Statistical Learning Experiment

Summary of datasets

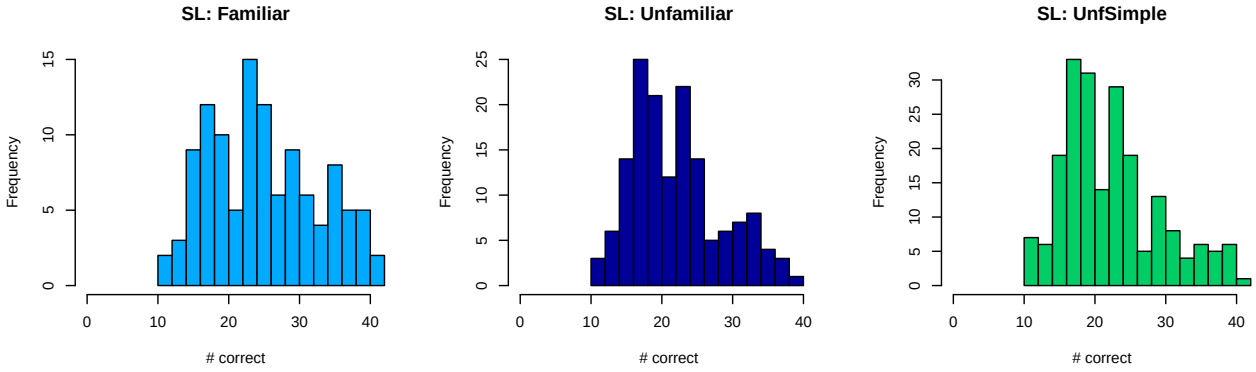
For Part 1, 160 people were run on AMT in an experiment that took 15-20 minutes and paid $3.40USD. Before analysing the data 14 were excluded for getting less than three of the four words in the sequence right (9 in **Familiar** and 5 in **Unfamiliar**). For Part 2, the remaining 146 were invited back to do a 20-minute follow-up that paid $4USD. Of those, 135 returned (a retention rate of 90%). Three of them reported less than three words in the sequence and were excluded. This left 132 people in the dataset, with 52 in the **Familiar** condition and 80 in the **Unfamiliar** condition. Across the whole dataset, 73 (i.e., 55.3%) of them were male and 129 (i.e., 97.7%) were from the US. Ages ranged from 20 to 69 with a mean of 36.1.

Statistical Learning Tests

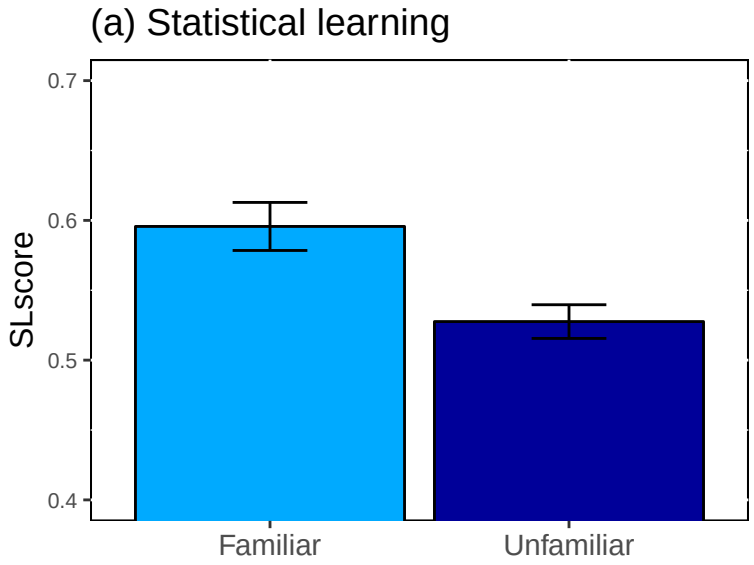
Our first step is just to make sure people are behaving sensibly on the SL tasks, and to check if there are differences in overall performance by condition.

Part 1

First let's look at the histograms of all responses to determine whether they look like those reported in Seigelman et al., 2017, and to see whether the data look normal. This includes the **Familiar** and **Unfamiliar** conditions, as well as **unfSimple** (explained below).



