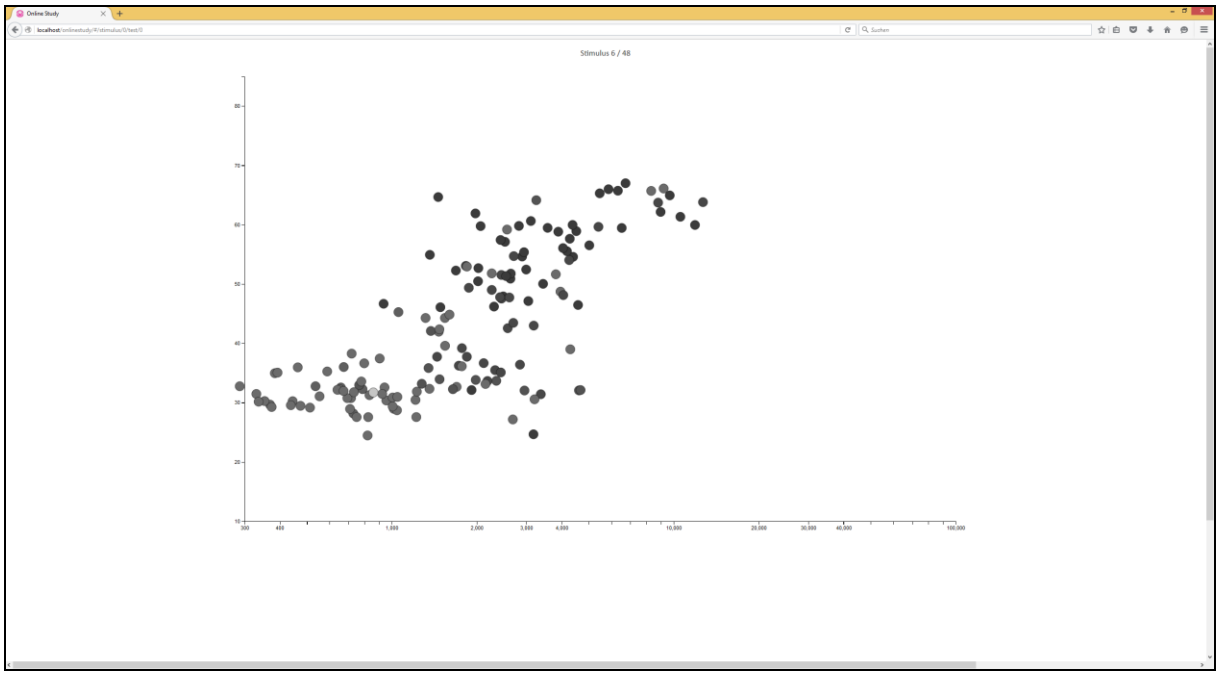
By clicking "continue", users were forwarded to a short demographic questionnaire:



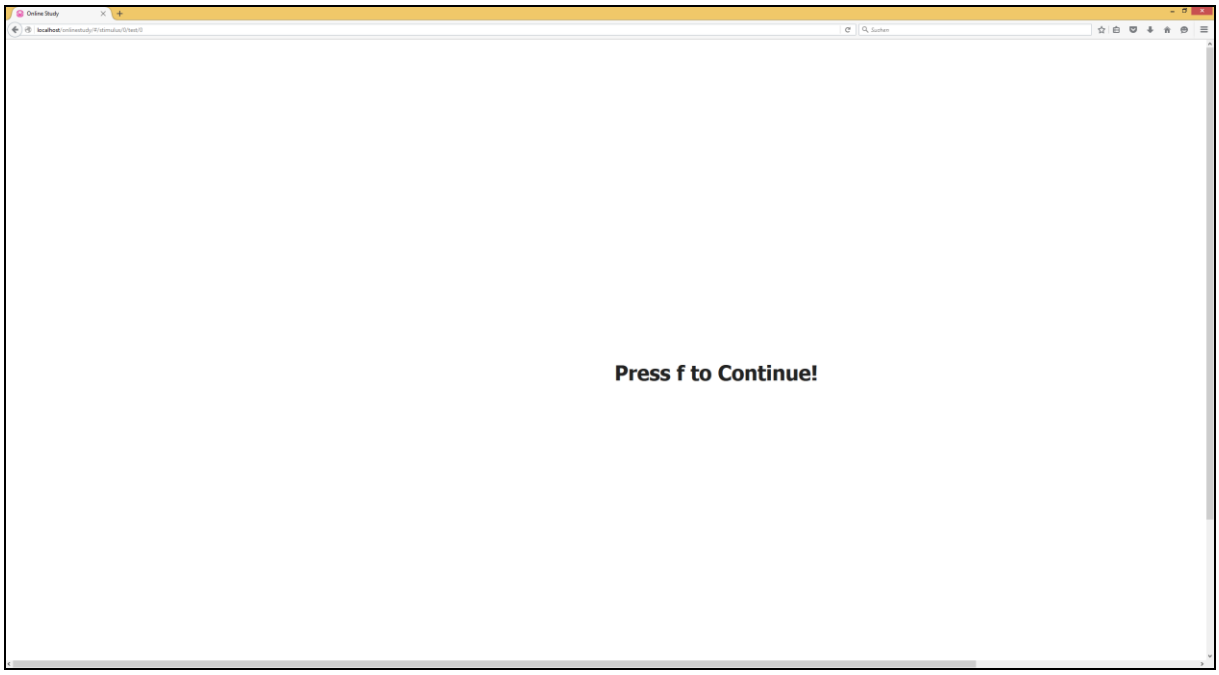
After that, a configuration interface was shown in which the experimenter chose the correct settings for the respective experiment.

Users could run the experiment without logging to fully understand the task and to get familiar with the controls. After telling the experimenter to run the actual study, the settings were reloaded and logging was initiated.

This is a screenshot of one stimulus in the single- vs. multi-channel highlighting study:



In the single-channel highlight study, the stimulus presentation was equivalent. After pressing return (found the target) or the space bar (user could not see any target), an intermediate screen was shown:



After pressing "f", we added a one-second delay so that users were forced to rest.

