Phantom-Words with simultaneous visual presentation - Results

Ansgar D. Endress
City, University of London

Abstract

Abstract (to be written)

1 Predictions

The predictions for the current experiment were unclear. On the one hand, it is plausible that observers might encode entire scenes when they are presented simultaneously. If so, they should not accept phantom-words. On the other hand, statistical learning might operate similarly for simultaneous as for sequential presentation. If so, the results with sequential presentations should be replicated, especially because the shapes appear as distinct individual shapes rather than wholes. Further, presenting the shapes as whole in an object (i.e., in the white on black presentation) might encourage observers to process the combination of shapes as a single hole, leading to the rejection of phantom words.

REMOVED INTERACTION TERM IN GLMM

MAKE SEPARATE TABLES FOR GLMMS

2 Analysis

2.1 Demographics

The main experiment recruited participants from testable minds (https://minds.testable.org/). I pilot experiment recruited participants from first year students at City, University of London (UK). In the latter population, other experiments typically need to exclude 30% to 50% of the sample due to insufficient attention. Unfortunately, the present experiment does not offer a clear performance-based criterion to make sure that participants paid attention to the stimuli, as the task might be genuinely difficult. However, given that our main interest lies in the performance on trials involving phantom-words for participants who succeeded in the statistical learning task, it is more conservative to exclude participants whom might not have paid attention to the task, even if this overestimates the statistical learning abilities.

As a result, I rely on the assumption that earlier statistical learning literature has shown that participants can learn statistical relations *in principle*, and exclude those participants not exceeding an accuracy of 50% on word vs. part-word trials. This criterion led to the removal of 23 and 53 participants from the students and testable samples, respectively.

The pattern of significance was very similar when all participants were excluded, with the following differences. First, for the testable minds sample, performance on the words vs. part-words trials was no longer greater than on the phantom-words vs part-words trials, both when shapes were presented in black on a white background and when the polarity was inverted. Given that earlier, sequential experiments involving phantom-words showed equivalent performance for both trial types, excluding participants is thus more conservative for the current purposes. Second, for the student sample, the performance difference between these trial types also ceased to be significant when all participants are included. Further, when items were presented as black shapes on white background, there was no significant preference for words over part-words, suggesting that some of these participants did not pay attention to the experiment.

Table 1: Demographics of the final sample, after excluding participants whose accuracy on word vs. part-words trials was below 50 percent. For the student population, age and gender have not been recorded due to experimenter error.

Population	Color polarity	N	Females	Males	Other	Age	Age range
testable	black on white	57	32	25	0	30.7	19-59
testable	white on black	51	22	29	0	31.9	18-57
students	black on white	12	0	0	12		Inf–Inf
students	white on black	15	0	0	15		Inf–Inf

The demographics of the remaining participants is given in Table 1; age and gender were not recorded due to experimenter error.

2.2 Analysis by accuracy

We will analyze the results using two types of analyses. First, I compare the performance in the different trial types to the chance level of 50% using Wilcoxon test. To compare performance across trial types, I calculate normalized difference scores, that is, $\frac{\text{accuracy}_{\text{trial type 1}}-\text{accuracy}_{\text{trial type 2}}}{\text{accuracy}_{\text{trial type 1}}+\text{accuracy}_{\text{trial type 2}}}$. These difference scores are the compared to the chance level of zero, again using Wilcoxon tests. I also ask whether any of these results is affected by the color polarity type (i.e., black on white vs. white on black).

Second, we will confirm these results using a set of generalized linear models.