These look pretty reasonable. Is the accuracy different between conditions?

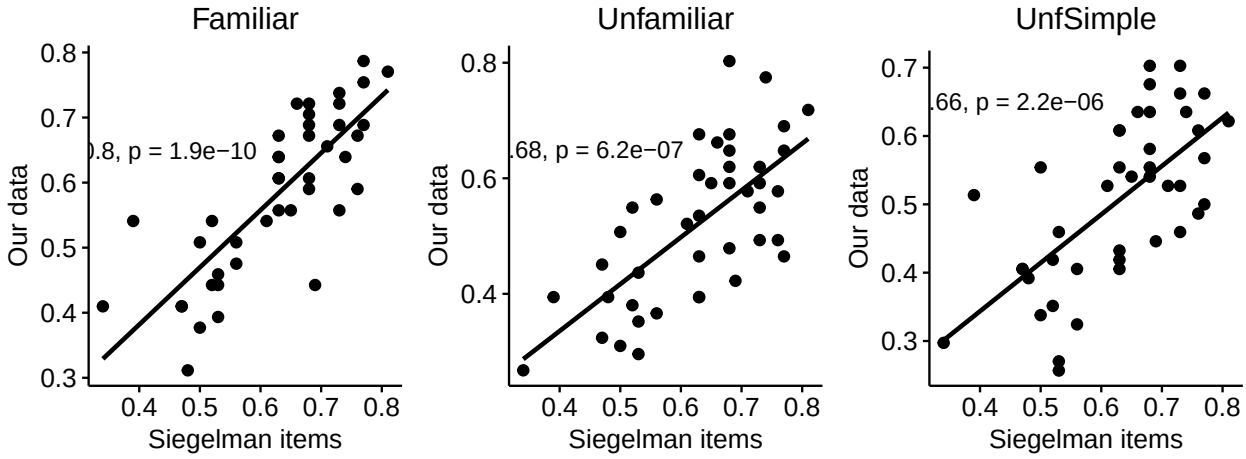


It looks like indeed there is a difference in performance. To evaluate this statistically we ran a t-test with accuracy as our outcome variable and condition as predictor.

Performance in both conditions is different from chance (**Familiar**: t(112)= 11.45, p=0, d=1.077; **Unfamiliar**: t(150)= 10.69, p=0, d=0.87).

The results of the t-test are also significant (t(211)=3.24, p=0.0014, d=0.415). A Bayesian t-test with default priors for effect size finds that the data were 25.0846356 in favour of the alternative hypothesis (i.e., were 25.0846356 times more likely to occur under the model that performance is different by condition).

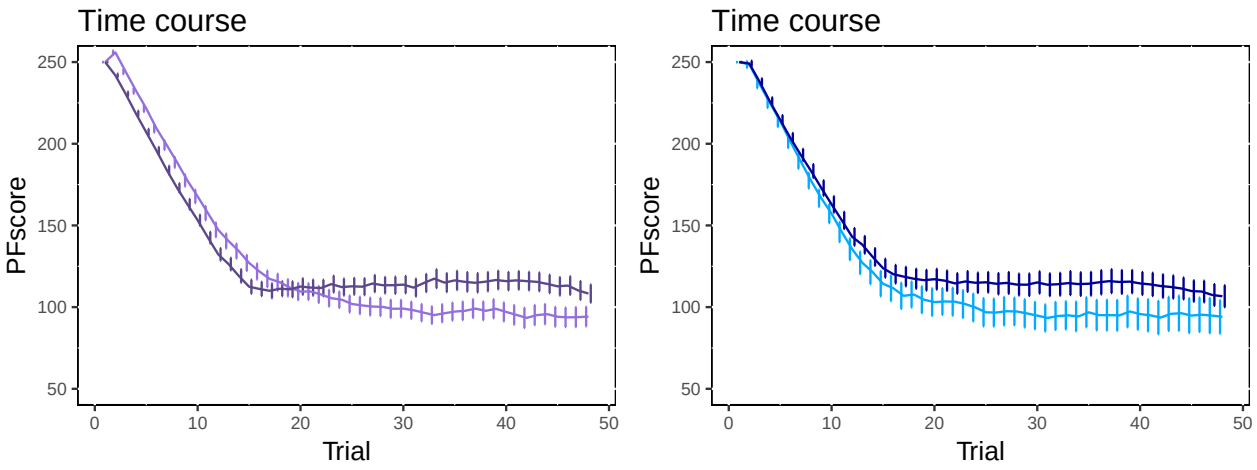
We might also care about whether people's behaviour showed the same *pattern* as Siegelman et al., 2017: that is, were the test items that their participants found hard the same test items that our participants found hard? We can evaluate this by calculating the correlation between performance on each item between them (which they reported) and our data. It is evident that all correlations are highly significant.