3 EXPERIMENT: VISUAL PROMINENCE IN A SINGLE HIGHLIGHT CHANNEL

3.1 TASK DESCRIPTION

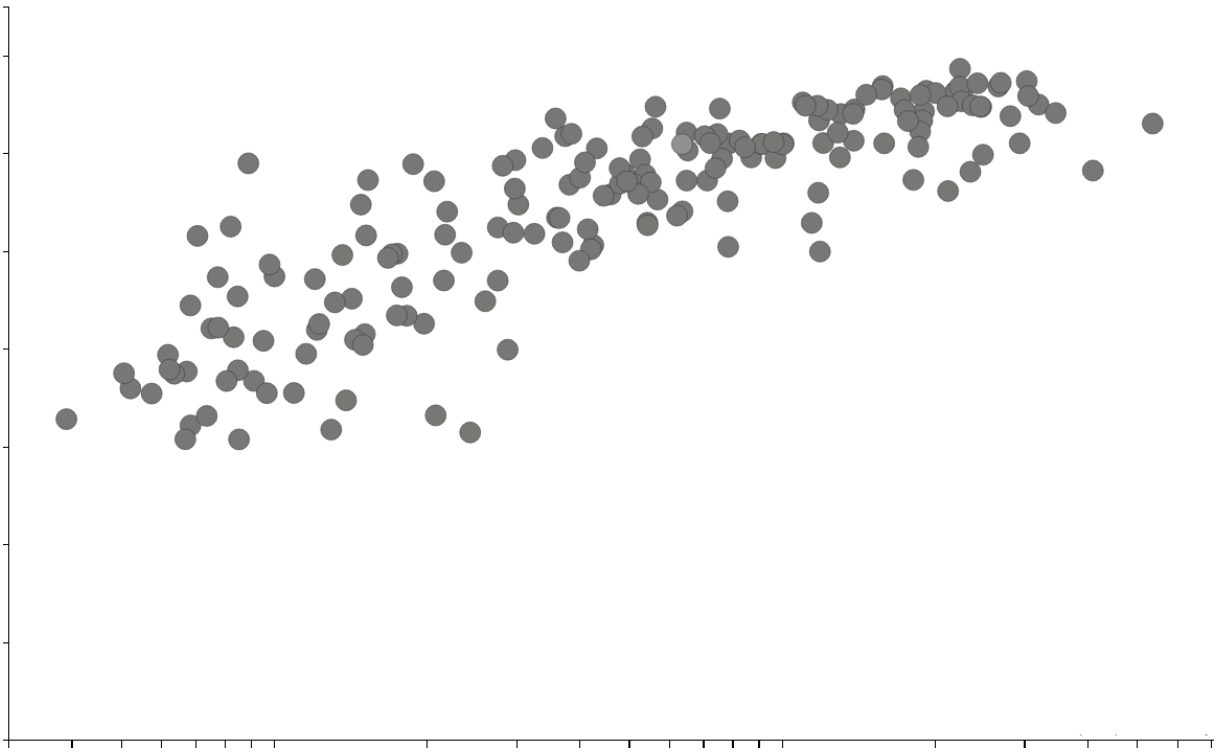
In this study, you will be presented with *scatterplots*. Scatterplots are used to show the relationship between two values. In each scatterplots, individual elements are represented by dots. In some of the scatterplots,

one of the dots is highlighted, and is brighter than the remaining dots. The lightness differences between the highlighted dots and the other dots will be varied throughout the experiment.

For each presented scatterplot, your task is to find the highlighted dot and **press return** as soon as you spotted it. If you cannot find any spot that is brighter than all the others, press the space bar. After pressing Return or the space bar, you will be forwarded to some intermediate display without scatterplot. If your response was correct, your reaction time until you pressed the correct button will be saved.If your response was incorrect, the trial has to be repeated at a later point. Proceed to the next scatterplot by pressing "f". **Press the "f" button only once!** There will be a short delay until the next scatterplot is loaded!

You can have as many test runs as you wish before the actual study starts. During these test runs, your performance will not be recorded. As soon as you feel confident, inform the study leader to start the study.

In sum, you will be presented with approximately 150 scatterplots, separated into two blocks.



3.2 RESULTS

Eight users had to perform 144 trials in total (3 T-N distances * 4 N ranges * 3 offsets * 2 target configurations * 2 repetitions), so we collected results for up to 1152 correct trials. 218 trials were answered incorrectly, so they had to be repeated until a correct response was gathered for the respective configuration.

Correctness

One participant was stopped after 180 trials, so only 1149 correct data points were collected:

correct * user Crosstabulation

Count

		user								
		2	3	4	5	6	7	8	9	Total
correct	FALSE	24	30	26	16	33	20	39	30	218
	TRUE	144	144	144	144	144	144	141	144	1149
Total		168	174	170	160	177	164	180	174	1367

There were only few false positive responses (target absent = A), but a lot of false negative responses (target present = P).

correct * target Crosstabulation

Count

		target		
		A	P	Total
correct	FALSE	20	198	218
	TRUE	576	573	1149
Total		596	771	1367

34% of targets with with T-N distance 10 were missed, which is close to chance level, but there were only 3% and 2% false negatives for T-N distance 20 and 30, respectively.

correct * T_D_distance Crosstabulation

Count

		T_D_distance			
		10	20	30	Total
correct	FALSE	198	12	8	218
	TRUE	382	383	384	1149
Total		580	395	392	1367

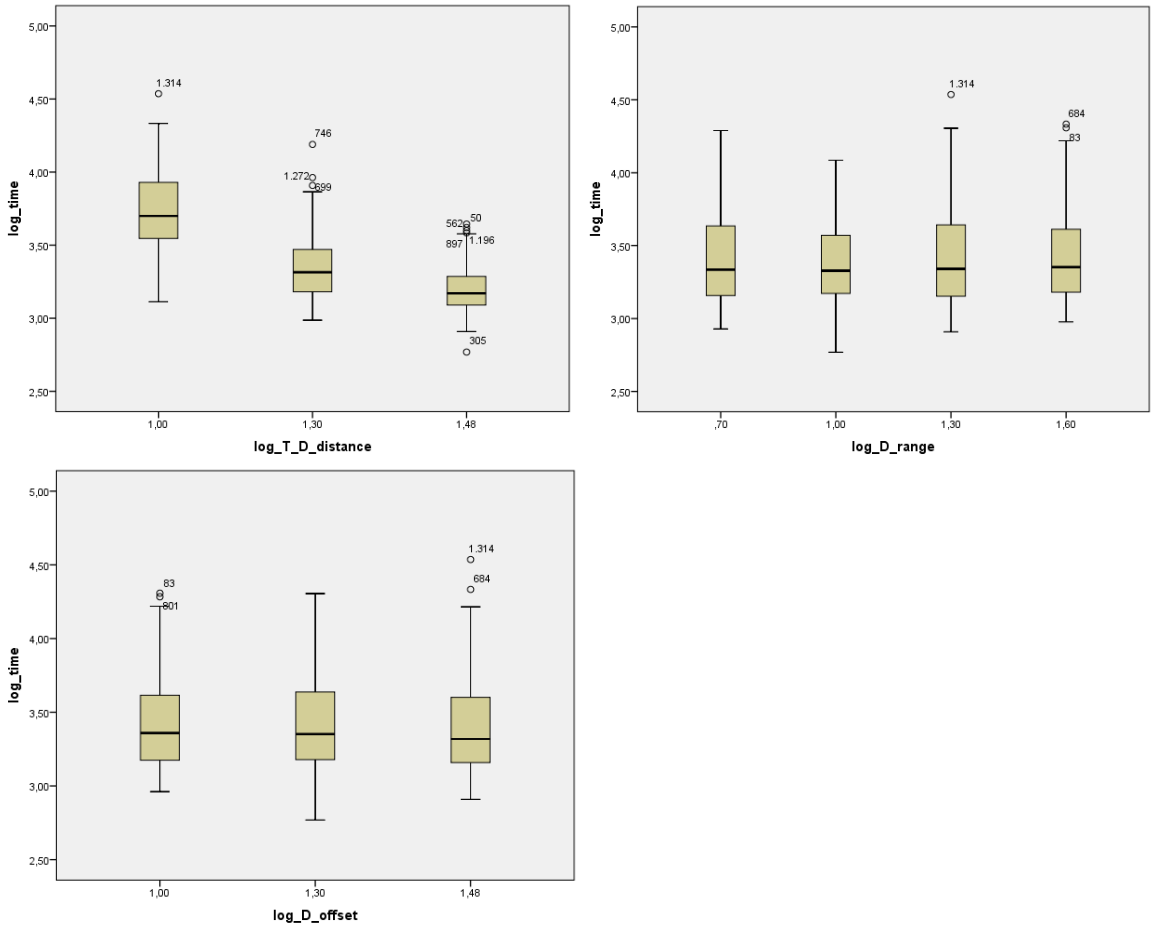
Mind that the higher total number for T_D_distance=10[1] stems from the fact that incorrect trials had to be repeated until a correct response was recorded.