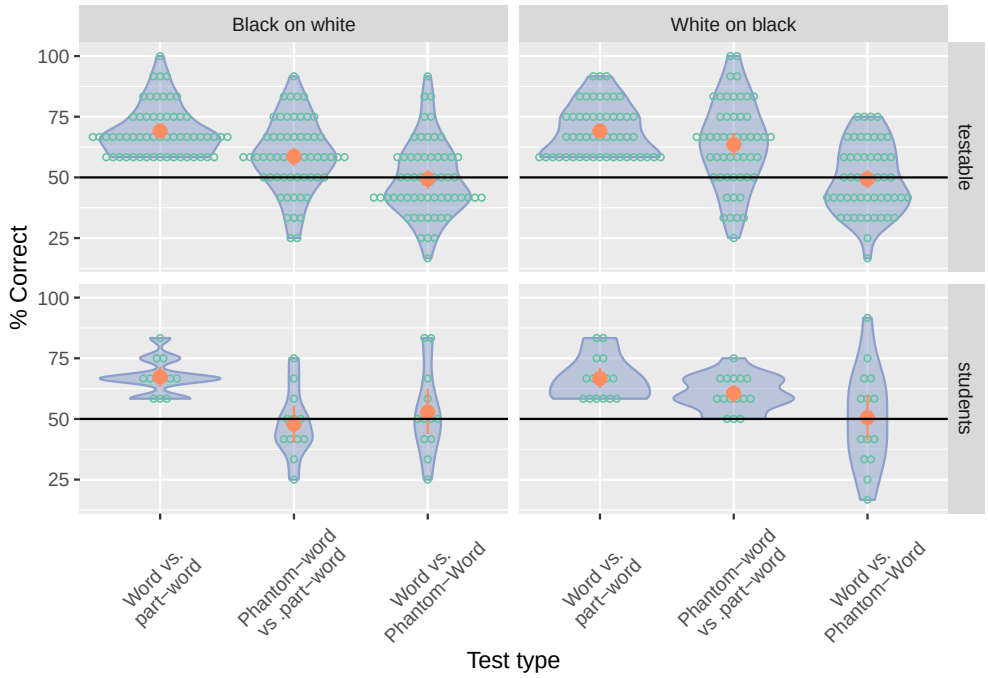

Figure 1: Accuracy in the different trial types (words vs. part-words, phantom-words vs. part-words, and words vs. phantom-words), after exclusion of participants whose performance was below 50% in the word vs. part-word trials. The dots, error bars and violion represent the sample averages, 95% bootstrap confidence intervals and the distribution of the average accuracy for individual participants, respectively. Empty circles represent individual participants.

As shown in Table 2 and Figure 1, participants from the testable minds sample preferred both words and

Table 2: Descriptives of accuracy scores and difference scores, after exclusion of participants whose performance was below 50 percent on word vs. part-word trials. The p value reflects a Wilcoxon test against the chance levels of 50 percent and of 0 for accuracies and difference scores, respectively. The effect of color polarity represents a Wilcoxon test comparing all of these dependent variables as a function of color polarity

test.type	M	SE	p.wilcox
testable - black.on.white (N = 57)			
w.pw	69.152	1.395	0.000
w.phw	49.269	2.065	0.905
phw.pw	58.626	1.985	0.000
d.relative.w.pw.w.phw	0.182	0.021	0.000
d.relative.w.pw.ph.pw	0.091	0.021	0.000
testable - white.on.black (N = 51)			
w.pw	69.118	1.496	0.000
w.phw	49.346	2.012	0.985
phw.pw	63.562	2.516	0.000
d.relative.w.pw.w.phw	0.178	0.021	0.000
d.relative.w.pw.ph.pw	0.056	0.021	0.026
testable - zEffect of color polarity			
w.pw			0.959
w.phw			0.878
phw.pw			0.125
d.relative.w.pw.w.phw			0.784
d.relative.w.pw.ph.pw			0.331
students - black.on.white (N = 12)			
w.pw	67.361	2.262	0.002
w.phw	52.778	5.392	0.623
phw.pw	47.917	4.167	0.765
d.relative.w.pw.w.phw	0.141	0.051	0.019
d.relative.w.pw.ph.pw	0.180	0.045	0.004
students - white.on.black (N = 15)			
w.pw	66.667	2.381	0.001
w.phw	50.556	5.355	0.875
phw.pw	60.556	1.968	0.002
d.relative.w.pw.w.phw	0.165	0.058	0.018
d.relative.w.pw.ph.pw	0.047	0.029	0.143
students - zEffect of color polarity			
w.pw			0.698
w.phw			0.825
phw.pw			0.008
d.relative.w.pw.w.phw			0.807
d.relative.w.pw.ph.pw			0.011

