

Evaluating Reordering Strategies for Cluster Identification in Parallel Coordinates

– Statistical Analysis –

CONTENTS

1 Task 1	2
1.1 Datasets and Scripts	2
1.2 Data Cleaning and Number of Trials	2
1.3 Results: Time to Identify the Number of Clusters	3
1.3.1 Summary Statistics and Data Distribution	3
1.3.2 Statistical Analysis	3
1.4 Results: Similarity of Selected Axes-Pair (numerical interpretation)	5
1.4.1 Summary Statistics and Data Distribution	5
1.4.2 Statistical Analysis	5
2 Task 2	7
2.1 Datasets and Scripts	7
2.2 Data Cleaning and Number of Trials	7
2.3 Results: Completion Time to Identify and Mark Clusters (Varying Amount of Clutter)	8
2.3.1 Summary Statistics and Data Distribution	8
2.3.2 Statistical Analysis	8
2.4 Results: Completion Time to Identify and Mark Clusters (Varying Clutter Level and Ordering Strategy)	10
2.4.1 Summary Statistics and Data Distribution	10
2.4.2 Statistical Analysis	10
2.5 Results: Quality of Identified and Mark Clusters (Varying Amount of Clutter)	12
2.5.1 Summary Statistics and Data Distribution	12
2.5.2 Statistical Analysis	12
2.6 Results: Quality of Identified and Mark Clusters (Varying Clutter Level and Ordering Strategy)	14
2.6.1 Summary Statistics and Data Distribution	14
2.6.2 Statistical Analysis	14
2.7 Results: Confidence of Marked Clusters (Varying Amount of Clutter)	16
2.7.1 Data Distribution and Summary Statistics	16
2.7.2 Statistical Analysis	17
2.8 Results: Confidence of Marked Clusters (Varying Amount of Clutter and Ordering Strategy)	18
2.8.1 Data Distribution and Summary Statistics	18
2.8.2 Statistical Analysis	18
3 Task 3	19
3.1 Datasets and Scripts	19
3.2 Data Cleaning and Number of Trials	19
3.3 Results: Preference of Users towards an Ordering Strategy (Varying Clutter Levels)	20
3.3.1 Data Distribution and Summary Statistics	20
3.3.2 Statistical Analysis	20

1 TASK 1

In the following, we provide all supplementary material for the first task in our study.

1.1 Datasets and Scripts

Datasets

- `TASK1.csv`. Data of all trials including the similarity of the selected axes pairs.

R-scripts We use the following R-scripts for our analysis. Together with the datasets above, the results can be reproduced.

- `t1-efficiency-to-identify-the-number-of-clusters.r`: Analyzing the time/efficiency to identify the number of clusters.
- `t1-similarity-of-selected-axes-pair.r`: Analyzing the similarity of the selected axes pair.

1.2 Data Cleaning and Number of Trials

We considered **all participants** (and trials) for our analysis, independent of whether the number of identified clusters is correct or not. Hence we have **31 participants \times 3 trials each = 93 trials** for task 1.

1.3 Results: Time to Identify the Number of Clusters

We show the data distribution and statistic for the *time to identify the number of clusters*.

1.3.1 Summary Statistics and Data Distribution

Use `t1-efficiency-to-identify-the-number-of-clusters.r` to reproduce the results.

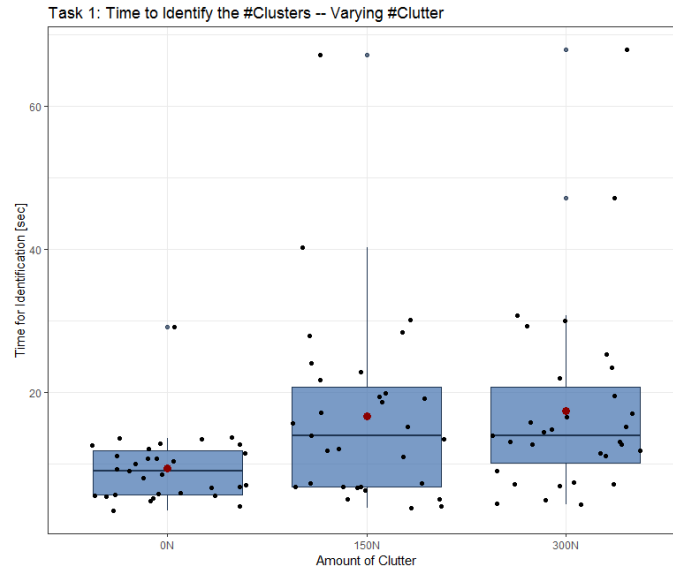


Figure 1: Distribution of time [sec] to identify the number of clusters. Mean is marked with a red dot.

	nbr.val	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N	31	3.52	29.11	25.59	292.74	9.08	9.44	0.87	1.77	23.34	4.83	0.51
150N	31	3.86	67.15	63.29	516.86	14.03	16.67	2.33	4.76	168.64	12.99	0.78
300N	31	4.35	67.88	63.53	541.71	13.96	17.47	2.37	4.84	174.37	13.21	0.76

Table 1: **Time to identify clusters.** Summary statistic for *different clutter levels*.

1.3.2 Statistical Analysis

We do have dependent trials/samples as every participant had to determine the number of clusters for all clutter levels (0N, 150N, and 300N). The summary statistic can be found in Table 1. In the following, we compute whether the time to identify the number of clusters differs significantly across the three clutter levels.

We use a repeated measures ANOVA to identify whether the mean values for the different clutter levels differ. Preconditions which need to be checked: (1) all groups need to be normally distributed, (2) sphericity (homogeneity of variances).

Check for normal distribution using Kolmogorov-Smirnov test and visual inspection. All $p - values > 0.05$, hence we can assume normal distribution for each of the three groups.