Task Completion Time

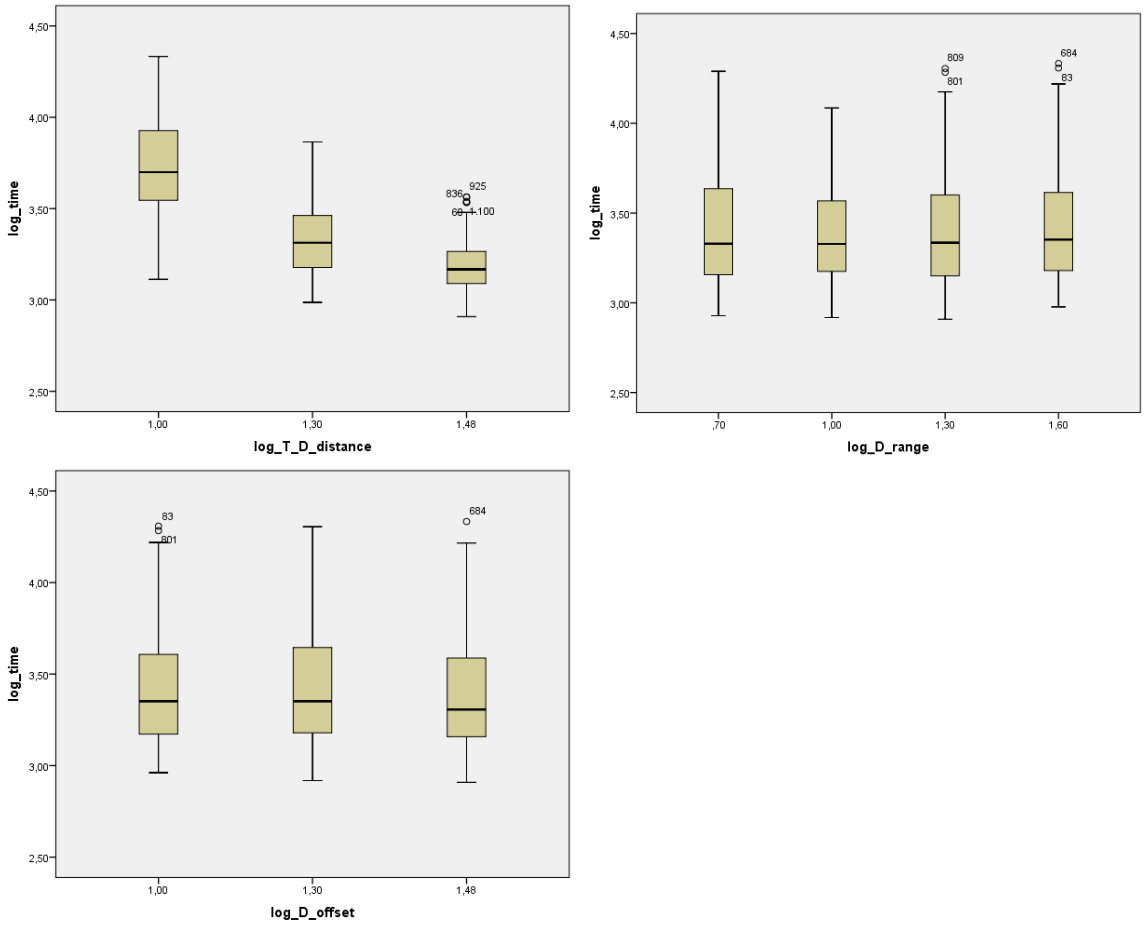
Since the obtained data is skewed, we applied a log-log-transformation on the data (i.e., we used the log-transformed values of the dependent variable and the independent variables). We only observed target present trials with correct responses.

Then, we removed all outliers. The following box plots show the log-transformed response times per factor levels with outliers.

[1] We used T_D_distance instead of T_N_distance in the SPSS analysis.



The following box plots show the log-transformed response times per factor levels with these outliers removed:



We then performed a linear regression of these data points with log-transformed response time as dependent variable, and log-transformed T-N-distance, N-range, and offset as independent variables. The linear regression yields a goodness-of-fit of $R^2 = 0.531$.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Statistikwerte ändern				
					Änderung in R-Quadrat	Änderung in F	df1	df2	Sig. Änderung in F
1	,728[a]	,531	,528	,21338	,531	210,340	3	558	,000

a. Einflußvariablen : (Konstante), log_D_offset, log_T_D_distance, log_D_range

b. Abhängige Variable: log_time

The regression is significant:

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	28,732	3	9,577	210,340	,000[b]
	Nicht standardisierte Residuen	25,407	558	,046		
	Gesamt	54,139	561			

a. Abhängige Variable: log_time

b. Einflußvariablen : (Konstante), log_D_offset, log_T_D_distance, log_D_range

T-N-distance is the only significant factor of the regression ($p < .001$).

Also, the regression coefficients of N-range and offset are very small:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressionskoeffizient B	Standardfehler	Beta		
1	(Konstante)	4,869	,087		55,789	,000
	log_T_D_distance	-1,146	,046	-,727	-25,078	,000
	log_D_range	,036	,027	,039	1,355	,176
	log_D_offset	-,043	,046	-,027	-,940	,348

a. Abhängige Variable: log_time

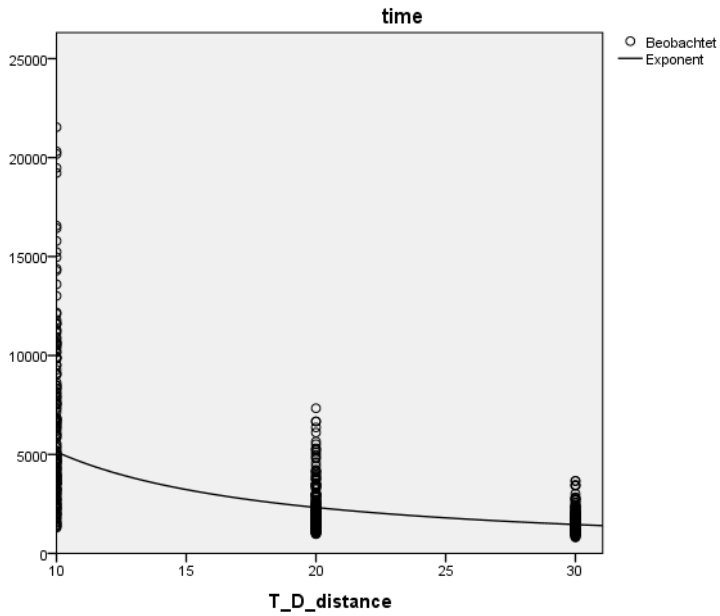
A linear regression on log-log-transformed data corresponds to a power regression on the original data. A power regression for response time and T_D_distance (i.e., the only factor explaining the model) yields a goodness-of-fit of R^2=.528. The constant is around 71,700, and the T-N coefficient (here named "b1) is -1.15.