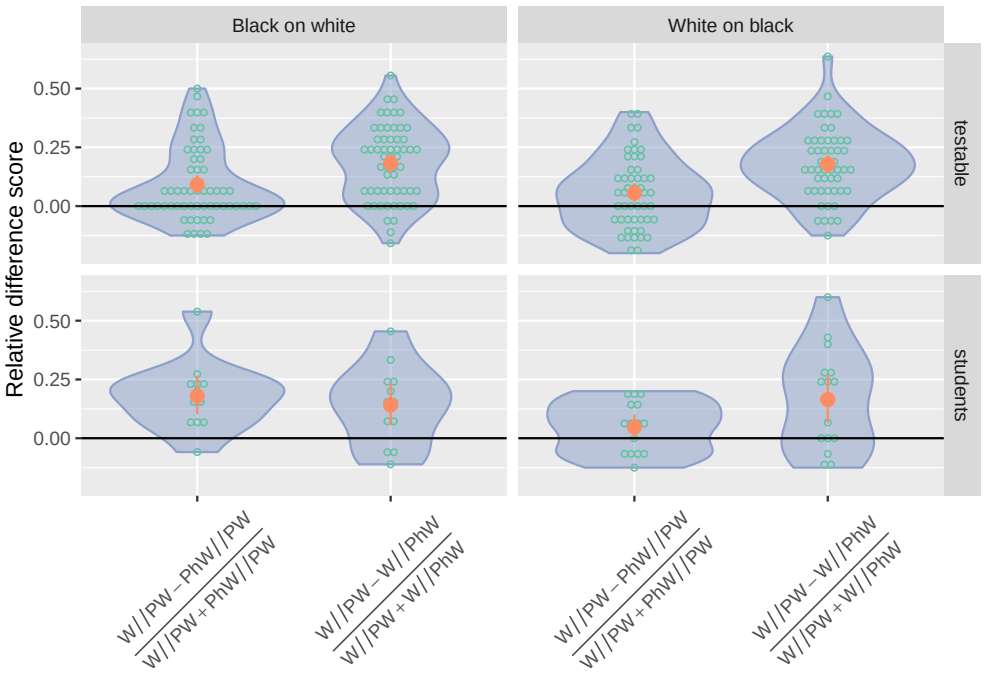

Figure 2: Relative difference scores for contrasts between different trial types (word vs. part-word trials vs. phantom-word vs. part-word trials, and word vs. part-word trials vs. word vs. phantom-word trials), after exclusion of participants whose performance was below 50% in the word vs. part-word trials. The dots, error bars and violion represent the sample averages, 95% bootstrap confidence intervals and the distribution of the difference scores for individual participants, respectively. Empty circles represent individual participants.

phantom-words to part-words. In contrast, the had no preference for words over phantom-words. Similar results were obtained for both color polarity types, with no discernible effect of color polarity type. The results for student sample were similar, except that there was no preference for phantom-words over part-words when black shapes were presented on a white background, and that the preference for phantom-words over part-words was significantly greater when white shapes were presented on a black background. However, given that the result relied on only 12 participants, I tentatively conclude that the current experiment replicate [Endress and Mehler, 2009] and [Endress and Langus, 2017], in that phantom-words are preferred to part-words, and there is no marked preference for words over phantom-words.

To compare performance in the different trial types, I calculated the difference scores mentioned above. As shown in Table 2 and Figure 2, participants from the testable sample performed much better on word vs. part-word trials than on word vs. phantom-word trials, irrespective of the color polarity type. This suggests that participants find discriminations based on TPs much easier than discriminations based on frequency of occurrence, which is problematic if statistical learning leads to memory for units. However, performance was also somewhat better for word vs. part-word trials than for phantom-word vs. part-word trials, suggesting that we cannot rule out that participants might also have some ability to track frequencies of occurrence. However, the corresponding difference score was much smaller than that comparing words vs. part-word and word vs. phantom-word trials, and ceased to be significant when all participants were included.

In the student sample, results were similar, except that I did not detect a performance difference between word vs. part-word and phantom-word vs. part-word trials when white shapes were presented on a black background.

Table 3: Results of generalized linear mixed models for trial-by-trial responses, after exclusion of participants whose performance was below 50% in the word vs. part-word trials.