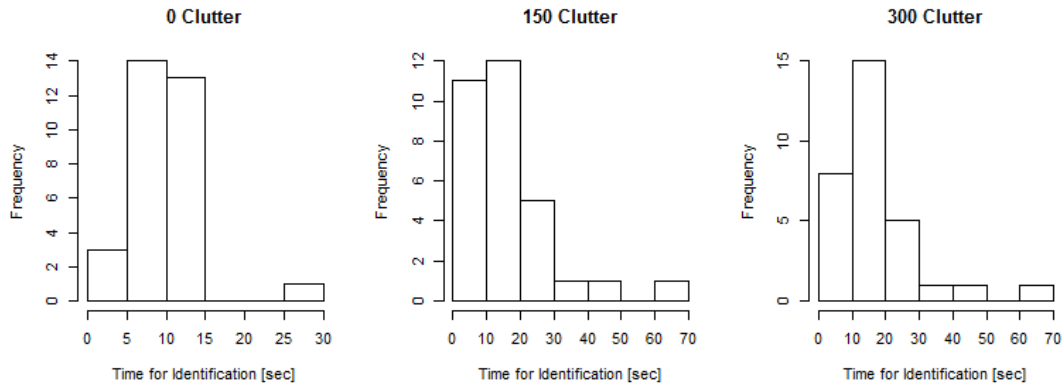


Figure 2: **Time to identify clusters.** Visual inspection to check for normal distribution.

	statistic	p.value	alternative	method	data.name
0N	0.16	0.38	two-sided	One-sample Kolmogorov-Smirnov test	clutter0\$time.identification
150N	0.16	0.39	two-sided	One-sample Kolmogorov-Smirnov test	clutter150\$time.identification
300N	0.22	0.08	two-sided	One-sample Kolmogorov-Smirnov test	clutter300\$time.identification

Table 2: **Time to identify clusters.** Results Kolmogorov-Smirnov test to check for normal distribution.

Check for statistical differences and sphericity using Repeated-measures ANOVA. We can see statistical significant differences (see Table 3).

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05	ges
1	(Intercept)	1.00	30.00	19634.82	4993.91	117.95	0.00	*	0.64
2	clutter	2.00	60.00	1213.26	5996.60	6.07	0.00	*	0.10

Table 3: **Repeated-measures ANOVA result.**

Mauchly's test for Sphericity is not significant ($p = 0.25 > 0.05$), hence sphericity can be assumed and we do not need to apply a correction on the degrees of freedom.

	Effect	W	p	p<.05
2	clutter	0.91	0.25	

Table 4: **Mauchly's test for Sphericity.** As $p > 0.05$, we can assume Sphericity.

Computing the effect size Eta-square (see field 'ges') in Table 3. Effect size $\eta^2 = 0.10$ which is between a medium and large effect.

Post hoc test using a t-test for dependent samples. Significant results between 0 and 150 and 0 and 300 clutter (but not between 150 and 300 clutter); see Table 5.

	method	data.name	p.value.0N	p.value.150N	p.adjust.method
150N	paired t tests	data\$time.identification and data\$clutter	0.01		bonferroni
300N	paired t tests	data\$time.identification and data\$clutter	0.01	1.00	bonferroni

Table 5: **Results Post hoc test.** Significant results between 0N and 150N as well as 0N and 300N.

1.4 Results: Similarity of Selected Axes-Pair (numerical interpretation)

We now analyze the *similarity* of the *selected neighboring axes-pairs*.

1.4.1 Summary Statistics and Data Distribution

Use `t1-similarity-of-selected-axes-pair.r` to reproduce the results.

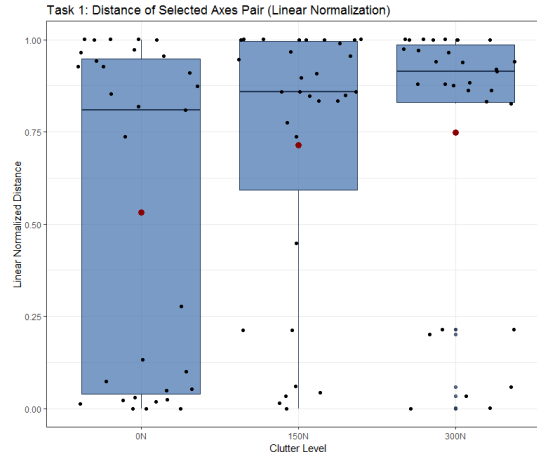


Figure 3: **Distribution of normalized distances** for different clutter levels (left). Mean values are marked with a red dot.

clutter	nbr.val	nbr.null	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N	31.00	3.00	0.00	1.00	1.00	16.48	0.81	0.53	0.08	0.16 0.20	0.20	0.44	0.84
150N	31.00	1.00	0.00	1.00	1.00	22.14	0.86	0.71	0.07	0.13 0.13	0.13	0.37	0.51
300N	31.00	1.00	0.00	1.00	1.00	23.19	0.91	0.75	0.06	0.13 0.13	0.13	0.36	0.48

Table 6: **Normalized distance of selected axes pair.** Summary statistic for different clutter levels.

1.4.2 Statistical Analysis

We now compare the mean values for the different clutter levels.

We use a Friedman test to check whether the mean values for the different clutter levels differ. Reason: A Repeated measures ANOVA has the precondition that the distributions of all groups need to follow a normal distribution. According to the Kolmogorov-Smirnov test below (+ visual inspection; see below), we cannot assume that the three clutter levels follow a normal distribution.

Check for normal distribution using Kolmogorov-Smirnov test and visual inspection.