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: time

Gleichung	Modellzusammenfassung					Parameterschätze	
	R-Quadrat	F	Freiheitsgrade 1	Freiheitsgrade 2	Sig.	Konstante	b1
Potenzfunktion	,528	627,473	1	560	,000	71657,387	-1,145

Die unabhängige Variable ist T_D_distance.


time
Beobachtet
Exponent
25000
20000
15000
10000
5000
0
10
15
20
25
30
T_D_distance

4 EXPERIMENT: VISUAL PROMINENCE IN MULTIPLE HIGHLIGHT CHANNELS

4.1 TASK DESCRIPTION VISUAL SEARCH PART

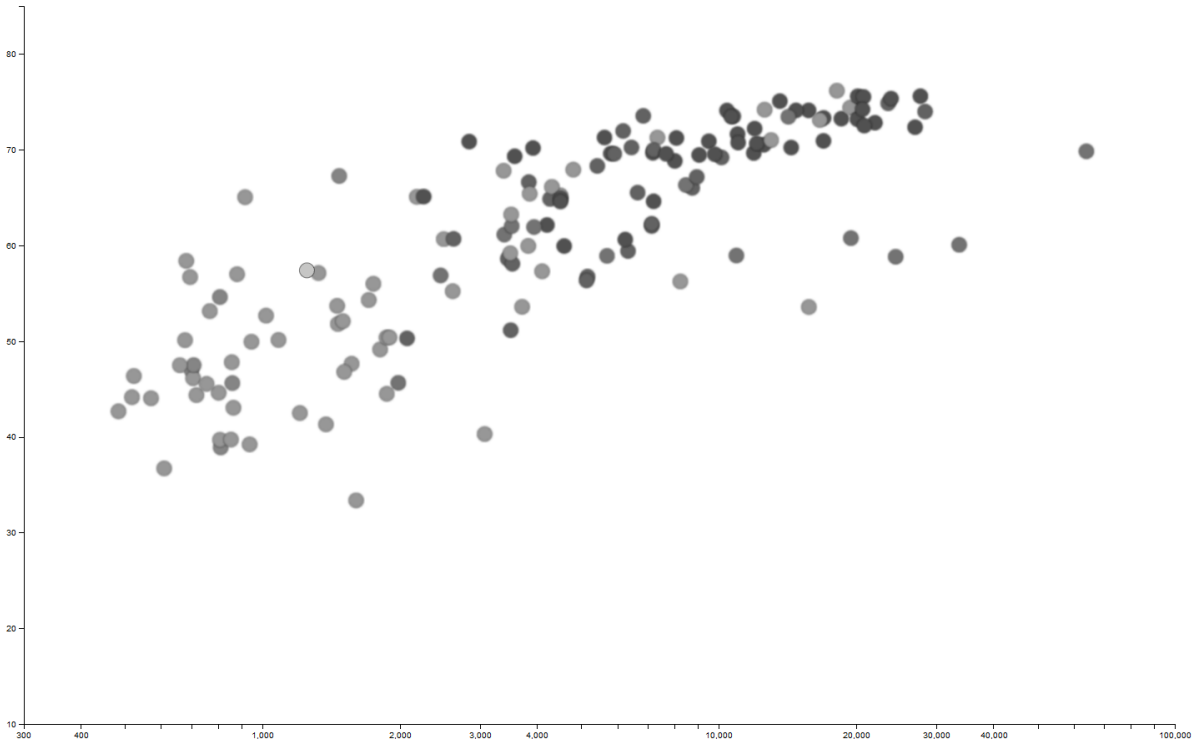
In this study, you will be presented with *scatterplots*. Scatterplots are used to show the relationship between two values. In each scatterplots, individual elements are represented by dots. In some of the scatterplots, one of the dots is highlighted, and is either brighter than the remaining dots, or the other dots are blurred, or both. The lightness differences between the highlighted dots and the other dots, as well as the blur factor of the non-highlighted dots, will be varied throughout the experiment.

For each presented scatterplot, your task is to find the highlighted dot and **press return** as soon as you spotted it. If you cannot find any spot that is brighter than all the others or a single non-blurred dot, press the space bar. After pressing Return or the space bar, you will be forwarded to some intermediate display without scatterplot. If your response was correct, your reaction time until you pressed the correct button will be saved. If your response was incorrect, the trial has to be repeated at a later point. Proceed to the next scatterplot by pressing "f".

Press the "f" button only once! There will be a short delay until the next scatterplot is loaded!

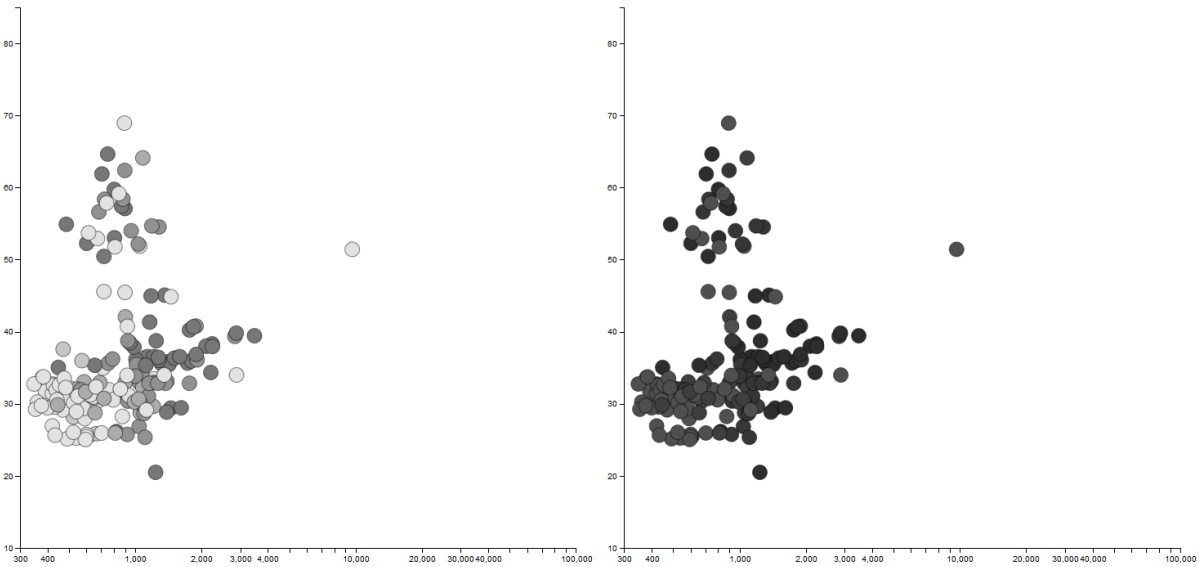
You can have as many test runs as you wish before the actual study starts. During these test runs, your performance will not be recorded. As soon as you feel confident, inform the study leader to start the study.

