A best-case scenario test of the Semantic Priming Stroop Task; can specific thoughts be found?

Bachelor thesis by:

Chiel Kienhuis – s1692151

c.kienhuis@student.utwente.nl

Supervisor:

Martin Schmettow

m.schmettow@utwente.nl

Second Supervisor

Simone Borsci

s.borsci@utwente.nl

Abstract

Throughout the years in the science of Psychology, implicit methods have been created in order to study the intentions behind human behavior. Testing whether they work is however also a very important part of this process. Recently, the replicability crisis within Psychology has disconfirmed many theories that were thought to be true, implicit methods included. More often than not, the low replicability is because of the p-value and the bad implications the common practices that go with it. In this study we look at an implicit method first used by a previous study of high reputation. This method is called the Semantic Priming Stroop task and is about how associations within the mind can be found through delay within the Stroop task. It has not been tested thoroughly since the original paper and it becomes clear in this study that the method does not work at its core. We calculated with, multilevel models and credibility intervals that the found effect is practically zero as the difference between associative and non-associative stimuli is around 3ms. It can be concluded that this version of the original Stroop task is not applicable in general.

Introduction

Reading someone’s mind is a common misconception about the skillset of a Psychologist. But what if we actually can? In a sense, this is what implicit methods are meant to do; finding out what a person is thinking without asking them directly. Finding out what a person is thinking gives more insight in the reasoning behind their behavior, as just observing it is not enough to draw conclusions from it.

Implicit Methods

Experiments that rely on self-reporting are usually biased due to some kind of deception that the participants apply, in some cases unintentional (Paulhus, 1984). The amount of deception tends to increase when the usage of the data is intended for the public, but even in anonymous studies participants tend to lie. These lies are usually in the form of socially desirable answers or answers as to how the participant would like to be ideally, or how the participant view themselves subconsciously without it being the reality (Paulhus, 1984). Experiments that concern how people think about things could be embarrassing to some if it is about specific topics that one does not want to speak about with strangers (Kihlstrom, 2004). Other things that people simply do not know about themselves are also impossible to test in self-report experiments.

Implicit methods are characterised by how they measure certain things without the participant knowing or based on, for example, reflexes. An example of the things that can be achieved by utilizing implicit measure tasks is the well-known Implicit Association test, which can be used for various topics such as racism, prejudice or your self-esteem (Greenwald, McGhee & Schwartz, 1998). Most of the implicit measures are assessed with reaction time and their delays, examples include both the Stroop task (Stroop, 1935) and the mplicit association test, although this is more reliant on what answers you give and is more accurate with lower response times. The implicit measures that have been discovered are relatively valid and to have another one added to this list with which you can know what someone is thinking about opens up a lot of new possibilities for research.

A notable example within the field of Implicit methods is the experimental paradigm called the Stroop task, a color naming task that tests the interference that mental load has on the thought process. This effect is also called the ‘Stroop effect’ (Stroop, 1935). At the time this implicit method was a big discovery and many researchers set out to verify whether the so-called ‘Stroop effect’ actually works. Furthermore it was also explored whether the ‘Stroop effect’ holds true in other applications or versions of the original task (MacLeod, 2011). More ways to use implicit methods is something that could help the field of Psychology as a whole.

Semantic Priming Stroop Task

One specific example of the variations that originated from the Stroop task (Stroop, 1935) is the ‘Semantic Priming Stroop task’, a variation explicitly focused on testing whether the participant undergoes a distinct thought. This is supposedly done by priming the participant with an idea or association and quickly follow up this priming with a color naming task. A delay in reaction time would then imply that the participant was still thinking about the used priming stimuli. This paradigm is special as it was first used in the study of Sparrow, Liu & Wagner (2011). In this study they made use of this Stroop task in order to test whether the participant thought about using search engines when asked certain trivia questions. Their results show that this was indeed the case.