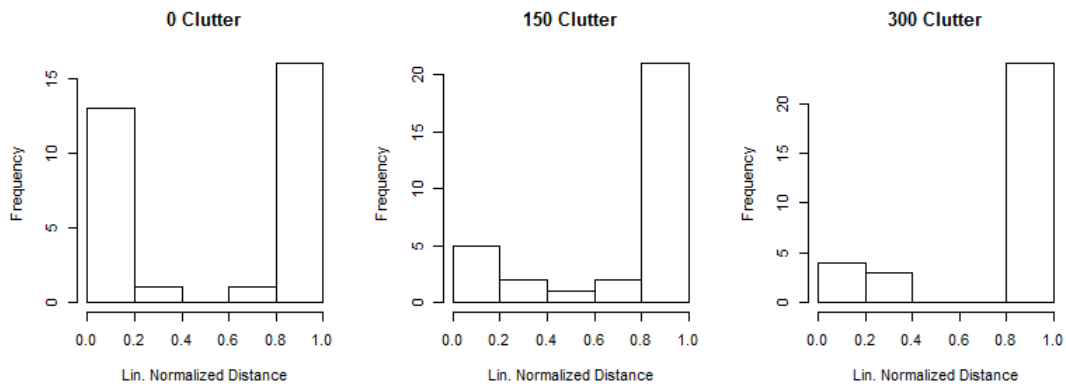


Figure 4: **Distribution of normalized distances.** Visual inspection to check normally distribution.

	statistic	p.value	alternative	method	data.name
0N	0.25	0.04	two-sided	One-sample Kolmogorov-Smirnov test	clutter0\$similarity.lin.norm
150N	0.31	0.01	two-sided	One-sample Kolmogorov-Smirnov test	clutter150\$similarity.lin.norm
300N	0.36	0.00	two-sided	One-sample Kolmogorov-Smirnov test	clutter300\$similarity.lin.norm

Table 7: **Distance of selected axes.** Results of Kolmogorov-Smirnov test to check for normally distribution. All $p < 0.05$, hence, none of the distributions can be assumed to be normally distributed.

Friedman test to check for statistical significance (no check for preconditions/assumptions required).

	statistic	parameter	p.value	method	data.name
Friedman chi-squared	4.77	2.00	0.09	Friedman rank sum test	dataFriedman

Table 8: **Distance of selected axes.** Results of Friedman test. Difference is not significant.

2 TASK 2

In the following, we provide all supplementary material for the second task in our study.

2.1 Datasets and Scripts

Datasets The following datasets are available to analyze the results of task 2:

- `TASK2.csv`. Time, cluster quality, and confidence of selected clusters. This file contains only the valid trials (see data cleaning below).

R-Scripts We use the following R-scripts for our analysis. Together with the dataset above, the results can be checked and reproduced.

- `t2-time-varying-clutter.r`
- `t2-time-varying-clutter-per-ordering.r`
- `t2-quality-varying-clutter.r`
- `t2-quality-varying-clutter-per-ordering.r`
- `t2-confidence-varying-clutter.r`
- `t2-confidence-varying-clutter-per-ordering.r`

2.2 Data Cleaning and Number of Trials

In task 2, participants had to select clusters in two neighboring dimensions. 132 trials were removed if participants selected the clusters in only one or more than two dimensions, or if the number of clusters was not consistent in the two neighboring dimensions. This cleaning led to 364 valid trials.

2.3 Results: Completion Time to Identify and Mark Clusters (Varying Amount of Clutter)

We show the distribution and statistic for the *time to identify and mark* the number of clusters.

2.3.1 Summary Statistics and Data Distribution

Use `t2-time-varying-clutter.r` to reproduce the results.

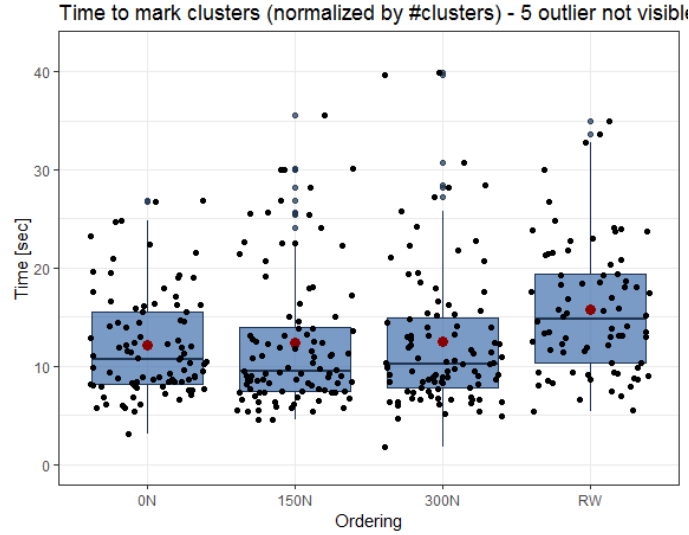


Figure 5: Distribution of time [sec] to mark the different clusters (normalized by the number of clusters). 5 outliers not visible. Mean is marked with a red dot.

	nbr.val	nbr.na	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N	92.00	0.00	3.08	57.72	54.64	1196.42	10.79	13.00	0.83	1.66	63.92	8.00	0.61
150N	96.00	0.00	4.51	35.55	31.04	1187.43	9.53	12.37	0.71	1.41	48.59	6.97	0.56
300N	96.00	0.00	1.75	53.88	52.12	1242.71	10.26	12.94	0.85	1.68	68.93	8.30	0.64
RW	80.00	0.00	5.41	49.92	44.52	1323.10	15.02	16.54	0.93	1.85	69.00	8.31	0.50

Table 9: **Time to identify and mark clusters.** Summary statistics for *different clutter levels*.

2.3.2 Statistical Analysis

We do have dependent trials/samples as every participant had to select clusters for the clutter levels 0N, 150N, and 300N as well as the real world dataset. However, due to our data cleaning process, the number of trials per clutter level differs. Hence, we cannot compute a repeated measure ANOVA. Instead, we would like to compute a regular ANOVA, assuming that we have independent samples. But, as our data is not normally distributed (see below), we use a Kruskal-Wallis test instead.

Check for normal distribution using Kolmogorov-Smirnov test and visual inspection. some p – values < 0.05 , hence we cannot assume normal distribution for each of the three four.