In sum, you will be presented with approximately 150 scatterplots, separated into two blocks.



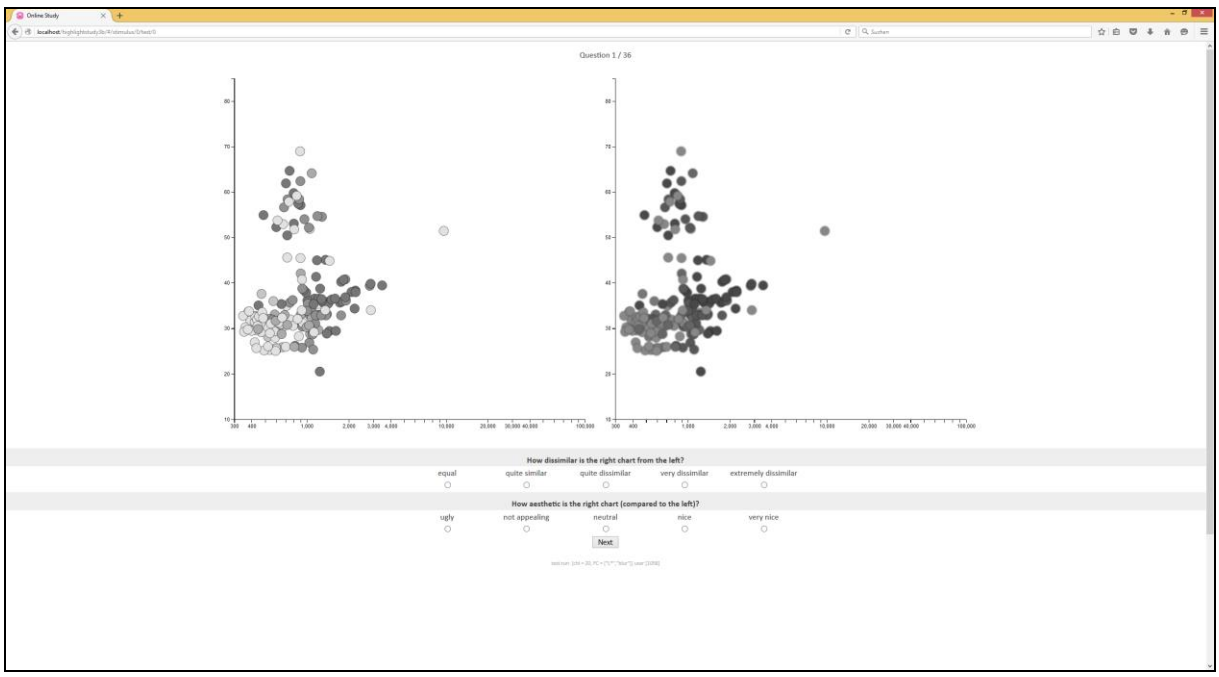
4.2 Task Description Subjective Dissimilarity Part

In this second part of the study, we will show you two identical scatterplots. Your task will be to rate how dissimilar the right on is from the left, and how aesthetic it looks, compared to the left (in a questionnaire). There will be no highlighted dot, but we will only show the dots that are not highlighted, and therefore darker and / or blurred. We will present 24 different configurations with different darkening and blur levels (the same as in the first part of the study). Only your questionnaire responses will be saved.



Stimulus Part 2

This is how the study was presented in the second part of the experiment:



4.3 RESULTS

Nine users had to perform 144 trials in total in the first experiment part (4 Chi steps * 3 channel configurations * 2 target configurations * 6 repetitions), so we collected results for 1296 correct trials. 160 trials were answered incorrectly, so they had to be repeated until a correct response was gathered for the respective configuration.

Correctness

There were only 10 false positive responses in total, but many false negatives (i.e., target misses):

target * correct Crosstabulation

Count

		correct		
		FALSE	TRUE	Total
target	A	10	648	658
	P	150	648	798
Total		160	1296	1456

Most false responses were gathered for the lowest Chi value (5):

Chi * correct Crosstabulation

Count

		correct		
		FALSE	TRUE	Total
Chi	5	148	324	472
	10	9	324	333
	15	1	324	325
	20	2	324	326
Total		160	1296	1456

There were more false responses in the multi-channel condition, and least in the sharpness (here called: "blur") condition:

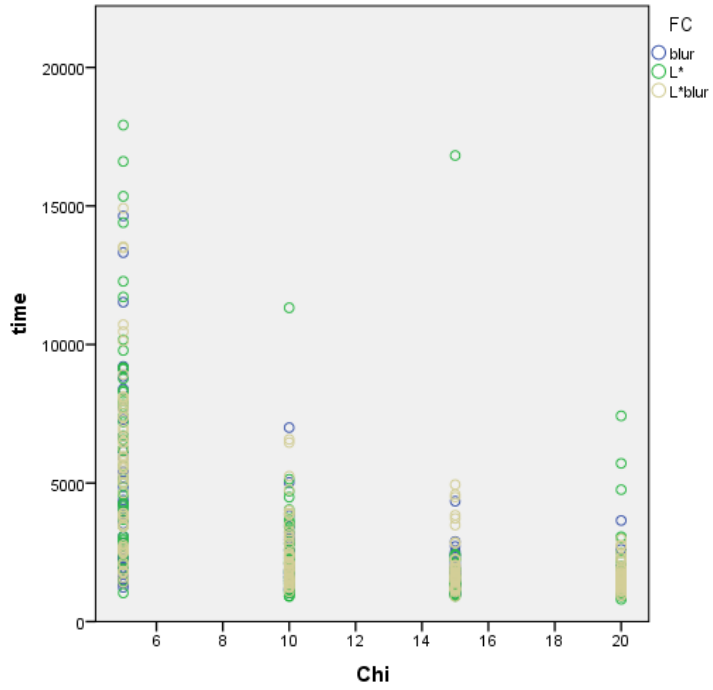
FC * correct Crosstabulation

Count

		correct		
		FALSE	TRUE	Total
FC	blur	21	432	453
	L*	43	432	475
	L*blur	96	432	528
Total		160	1296	1456