The Sparrow paper (2011) consisted out of 4 different studies, every study using a different method. The first study that was done made use of the Semantic Priming Stroop task, a version they proposed. They primed the participant semantically with several yes or no trivia questions were asked and each question was followed up by a color naming task. In this variation of the Stroop task, the words in the color naming task were not related to the color but instead to the stimuli the participant was primed to think about and would thus induce a mental load. The implication of their paper was big since, as of today, it has been cited over 900 times already. However, something to note is that, since Sparrow and colleagues (2011) were the first to use this Stroop task variation, it is unsure whether the results they got are valid or not. Furthermore, this specific paradigm has not been replicated a lot. One example where it was used successfully was in a study by Schmettow, Noordzij & Mundt (2013), just 1 replication however is not a good indication to how valid the task actually is and what parameters it requires.

Replication Crisis

There is still a lot of uncharted territory in the science of Psychology which requires research, including variations of the Stroop task specifically. Important within science as a whole is replicating studies to check whether the found effects are real or noise. The problem that we are now facing with the entire field of Psychology is that in recent years, many findings do not reproduce in follow-up studies (Gelman, 2018). Theories like Mindfulness (a state achievable through meditation), Embodied cognition (where aspects of your body contribute to your thoughts), and Ego depletion (where you supposedly have a finite amount of willpower), were first broadly accepted, but have been debunked in the last couple of years along with many other findings (Gelman, 2018).

There are a couple of reasons why the replication crisis is currently happening and it is not just because of various intentional fraudulent behaviours such as from Diederik Stapel who personally fabricated outcomes on studies in order to fit certain expectations. A much bigger contributor is the way academic papers in the field of Psychology are regarded. According to Giner-Sorolla (2012), the current bottleneck created by too many studies and too little publication outlets influences the way psychological findings are handled. In order to get a publication, researchers are inclined to make sure that their results are of statistical significance. Additionally, the significance is mostly determined by whether the found *p*-value is below the 0.05 threshold (Giner-Sorolla, 2012).

The need for statistical significance might seem good at first glance when starting out with the statistical methods, as it makes it so that researchers will do anything to get a good experiment going. However, researchers have a certain degree of freedom within their research design, which allows them to report seemingly good results while these may not be in line with their initial findings. The degree of freedom allows researchers to make decisions after their initial plan in order to find a result that is of statistical significance, and to only report what was considered to be true instead of reporting everything that happened during the study (Simmons, Nelson & Simonsohn, 2011).

Test subjects are often complex human beings, meaning there are many confounding variables that can influence the outcome of an experiment. Considering the simple ‘yes-or-no’ mentality of the 0.05 threshold for significant results, this may be the reason that a lot of significant results do not apply in the real world. Confounding variables that may have influenced the results can be overlooked or unreported in favor of reporting significant findings. To properly conduct an experiment, the experimenter should therefore be strict in reporting what happens during the experiment, which environmental factors might affect the test subject and so on.

To show how misuse of the p-value influences false-positive results, Simmons and colleagues (2011) conducted two experiments that had intentionally nonsensical hypotheses; the effect of old versus new songs on how old one feels. Several commonly used practices were applied to influence the results, such as making a distinction between genders even when this factor is irrelevant, increasing sample size and thereby increasing degrees of freedom, testing the average of the dependent variables on top of the regular testing, and dropping one condition that did not result in a significant *p*-value. After using these practices that are commonly used in regular studies, they reported a 60.7% rate of the experiment having a significant result, while the hypotheses were nonsensical from the start.

The experiment by Simmons and colleagues (2011) sheds a light on how the researcher's degree of freedom can influence the *p*-value and create significant results. Besides the methods already researched, Simmons and colleagues (2011) name several more ways in which researchers may influence their results, such as choosing among dependent variables, stopping the data collection early because a significant result was already found, excluding certain participants, deciding whether early data was part of a pilot study or the real one. With all of these options it is understandable how the replication crisis started.