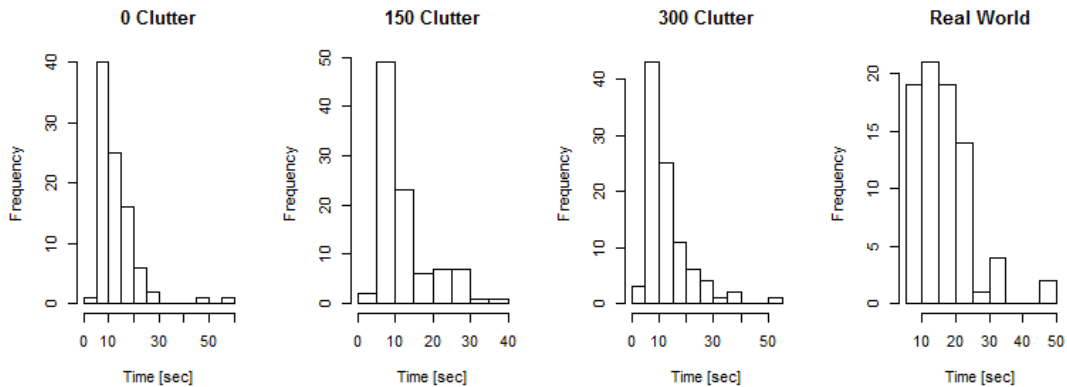


Figure 6: **Time to mark clusters.** Visual inspection to check normally distribution.

	statistic	p.value	alternative	method	data.name
D	0.16	0.02	two-sided	One-sample Kolmogorov-Smirnov test	data0\$time.marking.normalized
D	0.17	0.01	two-sided	One-sample Kolmogorov-Smirnov test	data150\$time.marking.normalized
D	0.17	0.01	two-sided	One-sample Kolmogorov-Smirnov test	data300\$time.marking.normalized
D	0.11	0.31	two-sided	One-sample Kolmogorov-Smirnov test	dataRW\$time.marking.normalized

Table 10: **Time to mark clusters.** Results of Kolmogorov-Smirnov test to check for normally distribution. Most $p - values < 0.05$, hence we cannot assume the groups follow a normal distribution.

Kruskal-Wallis test. Result: significant effect of clutter to time.

Kruskal-Wallis rank sum test

data: time.marking.normalized by clutter

Kruskal-Wallis chi-squared = 23.308, df = 3, **p-value = 3.484e-05**

Posthoc test: Pairwise comparisons using Wilcoxon rank sum test

data: data[, "time.marking.normalized"] and data[, "clutter"]

```

      ON      150N      300N
150N 1.00000 -          -
300N 1.00000 1.00000 -
RW   0.00140 0.00012 0.00058

```

P value adjustment method: bonferroni

Only clutter vs. RW has a statistical effect.

2.4 Results: Completion Time to Identify and Mark Clusters (Varying Clutter Level and Ordering Strategy)

We show the distribution and statistic for the *time to identify and mark* the number of clusters for the different axes orderings within each clutter level.

2.4.1 Summary Statistics and Data Distribution

Use `t2-time-varying-clutter-per-ordering.r` to reproduce the results.

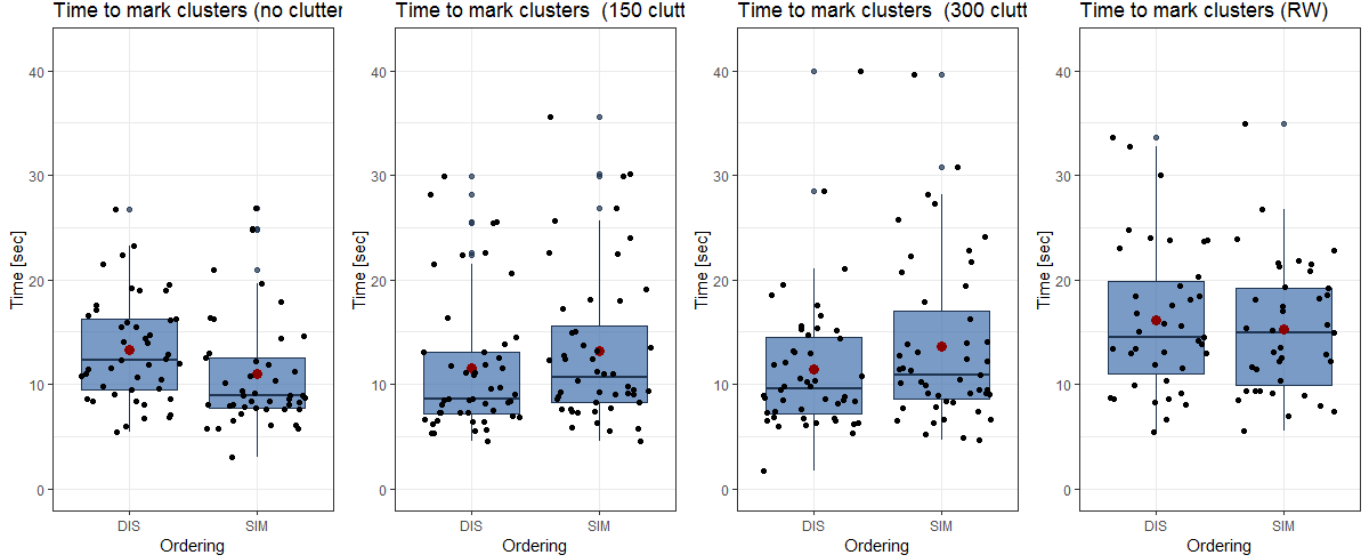


Figure 7: Distribution of time [sec] to mark the different clusters (normalized by the number of clusters, vary ordering strategy). 2 outliers not visible.

	nbr.val	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N - DIS	46.00	5.43	57.72	52.29	655.66	12.35	14.25	1.21	2.44	67.49	8.21	0.58
0N - SIM	46.00	3.08	47.77	44.69	540.76	8.90	11.76	1.13	2.27	58.59	7.65	0.65
150N - DIS	48.00	4.51	29.95	25.44	552.07	8.58	11.50	0.95	1.90	43.01	6.56	0.57
150N - SIM	48.00	4.51	35.55	31.04	635.36	10.67	13.24	1.06	2.13	53.66	7.33	0.55
300N - DIS	48.00	1.75	39.93	38.17	548.73	9.62	11.43	0.94	1.88	42.06	6.49	0.57
300N - SIM	48.00	4.65	53.88	49.23	693.99	11.12	14.46	1.39	2.79	92.59	9.62	0.67
RW - DIS	40.00	5.41	49.92	44.52	680.21	14.79	17.01	1.39	2.81	77.16	8.78	0.52
RW - SIM	40.00	5.52	45.98	40.46	642.89	15.07	16.07	1.25	2.52	62.17	7.88	0.49

Table 11: **Time to identify and mark clusters.** Summary statistic for the two orderings / clutter level. Mean is marked with a red dot.