Perceptual Fluency Tasks

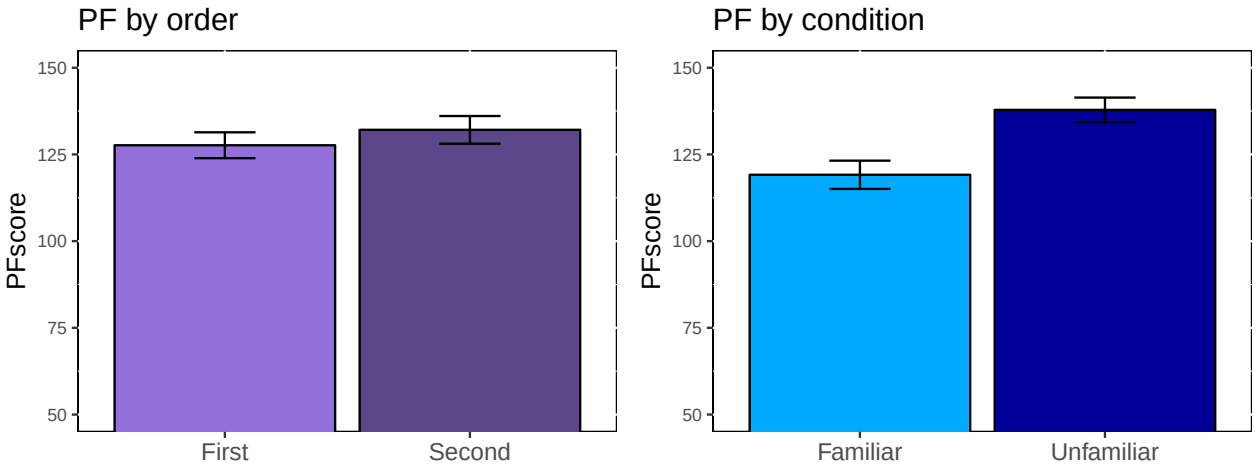
So far it seems that the statistical learning tasks are behaving as expected, and we're getting significant differences in performances based on the visual complexity / familiarity of the stimuli involved. Let's also check that the perceptual fluency tasks are behaving in a sensible fashion. This is a bit harder because they are new tasks so there's no published literature to compare to, but as a first stab we can look at what happens to the duration that the target item flashes over the course of the task (which I'll refer to as the *lag time*). Remember that it gets faster when people are more accurate, so we should expect that in easier tasks it should end up being on average faster. We should also expect it to stabilise somewhere.

We can make two comparisons. First, remember that each person saw two perceptual fluency tasks in a row. We might be curious about whether performance changes between the tasks, either due to fatigue or practice effects. Thus, the figure on the left below shows overall performance on the First and Second task. Second, some people saw perceptual fluency tasks with **Familiar** items, and some with **Unfamiliar** items. If the length that the item flashes is really related to their complexity, we would expect that in the **Familiar** condition it doesn't need to flash as long. This is shown in the figure on the right.



Some things are immediately obvious. First, reassuringly, the time each stimulus flashed does go down, and doesn't bobble around. Second, it appears as though there are differences in both order (people are faster for the first task, so fatigue may be a factor) and condition (people are faster for the **Familiar** items, which is nice). However, we don't know if these effects are significant.