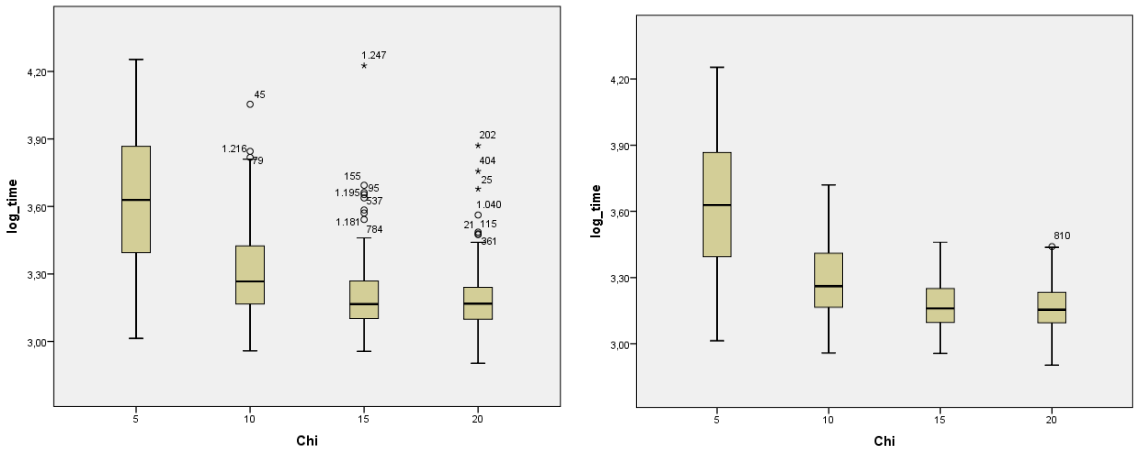
Response Time

Like in the first experiment, the response time per highlight strength (for correctly answered target present trials) is apparently skewed[2]:



Therefore, we log-transformed the response time values.

Then, we removed outlier cases for each highlight strength level (left: before outlier removal, right: after outlier removal):



[2] Chi here corresponds to Psi in the paper.

The Shapiro-Wilk test is significant for three highlight levels, so the normality assumption is violated:

Tests auf Normalverteilung

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Chi	Statistik	df	Signifikanz	Statistik	df	Signifikanz
log_time	5	,076	162	,023	,978	162	,010
	10	,085	158	,008	,973	158	,004
	15	,064	154	,200*	,982	154	,038
	20	,048	155	,200*	,987	155	,146

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

We therefore performed a Friedman test on the aggregated response times per user and highlight condition (i.e., the mean response times over all highlight strengths and repetitions).

Descriptive statistics (N, average, standard deviation, minimum, and maximum):

Deskriptive Statistiken

	N	Mittelwert	Standardabweichung	Minimum	Maximum
blur	9	3,2815	,11220	3,12	3,45
L*	9	3,3305	,11552	3,12	3,48
L*blur	9	3,3545	,11037	3,20	3,50

Ranks:

Ränge

	Mittlerer Rang
blur	1,44
L*	2,00
L*blur	2,56

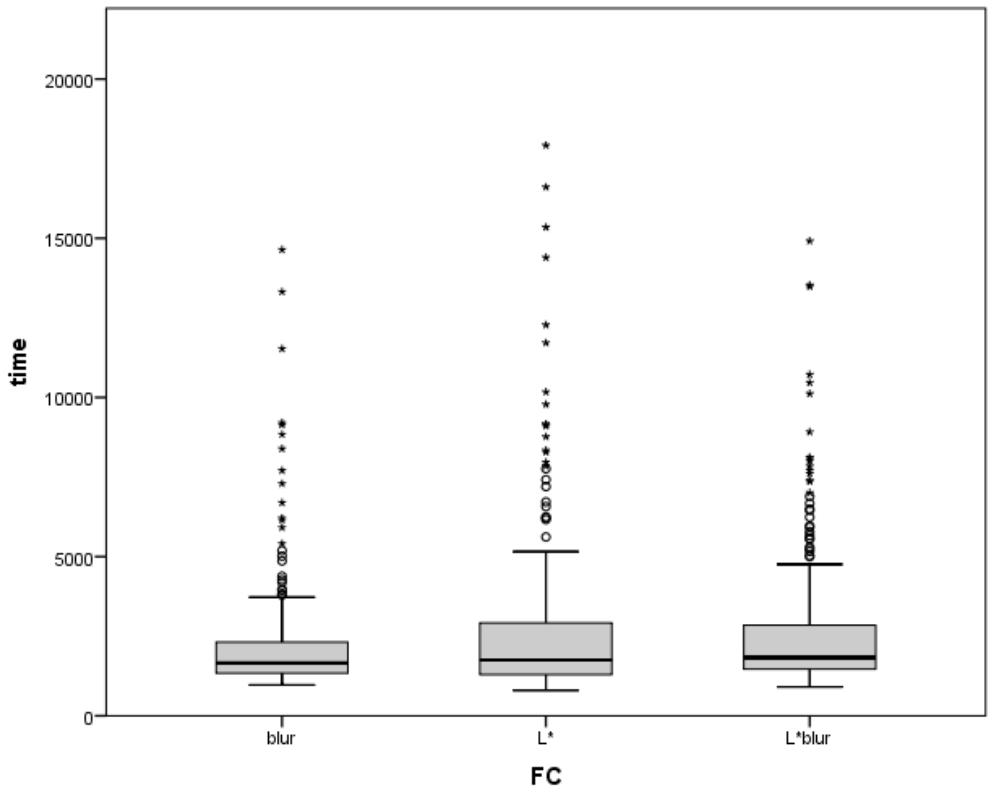
There is no significant difference between the response times.

Statistik für Test[a]

N	9
Chi-Quadrat	5,556
df	2
Asymptotische Signifikanz	,062

a. Friedman-Test

The box plots below show the response times per highlight condition:



Descriptive statistics show that, on average, blur (s) led to 560 ms (~ 24%) faster responses, compared to the the L*-condition (report columns: average, N, standard deviation).

Bericht

time

FC	Mittelwert	N	Standardabweichung
blur	2291,91	213	1978,929
L*	2852,26	209	2865,009
L*blur	2815,08	207	2403,805
Insgesamt	2650,27	629	2449,937

Perceived Dissimilarity

Users were asked to rate the dissimilarity between the two juxtaposed scatterplots, i.e., of the blurred and / or darkened context dots from their original appearance. The 5-point Likert scale had the following labels:

1	Equal
2	Quite similar
3	Quite dissimilar
4	Very dissimilar
5	Extremely dissimilar

We aggregated all dissimilarity responses per highlight condition and user, and performed a Friedman test.

Descriptive statistics (N, average, standard deviation, minimum, and maximum):

Deskriptive Statistiken

	N	Mittelwert	Standardabweichung	Minimum	Maximum
blur	9	3,1111	,73598	2,00	4,17
L*	9	3,4352	,46915	2,75	4,00
L*blur	9	3,0093	,25835	2,67	3,42

Ranks:

Ränge

	Mittlerer Rang
blur	2,00
L*	2,39
L*blur	1,61

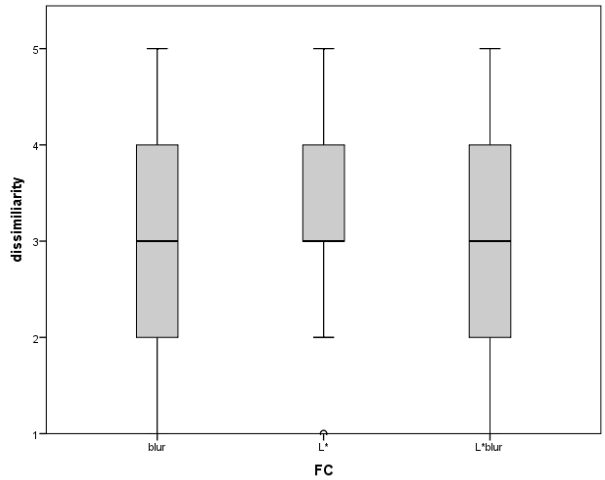
There is no significant difference in the dissimilarity responses.