2.4.2 Statistical Analysis

Only the combination of RW follows a normal distribution and we will test the similarity using a one-sided t-test for dependent samples. For the other combinations, we will use a Wilcoxon Signed-rank test, which does not assume normal distribution (also for 150 DIS: $p = 0.04903$, do be comparable with the others).

Check for normal distribution using Kolmogorov-Smirnov test.

	statistic	p.value	alternative	method	data.name
0N - SIM	0.21	0.03	two-sided	One-sample Kolmogorov-Smirnov test	data0SIM\$time.marking.normalized
0N - DIS	0.15	0.22	two-sided	One-sample Kolmogorov-Smirnov test	data0DIS\$time.marking.normalized
150N - SIM	0.18	0.08	two-sided	One-sample Kolmogorov-Smirnov test	data150SIM\$time.marking.normalized
150N - DIS	0.19	0.05	two-sided	One-sample Kolmogorov-Smirnov test	data150DIS\$time.marking.normalized
300N - SIM	0.23	0.02	two-sided	One-sample Kolmogorov-Smirnov test	data300SIM\$time.marking.normalized
300N - DIS	0.17	0.14	two-sided	One-sample Kolmogorov-Smirnov test	data300DIS\$time.marking.normalized
RW - SIM	0.11	0.74	two-sided	One-sample Kolmogorov-Smirnov test	dataRWSIM\$time.marking.normalized
RW - DIS	0.13	0.44	two-sided	One-sample Kolmogorov-Smirnov test	dataRWDIS\$time.marking.normalized

Table 12: Check if time distributions follow a normal distribution for $\{0N, 150N, 300N, RW\} \times \{SIM, DIS\}$

0N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = -2.4637, **p-value = 0.01298**

alternative hypothesis: true mu is not equal to 0

150N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = 1.959, p-value = **0.05011**

alternative hypothesis: true mu is not equal to 0

300N SIM vs. DIS

Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = 1.959, p-value = **0.05011**

alternative hypothesis: true mu is not equal to 0

RW SIM vs. DIS - we are allowed to compute a t-test, but for comparability, we also compute a Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = -0.24194, **p-value = 0.816**

alternative hypothesis: true mu is not equal to 0

One sided Paired t-test

data: dataRWSIM[, "time.marking.normalized"] and dataRWDIS[, "time.marking.normalized"]

t = -0.65658, df = 39, **p-value = 0.2577**

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 1.461225

sample estimates:

mean of the differences

-0.9330208

2.5 Results: Quality of Identified and Mark Clusters (Varying Amount of Clutter)

We show the data distribution and statistic for the *quality to identify and mark clusters*.

2.5.1 Summary Statistics and Data Distribution

Use `t2-quality-varying-clutter.r` to reproduce the results.

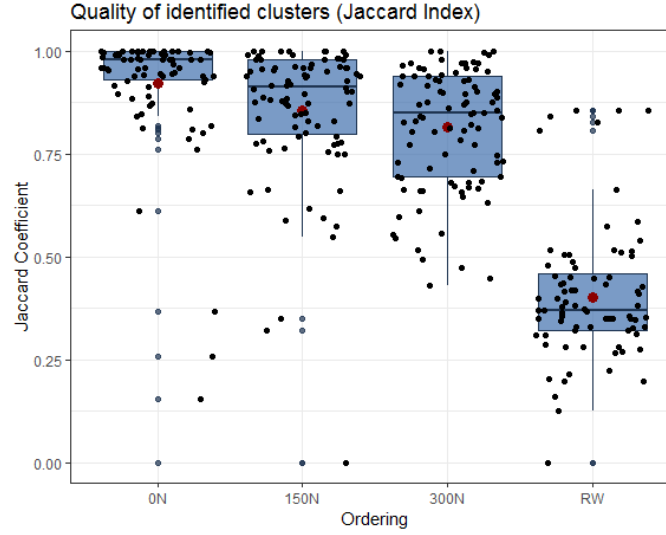


Figure 8: Distribution of the quality marking the clusters. Mean is marked with a red dot.

	nbr.val	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N	92.00	0.00	1.00	1.00	84.65	0.98	0.92	0.02	0.04	0.03	0.17	0.19
150N	96.00	0.00	1.00	1.00	82.08	0.91	0.86	0.02	0.04	0.03	0.19	0.22
300N	96.00	0.43	1.00	0.57	78.34	0.85	0.82	0.02	0.03	0.02	0.15	0.18
RW	80.00	0.00	0.85	0.85	32.15	0.37	0.40	0.02	0.04	0.03	0.17	0.42

Table 13: **Quality to identify clusters.** Summary statistics for *different clutter levels*.

2.5.2 Statistical Analysis

We do have dependent trials/samples as every participant had to select clusters for the clutter levels 0N, 150N, and 300N as well as the real world dataset. However, due to our data cleaning process, the number of trials per clutter level differs. Hence, we cannot compute a repeated measure ANOVA. Instead, we would like to compute a regular ANOVA, assuming that we have independent samples. But, as our data is not normally distributed (see below), we use a Kruskal-Wallis test.

Check for normal distribution using Kolmogorov-Smirnov test and visual inspection. Two p – values < 0.05 , hence we cannot assume all groups follow a normal distribution.