Unfortunately, we can't run statistical tests on multiple different trials, because the trials are all very non-trivially dependent on each other. We therefore feed each person's average lag time in and test whether these differ by order or condition.



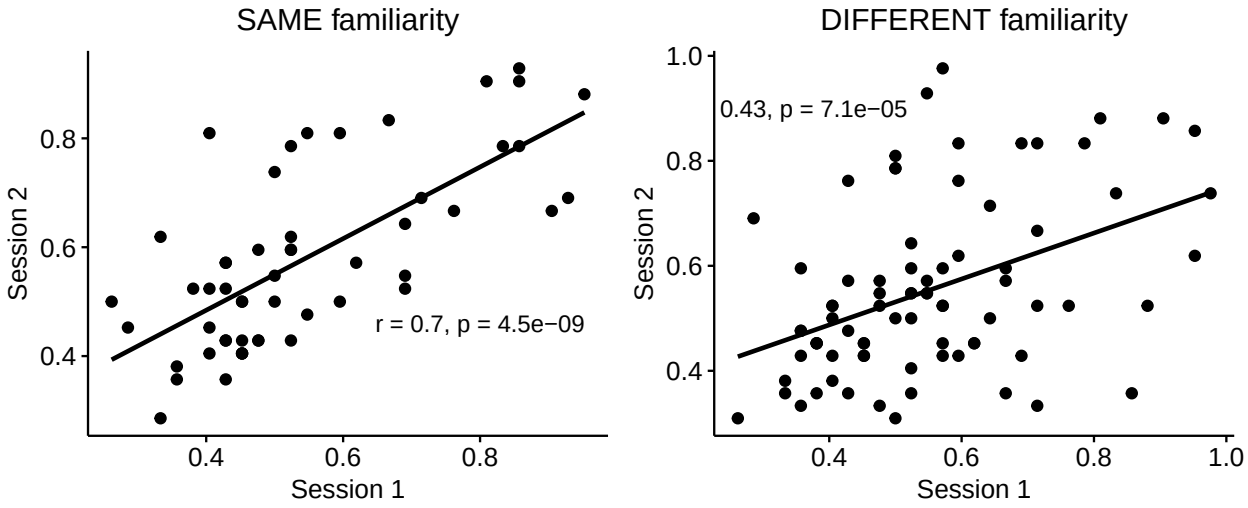
Is this significant? In order to evaluate this, I ran two separate t-tests, one looking at lagtime by test order (i.e., corresponding to the figure on the left) and one at lagtime by condition (i.e., corresponding to the figure on the right). Order was not significant (t(260.1)=-0.81, p=0.4172, d=0.1) but condition was (t(240.1)=-3.46, p=0.0006, d=0.431). This is pretty cool: it means that, as predicted, people did better on the perceptual fluency task when the stimulus items involved were **Familiar** ones. That said, performance on the two back-to-back perceptual fluency tasks was of course also highly correlated with one another (r = 0.5424, p<0.0001).

Finally, next, we really get to look at the meat of our experiment: how these measures relate to each other.

Relation of statistical learning to statistical learning

Our first question is how people's statistical learning performance on the two separate days compares to each other. As you'll recall, some people saw the **Same** kinds of stimulus sets on both days (either both **Familiar** or both **Unfamiliar**) and some people saw **Different** types of stimulus sets (some saw **Familiar** on Session 1 and **Unfamiliar** on Session 2, and some vice-versa). In no case did anybody see the same *items* on both days – even if it was the same kind of stimulus, the items were different. Thus, for instance, if your stimulus contained a star on Session 1 then it did not contain a star on Session 2. This means that any correlation in performance cannot be explained by trivial familiarity with the specific stimulus items.