Statistik für Test[a]

N	9
Chi-Quadrat	2,800
df	2
Asymptotische Signifikanz	,247

a. Friedman-Test

The box plot shows the dissimilarity scores per highlight condition:



For all three conditions, the median score was "quite dissimilar".

Aesthetics

Users were asked to rate the aesthetics of the the blurred and / or darkened context dots compared to the original scatterplot dots. The 5-point Likert scale had the following labels:

1	Ugly
2	Not appealing
3	Neutral
4	Nice
5	Very nice

We aggregated all aesthetics responses per highlight condition and user, and performed a Friedman test.

Descriptive statistics (N, average, standard deviation, minimum, and maximum):

Deskriptive Statistiken

	N	Mittelwert	Standardabweichung	Minimum	Maximum
blur	9	2,1204	,48967	1,42	3,25
L*	9	2,5000	,60953	1,67	3,75
L*blur	9	2,7037	,47891	2,17	3,67

Ranks:

Ränge

	Mittlerer Rang
blur	1,33
L*	1,83
L*blur	2,83

There is a significant difference in the aesthetics responses.

Statistik für Test[a]

N	9
Chi-Quadrat	10,800
df	2
Asymptotische Signifikanz	,005

a. Friedman-Test

We performed pairwise Wilcoxon-Signed Rank Test post-hoc comparisons with a Bonferroni-corrected critical p-value of 0.05 / 3 = 0.0167. There is a significant difference between blur and L*blur (s and L*s), with a higher aesthetics rating for L*blur than for blur alone.

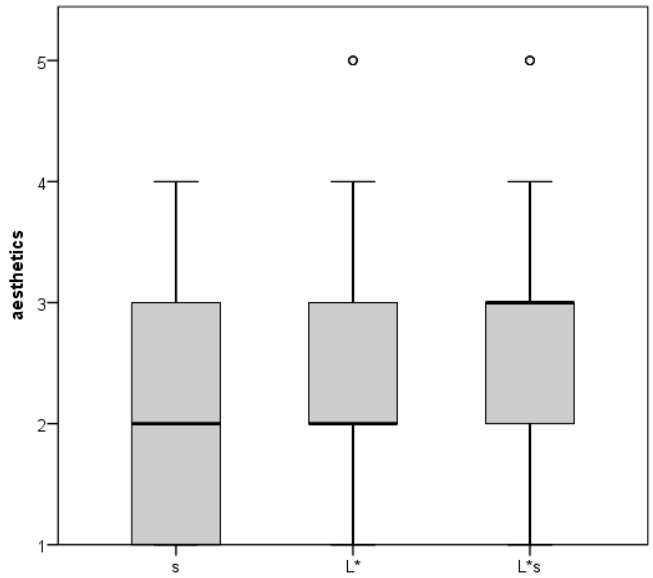
Statistik für Test[a]

	L* - blur	L*blur - L*	L*blur - blur
Z	-1,722[b]	-1,832[b]	-2,677[b]
Asymptotische Signifikanz (2-seitig)	,085	,067	,007

a. Wilcoxon-Test

b. Basiert auf negativen Rängen.

The box plot shows the subjective aesthetics ratings of the distorted context dots, compared to their original appearance:



While the multi-channel condition L*s received mostly neutral reponses (average: 2.83), the single-channel conditions L* and s received more votes for "not appealing".

Minkowski-r

In a log-log transformed plot, the power regression curve of response time becomes linear, where the slope represents beta and the intercept represents 1/k by Stephens' Power Law.

Below, the fitted linear regression lines of the three highlight conditions (s, L*, L*s; all log-log-transformed) are shown. The goodness-of-fit R^2 of the regressions are 0.371 (s), 0.537 (L*), and 0.542 (L*s), respectively.

Model summary (R squared, F, degree of freedeom e1, degree of freedom e2, significance) and parameter estimation (constant and b1); dependent variable: log_time; independent variable is log_Chi:

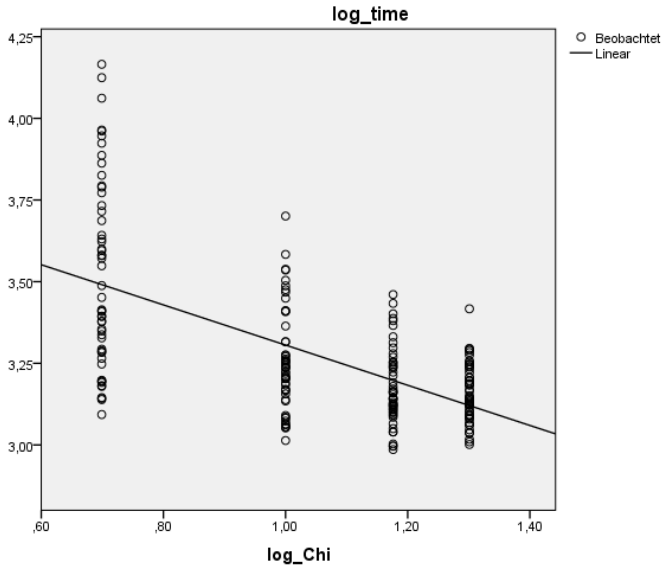
Modellzusammenfassung und Parameterschätzer

Abhängige Variable: log_time

Gleichung	Modellzusammenfassung					Parameterschätzer	
	R-Quadrat	F	Freiheitsgrad e 1	Freiheitsgrad e 2	Sig.	Konstante	b1
Linear	,371	124,479	1	211	,000	3,921	-,615

Die unabhängige Variable ist log_Chi.

It is not surprising that the multi-channel condition L*s caused lower performance, since we used the lowest possible channel combination factor (Minkowski-r = 1.0).

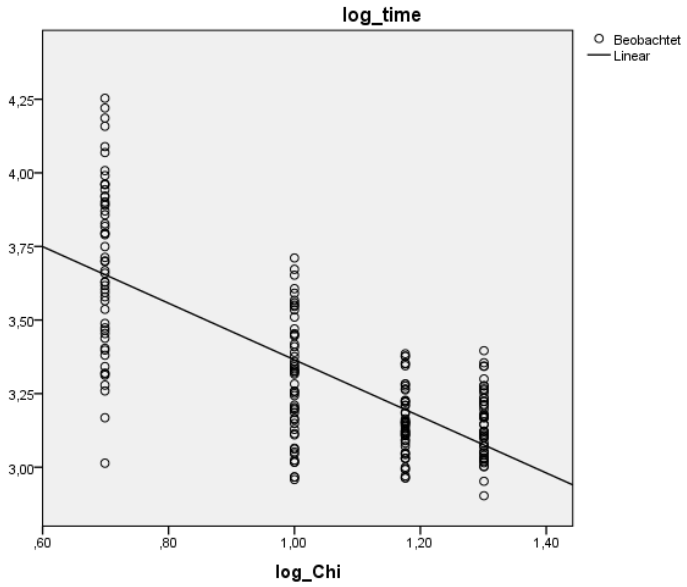


Modellzusammenfassung und Parameterschätzer

Abhängige Variable: log_time

Gleichung	Modellzusammenfassung					Parameterschätzer	
	R-Quadrat	F	Freiheitsgrad e 1	Freiheitsgrad e 2	Sig.	Konstante	b1
Linear	,537	239,904	1	207	,000	4,326	-,961

Die unabhängige Variable ist log_Chi.