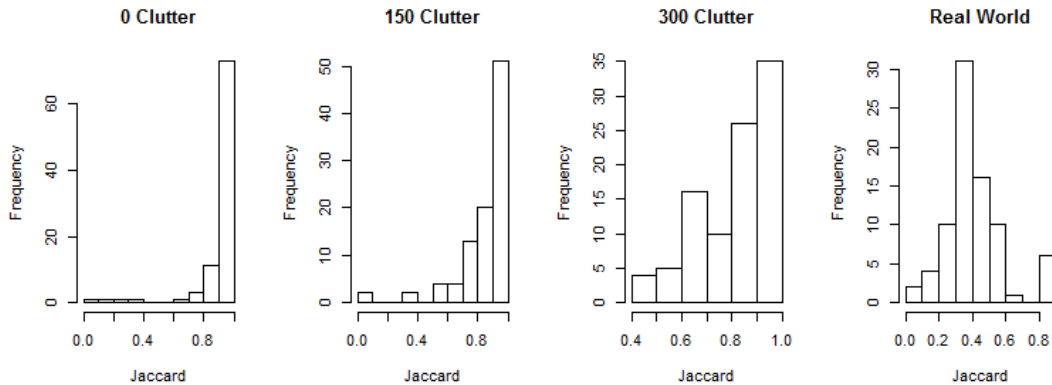


Figure 9: **Quality of marked clusters.** Visual inspection to check normally distribution.

	statistic	p.value	alternative	method	data.name
D	0.32	0.00	two-sided	One-sample Kolmogorov-Smirnov test	data0\$jaccard.coefficient
D	0.22	0.00	two-sided	One-sample Kolmogorov-Smirnov test	data150\$jaccard.coefficient
D	0.11	0.18	two-sided	One-sample Kolmogorov-Smirnov test	data300\$jaccard.coefficient
D	0.13	0.14	two-sided	One-sample Kolmogorov-Smirnov test	dataRW\$jaccard.coefficient

Table 14: **Quality of marked clusters.** Results of Kolmogorov-Smirnov test to check for normally distribution. Two $p - values < 0.05$, hence we cannot assume the groups follow a normal distribution.

Kruskal-Wallis test. Results:

Kruskal-Wallis rank sum test

data: jaccard.coefficient by clutter

Kruskal-Wallis chi-squared = 181.56, df = 3, $p - value < 2.2e - 16$

Posthoc test Pairwise comparisons using Wilcoxon rank sum test

data: data[, "jaccard.coefficient"] and data[, "clutter"]

```

      0N      150N      300N
150N 6.0e-05 -        -
300N 1.7e-10 0.037   -
RW   < 2e-16 < 2e-16 < 2e-16

```

P value adjustment method: bonferroni

All combinations are statistical significant.

2.6 Results: Quality of Identified and Mark Clusters (Varying Clutter Level and Ordering Strategy)

We show the data distribution and statistic for the *quality to identify and mark clusters* for the different reordering strategies.

2.6.1 Summary Statistics and Data Distribution

Use `t2-quality-varying-clutter-per-ordering.r` to reproduce the results.

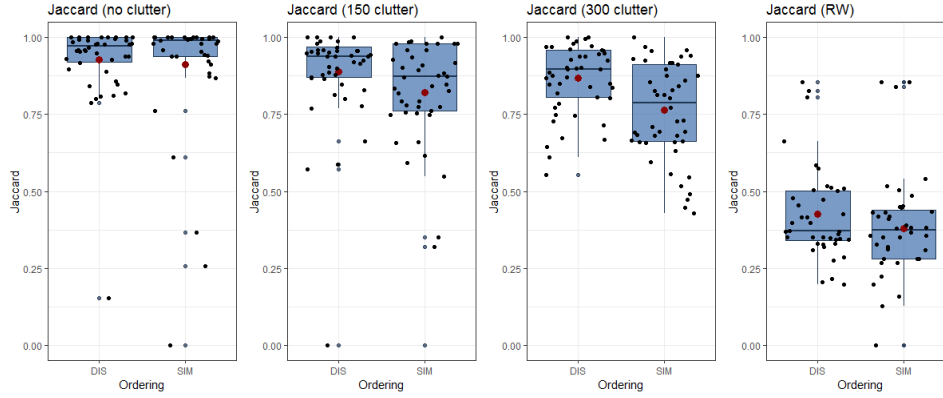


Figure 10: Distribution of the quality score to identify and mark the clusters. Mean is marked with a red dot.

	nbr.val	min	max	range	sum	median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
0N - DIS	46.00	0.15	1.00	0.85	42.71	0.97	0.93	0.02	0.04	0.02	0.13	0.14
0N - SIM	46.00	0.00	1.00	1.00	41.94	0.99	0.91	0.03	0.06	0.04	0.20	0.22
150N - DIS	48.00	0.00	1.00	1.00	42.70	0.94	0.89	0.02	0.05	0.03	0.17	0.19
150N - SIM	48.00	0.00	1.00	1.00	39.38	0.87	0.82	0.03	0.06	0.04	0.20	0.24
300N - DIS	48.00	0.55	1.00	0.45	41.64	0.90	0.87	0.02	0.03	0.01	0.11	0.13
300N - SIM	48.00	0.43	1.00	0.57	36.70	0.79	0.76	0.02	0.05	0.03	0.16	0.21
RW - DIS	40.00	0.20	0.85	0.66	17.00	0.37	0.42	0.02	0.05	0.02	0.15	0.36
RW - SIM	40.00	0.00	0.85	0.85	15.15	0.37	0.38	0.03	0.06	0.03	0.18	0.48

Table 15: **Quality to identify clusters.** Summary statistics for *different clutter levels and ordering strategies*.

2.6.2 Statistical Analysis

Check for normal distribution using Kolmogorov-Smirnov test.

	statistic	p.value	alternative	method	data.name
0N - SIM	0.33	0.00	two-sided	One-sample Kolmogorov-Smirnov test	data0SIM\$jaccard.coefficient
0N - DIS	0.30	0.00	two-sided	One-sample Kolmogorov-Smirnov test	data0DIS\$jaccard.coefficient
150N - SIM	0.19	0.05	two-sided	One-sample Kolmogorov-Smirnov test	data150SIM\$jaccard.coefficient
150N - DIS	0.25	0.00	two-sided	One-sample Kolmogorov-Smirnov test	data150DIS\$jaccard.coefficient
300N - SIM	0.10	0.76	two-sided	One-sample Kolmogorov-Smirnov test	data300SIM\$jaccard.coefficient
300N - DIS	0.12	0.47	two-sided	One-sample Kolmogorov-Smirnov test	data300DIS\$jaccard.coefficient
RW - SIM	0.17	0.20	two-sided	One-sample Kolmogorov-Smirnov test	dataRWSIM\$jaccard.coefficient
RW - DIS	0.16	0.26	two-sided	One-sample Kolmogorov-Smirnov test	dataRWDIS\$jaccard.coefficient