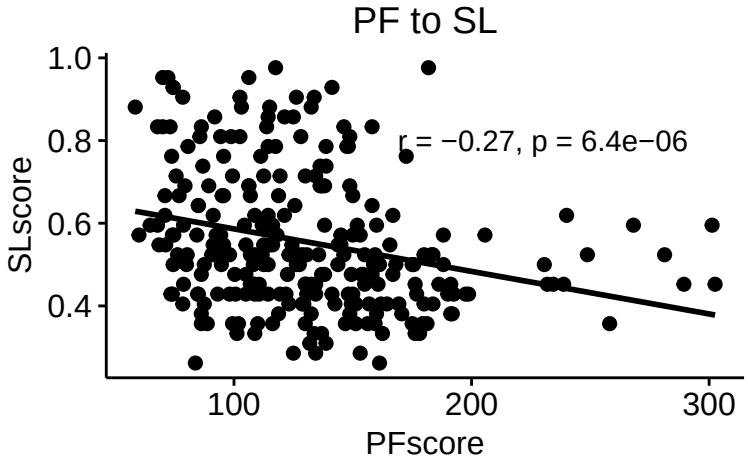
Our key question is whether people who saw the **Same** kind of stimulus set on both days had a higher correlation in their statistical learning accuracy between tasks than people who saw **Different** stimulus sets. If individual differences in performance on this task are *just* about statistical learning, these correlations should be the same, since the underlying statistical learning is exactly the same – all that differs is the stimulus type. If, on the other hand, perceptual fluency plays an important role, then one would expect a larger correlation for people who saw the **Same** stimulus sets and a smaller (or nonexistent) one for people who saw **Different** ones. These correlations are shown below.



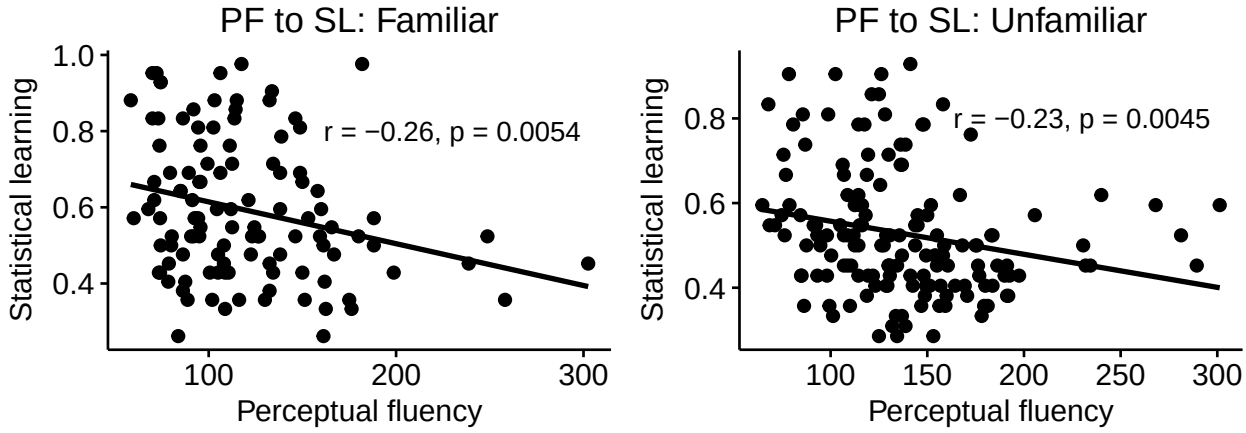
It is indeed apparent that **Same** appears to correlate with each other much better than **Different**. We can statistically test this with the Fisher r-to-z transformation: z=2.2554213, p=0.0241069. So these are indeed significantly different.

Relation of statistical learning to perceptual fluency

Given the hypotheses of our study, we're also interested in determining whether performance on the perceptual fluency task predicts performance on the statistical learning task. I set it up so that for everyone, the first perceptual fluency task used the same stimuli as the statistical learning test on Session 1. Conversely, the second perceptual fluency task used the same stimuli as the stastistical learning test on Session 2. The obvious thing to do first is just to pool everything together, and see to what extent each perceptual fluency task predicts statistical learning. That is shown below.



So having a shorter lag time on the perceptual fluency task is correlated with being more accurate on the statistical learning task. We can also break it down by whether the stimuli were **Familiar** or **Unfamiliar**. .