Modellzusammenfassung und Parameterschätzer

Abhängige Variable: log_time

	Modellzusammenfassung					Parameterschätzer	
Gleichung	R-Quadrat	F	Freiheitsgrade 1	Freiheitsgrade 2	Sig.	Konstante	b1
Linear	,542	242,505	1	205	,000	4,235	-,851

Die unabhängige Variable ist log_Chi.

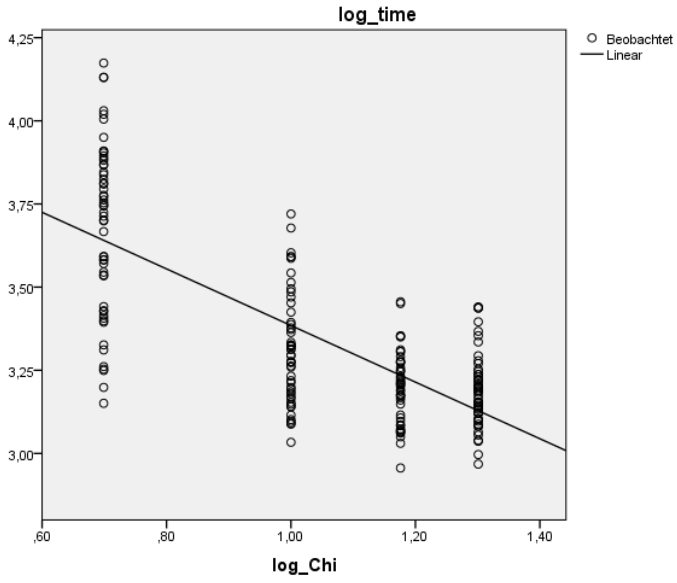


Figure 1: Regression of L*s.

To find the optimum Minkowski-r, we first combined all samples from the single-channel conditions (s and L*), and fitted a single regression line for these two conditions ($R^2 = 0.450$):

Modellzusammenfassung und Parameterschätzer

Abhängige Variable: log_time

	Modellzusammenfassung					Parameterschätzer	
Gleichung	R-Quadrat	F	Freiheitsgrad e 1	Freiheitsgrad e 2	Sig.	Konstante	b1
Linear	,450	343,173	1	420	,000	4,123	-,787

Die unabhängige Variable ist log_Chi.

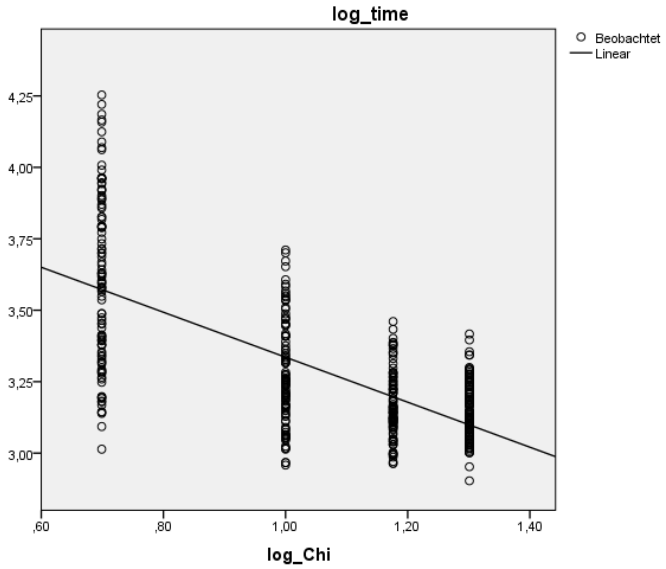
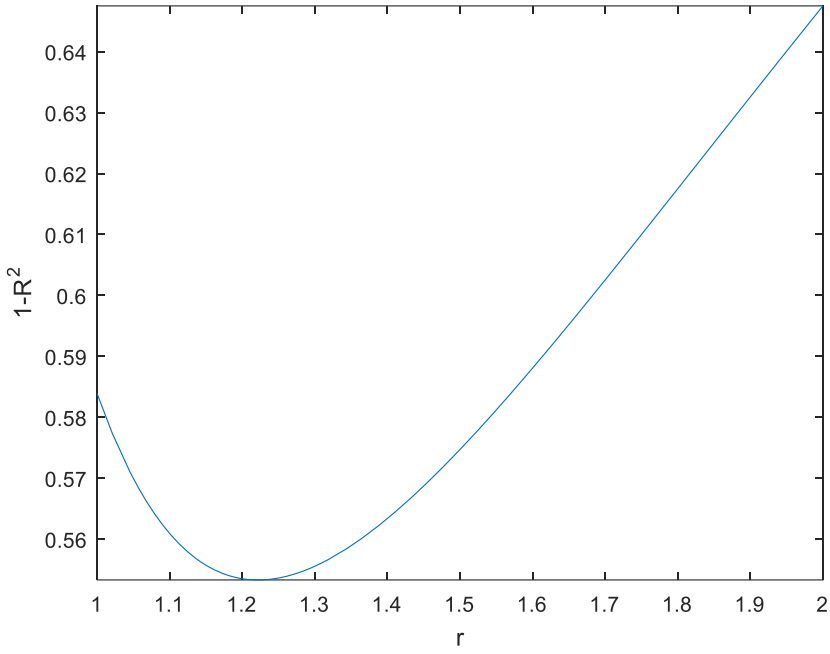


Figure 2: Combined regression of both single-channel conditions (L* and s)

With r=1.0, as used in the experiment, the two highlight channels are treated as fully separable dimensions. It can be expected, however, that sharpness and luminance slightly influence each other, so that r>1.0. By raising r, the combined sensation magnitude shrinks. Therefore, the intercept of the linear regression on the log-log-transformed values will decrease. Graphically speaking, the sample points and the fitted line shift towards the left in the scatterplot shown above.

To find the best match of the multi-channel condition regression (Figure 1) and the single-channel regression (Figure 2), we calculated the goodness-of-fit R^2 of the samples obtained from the two single-channel conditions (dots in Figure 2) with respect to the multi-channel model (line in Figure 1) as a function of r. By using $1-R^2$ as goodness-of-fit, we could analytically find a minimum of r within the interval [1, 2].

The minimum inverted R^2 was found for r=1.22, as illustrated below.


0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
1-R2
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
r