Table 16: Check if quality distributions follow a normal distribution for $\{0N, 150N, 300N, RW\} \times \{SIM, DIS\}$

0N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = 0.98081, **p-value = 0.3327**

alternative hypothesis: true mu is not equal to 0

150N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = -2.6156, **p-value = 0.008126**

alternative hypothesis: true mu is not equal to 0

300N SIM vs. DIS – we are allowed to compute a t-test, but for comparability, we also compute a Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = -3.3641, **p-value = 0.0005425**

alternative hypothesis: true mu is not equal to 0

One-sided, Paired t-test

data: data300SIM[, "jaccard.coefficient"] and data300DIS[, "jaccard.coefficient"]

t = -3.5747, df = 47, **p-value = 0.000412**

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.05454405

sample estimates:

mean of the differences

-0.1027938

RW SIM vs. DIS – we are allowed to compute a t-test, but for comparability, we also compute a Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

Z = -1.156, **p-value = 0.2535**

alternative hypothesis: true mu is not equal to 0

One-sided, Paired t-test

data: dataRWSIM[, "jaccard.coefficient"] and dataRWDIS[, "jaccard.coefficient"]

t = -1.234, df = 39, **p-value = 0.1123**

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 0.01691751

sample estimates:

mean of the differences

-0.0463

2.7 Results: Confidence of Marked Clusters (Varying Amount of Clutter)

We show the distribution and statistic for the confidence of the marked clusters in task 2.

2.7.1 Data Distribution and Summary Statistics

Use `t2-confidence-varying-clutter.r` to reproduce the results. The charts have been created with d3.js.

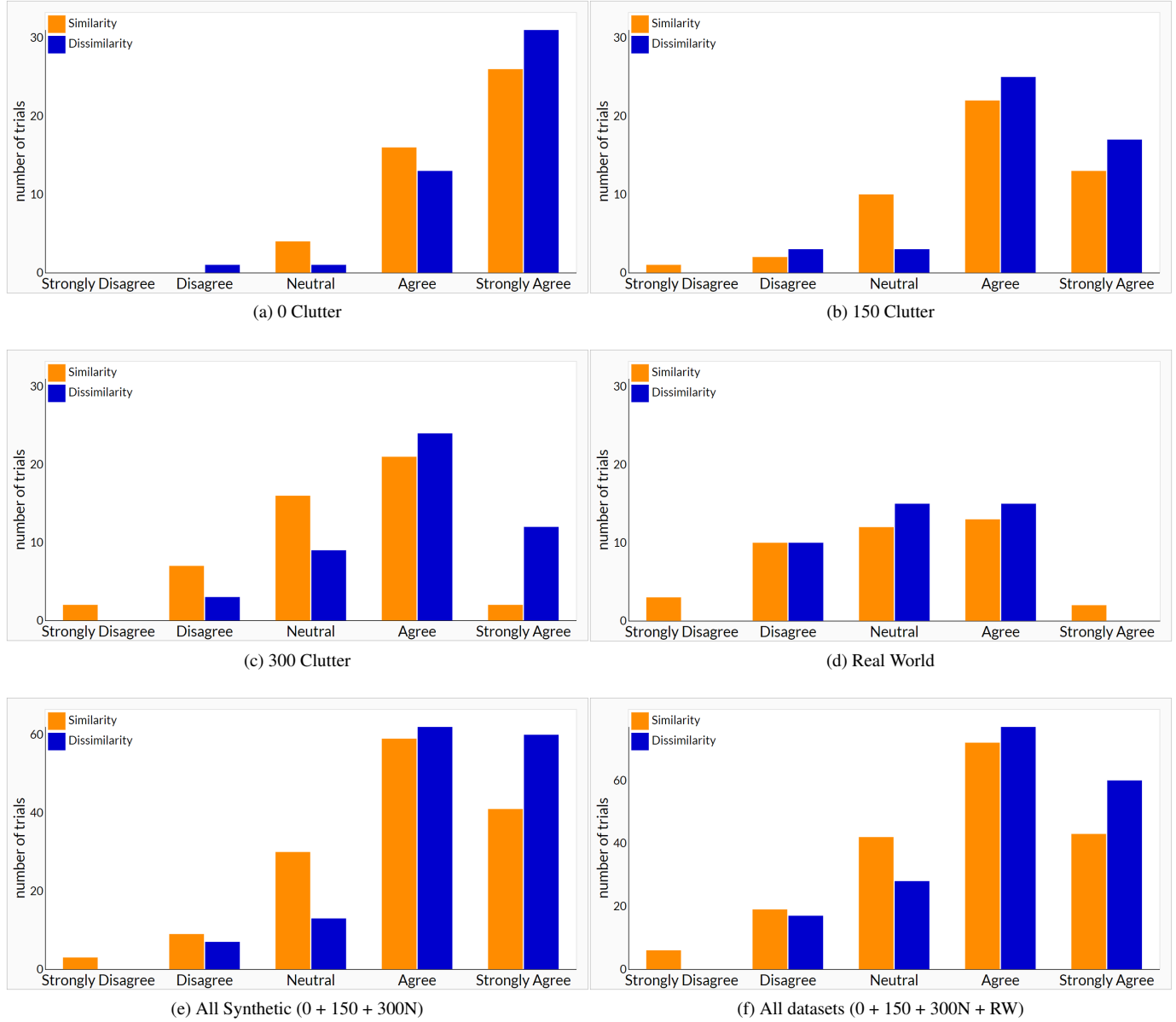


Figure 11: Distribution of the confidence score across the different datasets.

Medians of Likert scale.

- 0N: 2; SIM: 2; DIS: 2
- 150N: 1; SIM: 1; DIS: 1
- 300N: 1; SIM: 0; DIS: 0
- RW: 0; SIM: 0; DIS: 0

2.7.2 Statistical Analysis

We test for significant change across clutter levels (independent of the ordering) using a Chi-square test.