	Log-odds			Odd ratios				
term	Estimate	SE	CI	Estimate	SE	CI	t	p
testable - w.pw vs. w.phw								
test.typew.pw	0.834	0.082	[0.674, 0.995]	2.303	0.189	[1.96, 2.7]	10.190	0.000
color.typewhite.on.black	0.001	0.082	[-0.159, 0.161]	1.001	0.082	[0.853, 1.17]	0.011	0.991
testable - w.pw vs. phw.pw								
test.typew.pw	0.460	0.114	[0.237, 0.683]	1.584	0.180	[1.27, 1.98]	4.047	0.000
color.typewhite.on.black	0.209	0.117	[-0.02, 0.437]	1.232	0.144	[0.98, 1.55]	1.788	0.074
test.typew.pw:color.typewhite.on.black	-0.210	0.166	[-0.536, 0.116]	0.810	0.135	[0.585, 1.12]	-1.263	0.207
students - w.pw vs. w.phw								
test.typew.pw	0.646	0.163	[0.328, 0.965]	1.909	0.310	[1.39, 2.62]	3.977	0.000
color.typewhite.on.black	-0.062	0.166	[-0.387, 0.263]	0.940	0.156	[0.679, 1.3]	-0.374	0.708
students - w.pw vs. phw.pw								
test.typew.pw	0.808	0.244	[0.33, 1.29]	2.243	0.547	[1.39, 3.62]	3.315	0.001
color.typewhite.on.black	0.512	0.226	[0.0691, 0.955]	1.669	0.377	[1.07, 2.6]	2.266	0.023
test.typew.pw:color.typewhite.on.black	-0.543	0.328	[-1.19, 0.0997]	0.581	0.191	[0.305, 1.1]	-1.656	0.098

I confirmed these results using generalized linear mixed models with the fixed factor predictors trial type and and color polarity as well as their interaction, and a random intercept for participants. I fitted separate model for each sample (testable vs. students) and trial contrast (word vs. part-word trials vs. word vs. phantom-word trials and word vs. part-words and phantom-word vs. part-word trials). The models showed that performance on word vs. part-word trials is significantly better than for word vs. phantom-word trials. In the testable population, they also showed that performance on word vs. part-word trials was significantly better than on phantom-word vs. part-word trials, though this predictor was not significant in the student population. Further, the odds ratio associated with the former contrast was almost twice as high as that from the latter contrast.

There were generally no main effects or interactions with polarity type, though students performed somewhat better for black on white displays.

3 Discussion

4 Appendix 1: Results with the full sample

As shown in Table **??** and Figure **??**, participants from the testable minds sample preferred both words and phantom-words to part-words. In contrast, the had no preference for words over phantom-words. Similar results were obtained for both color polarity types, with no discernible effect of color polarity type. In the complete student sample, in contrast, the only significant preference was that for words over part-words, but only when black shapes were presented on a white background. These contrasting results presumably reflect the finding that, in other experiments that can implement attention check manipulations, 30% to 50% of such samples need to be excluded due to insufficient attention.

As a result, I tentatively conclude that, at least in the testable minds sample, the current experiment replicate [Endress and Mehler, 2009] and [Endress and Langus, 2017], in that phantom-words are preferred to part-words, and there is no marked preference for words over phantom-words.

To compare performance across trial types, I calculated the difference scores mentioned above. As shown in Table **??** and Figure **??**, participants from the testable sample performed much better on word vs. part-word trials than on word vs. phantom-word trials, irrespective of the color polarity type. This suggests that participants find discriminations based on TPs much easier than discriminations based on frequency of occurrence, which is problematic if statistical learning leads to memory for units. In contrast to the results with the restricted sample, performance was not significantly different between word vs. part-word trials and phantom-word vs. part-word trials.

In the student sample, only the comparison between word vs. part-word trials and word vs. phantom-word trials was statistically different from zero, but only when black shapes were presented on a white background. However, and as mentioned above, this sample likely contained a sizable proportion of participants who paid little attention to the stimuli.

I confirmed these results using generalized linear mixed models with the fixed factor predictors trial type and and color polarity as well as their interaction, and a random intercept for participants. I fitted separate model for each sample (testable vs. students) and trial contrast (word vs. part-word trials vs. word vs. phantom-word trials and word vs. part-words and phantom-word vs. part-word trials). As shown in Table **??**, the models showed that performance on word vs. part-word trials is significantly better than for word vs. phantom-word trials. In contrast, there was no difference between word vs part-word and phantom-word vs. part-word trials.

There were generally no main effects or interactions with polarity type.

References

Ansgar D. Endress and A Langus. Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, 92:37–64, 2017. doi: 10.1016/j.cogpsych.2016.11.004.

Ansgar D. Endress and Jacques Mehler. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, 60(3):351–367, 2009. doi: 10.1016/j.jml.2008.10.003.