This is kind of interesting. It's cool that the PF test does a reasonable job predicting statistical learning performance. It's also interesting that the relationship is stronger for the **Unfamiliar** stimuli (possibly because the **Familiar** ones are so overlearned there's less room for individual differences?). It should be noted that the strength of the relationship between perceptual fluency and statistical learning is less strong than between statistical learning on two consecutive days, even when the stimuli are different. So that's an indication that there is some important component of statistical learning the perceptual fluency doesn't touch. Still, it's kind of impressive how much does just come down to perceptual fluency.

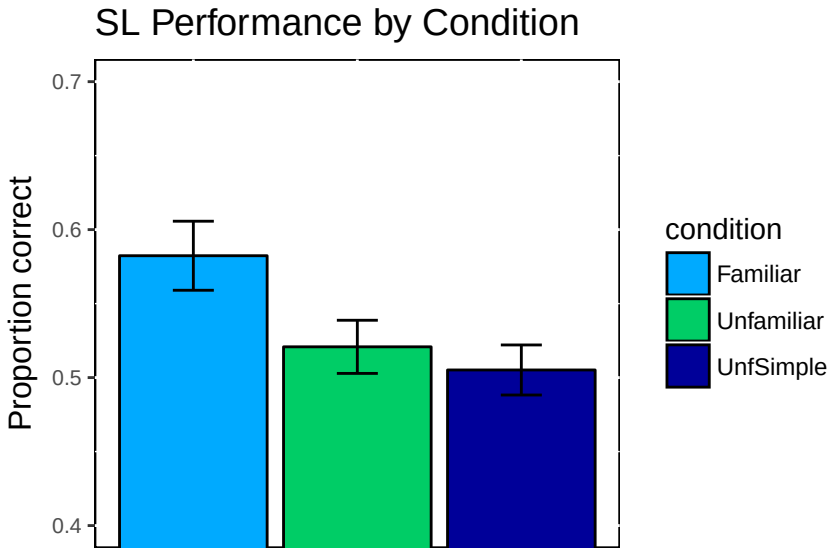
Overall

(1) People don't do as well on statistical learning tasks that are identical in terms of statistical complexity but that have more complex stimuli. (2) Statistical learning performance from one day to another is more highly correlated when the items involved are similar, even though in both cases the statistical complexity is the same. (3) Performance on an independent task of perceptual fluency predicts accuracy on the statistical learning task from a previous day.

Compare to simpler unknown stimuli

Is this familiarity or complexity? Let's compare SL performance in the **Familiar** and **Unfamiliar** conditions to a condition that uses unfamiliar stimuli that are also substantially simpler visually. For this we ran 80 people on AMT in exactly the same experiment but with different stimuli, which again took 15-20 minutes and paid $3.40USD. Before analysing the data 6 were excluded for getting less than three of the four words in the sequence right. This left 74 people in the dataset, in what well call the **unfSimple** condition (to contrast with the previous unfamiliar hard condition). In this condition, 46 (i.e., 62.2%) were male and 70 (i.e., 94.6%) were from the US. Ages ranged from 21 to 62 with a mean of 36.4.

Is the accuracy different between conditions? To look at this we compare SL performance on all of the *first session* datasets (not like the one before for **Familiar** and **Unfamiliar**, which included both days). The reason we just look at first session is to make **unfSimple** comparable to the others, since it had only one session – it doesn't matter much but seems more principled. Anyway, the result is shown below.



It looks like indeed there is a difference in performance. To evaluate this statistically we ran a one-way ANOVA with accuracy as our outcome variable and condition as predictor.

The results of the ANOVA are significant ($F(2,203)=4.2743$, $p=0.0152$, $etaSq = 0.0404$). We did three post-hoc t-tests. The difference between **Familiar** and **unfSimple** was significant ($t(113.8)=2.68$, $p=0.0085$, $d=0.473$), as was the difference between **Familiar** and **Unfamiliar** ($t(117)=2.09$, $p=0.0387$, $d=0.37$). However, the two unfamiliar conditions were not significantly different from each other ($t(142.1)=-0.63$, $p=0.5271$, $d=0.105$).