Pearson's Chi-squared test

data: tbl

X-squared = 120.97, df = 12, $p\text{-value} < 2.2e-16$

PostHoc Tests.

0N – 150N Pearson's Chi-squared test

X-squared = 19.789, df = 4, **p-value = 0.0005498**

0N – 300N Pearson's Chi-squared test

X-squared = 52.137, df = 4, **p-value = 1.291e-10**

0N – RW Pearson's Chi-squared test

X-squared = 86.187, df = 4, $p\text{-value} < 2.2e-16$

150N – 300N Pearson's Chi-squared test

X-squared = 11.651, df = 4, **p-value = 0.02014**

150N – RW Pearson's Chi-squared test

X-squared = 43.115, df = 4, **p-value = 9.794e-09**

300N – RW Pearson's Chi-squared test

X-squared = 15.241, df = 4, **p-value = 0.004227**

2.8 Results: Confidence of Marked Clusters (Varying Amount of Clutter and Ordering Strategy)

We show the distribution and statistic for the confidence of the marked clusters in task 2.

2.8.1 Data Distribution and Summary Statistics

See Fig. 11 for data data distribution. Use `t2-confidence-varying-clutter-per-ordering.r` to reproduce the results. The charts have been created with `d3.js`.

2.8.2 Statistical Analysis

The Likert scale results are not ratio scaled. Therefore, we compute Wilcoxon Signed-rank test to see whether there are differences for the different clutter levels.

0N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

$Z = -1.2326$, **p-value = 0.2482**

alternative hypothesis: true mu is not equal to 0

150N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

$Z = -2.5222$, **p-value = 0.01058**

alternative hypothesis: true mu is not equal to 0

300N SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

$Z = -3.7464$, **p-value = 0.0001153**

alternative hypothesis: true mu is not equal to 0

RW SIM vs. DIS Exact Wilcoxon-Pratt Signed-Rank Test

data: y by x (pos, neg)

stratified by block

$Z = -0.75954$, **p-value = 0.4869**

alternative hypothesis: true mu is not equal to 0

3 TASK 3

In the following, we provide all supplementary material for task 3 in our study.

3.1 Datasets and Scripts

Datasets The following datasets are available to analyze the results:

- `TASK3.csv`. Data of all trials including the free text with a description why participants preferred a particular chart. We changed the preference of 4 participants to *NP* – (*no preference*) as they reported that they do not have any preference within their free text field.

R-Scripts We use the following R-scripts for our analysis. Together with the dataset above, the results can be checked and reproduced.

- `t3-statistic-preferences.r` Check statistics for the preferences of users for different clutter levels.

3.2 Data Cleaning and Number of Trials

4 participants ($p.id \in \{10, 12, 43, 45\}$) reported in at least one trial that they did not have any preference (all selected DIS but reported in the full text description that they do not have a preference; 3 with 0 clutter + 1 in 150 clutter.) We removed these participants from the following analysis and from Fig. 12. As a result, **our analysis is based on 27 participants \times 2 trials = 54 trials** (4 participants / 8 trials ignored as no preference).

3.3 Results: Preference of Users towards an Ordering Strategy (Varying Clutter Levels)

In the following, we analyze the preferences of the users towards a specific reordering technique for different clutter levels.

3.3.1 Data Distribution and Summary Statistics

Distribution of preferences, see Fig. 12.

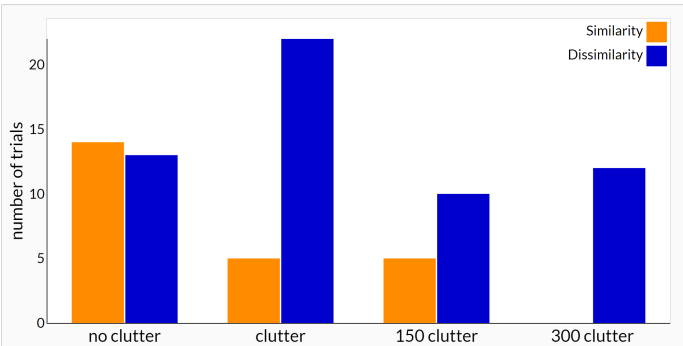


Figure 12: **Task 3: Preference of reordering strategy for different clutter levels.** Only the trials of participants with a clear preference in both trials are considered.

3.3.2 Statistical Analysis

To compute a statistical significance whether people change from a similarity based reordering to a dissimilarity based reordering when changing the clutter level, a McNemar’s Chi-Squared test can be computed as we have dependent samples. The contingency table with all factors can be found in Table 17 and the results of McNemar’s test in Table 18. However, as we have small samples sizes, in particular $b + c = 12 + 3 = 15$ (should be at least 10 or even 25 according to different authors), an exact binomial test is more reliable compared to McNemar’s test [1]. The results can be found in Table 19, which shows a significant result (same as McNemar’s test). The statistic can be found in `t3-statistic-preferences.r`.

		without clutter	
		SIM	DIS
with clutter	SIM	$a = 2$	$b = 12$
	DIS	$c = 3$	$d = 10$

Table 17: **Task 3 Contingency table** illustrating the consistency between responses (preferred layout for no clutter vs. 150+300 clutter).

	statistic	df	p.value	method	data.name
McNemar’s chi-squared	4.27	1.00	0.04	McNemar’s Chi-squared test with continuity correction	contingency.table

Table 18: **Results McNemar’s Chi-squared test** with continuity correction. Significant result.

	statistic	parameter	p.value	conf.int	estimate	null.value	alternative	method	data.name
1	12.00	15.00	0.04	0.52	0.80	0.50	two.sided	Exact binomial test	b and (b + c)
2	12.00	15.00	0.04	0.96	0.80	0.50	two.sided	Exact binomial test	b and (b + c)

Table 19: **Results Exact Binomial test.** Significant difference.

REFERENCES

[1] K. Rufibach. Assessment of paired binary data. *Skeletal Radiology*, 40(1):1–4, Jan 2011. doi: 10.1007/s00256-010-1006-1