One category of experiments that is especially fallible is priming, which includes the Semantic Stroop Task by Sparrow and colleagues (2011). Various replication studies have reported that the results of the original studies were not found after all (Doyen, Klein, Pichon & Cleeremans, 2012). This in turn has spawned scepticism about the replications themselves in the context of experiments that include priming. Researchers like Cesario (2014) and Klein and his colleagues (2014) made claims that the replications are not direct copies of the original experiment and therefore fallible. In the case of the Sparrow paper however, the big issue related to the replication crisis is that their used Stroop task variation has not been tested enough times to know whether or not is a valid test. This is why this study will not replicate the study of Sparrow and colleagues (2011), but instead put the Semantic Priming Stroop task itself to the test.

The Current Study

In this research, an effort will be made to test the implicit methods test used in the Sparrow study, a paper with a high reputation but with a low amount of replication. The experiment will be conducted in a way with a best-case scenario. This scenario contains stimuli within the semantic priming that use general knowledge in order to make it unlikely that the person does not think about the associations that are implied. The experiment is very basic in the sense that it uses only 1 independent variable, which is the stimuli that either has an implied association or not. If it does not work in this very basic structure with a minimum amount of possible confounding variables, it casts doubt on the previous studies that used this task in its experiments. Considering the replication crisis, an experimental paradigm that barely has any replications on its name plus has been used without verifying its validity seems questionable. Therefore, we expect the experiment to fail in the sense that there will not be a big difference in response times between the two conditions.

This leads to the actual experiment, requiring optimal conditions to achieve the best-case scenario that was mentioned before. An important condition is having repeated measures, meaning that the same trial will be performed at least twice to prevent random delays to be interpreted as an effect. In the original experiment by Sparrow et al. (2011) there were only 16 trials, which is not a high number by any means. The picture-word pairs that associate with a certain thing should be known by many, if not all, participants so the association should be well-known. These include things like for example having a picture of a wolf and the word 'Grandma' to indicate the story of 'Red Riding Hood'. Furthermore the environment should be as non-distracting as possible to minimize the possible confounding variables in the mix. Lastly, the analysis of the data is important; all the data will be tested both on the population level (where it tests if the effect is present on average) and the participant level (where it tests if the effect is present for some should it not be on average). When all of these conditions are applied, the following research question can be answered by checking whether an effect can be found:

'**Under optimal conditions, does the Semantic Priming Stroop task work?**'

Looking at previous experiments and how they could be so easily biased by common practices involving the degrees of freedom and the *p*-value (Simmons et al., 2011), the data analysis will be using a different method to calculate the results. The method of Multi-level models using the Ex-Gaussian sample will give estimates on both the individual level and on the population level that do not rely on *p*-values and degrees of freedom. Instead the analysis of the data will make use of credibility intervals as these give more information about the findings in general, for example to what extent the effect exists if at all, and open up interpretation instead of giving a yes or no answer (Cummings, 2008).

Methods

Stimuli

A total of 100 different stimuli in the form of word picture pairs were used. 50 of them had an association connected to them and 50 of them did not. These stimuli were created and selected in a specific procedure by two separate researchers. This is important since the stimuli are vital to the best-case scenario that was proposed in the introduction. The associative stimuli should be good enough that most people know about the association without the researchers having to inform them about it.

First of all, the associations were considered to be something more specific than a simple thought. For example, a picture of a bowl with yoghurt and the word 'Spoon' to indicate 'Breakfast' or 'Dessert' was not enough of an association to get the participants attention. The association needed to be something that has a narrative. Good examples of these are 'Grandma & Wolf' for 'Red Riding Hood', 'Iceberg & Ship' for 'Titanic'. These narratives impose a stronger mental load than 'Dessert' due to the story it carries. This detail and the idea that these are well known stories made sure there were a lot of fairy tales and movies among the stimuli due to the high chance of people knowing about these stories. Other narratives within the stimuli concerned controversial or shocking topics like: 'Slavery', '9/11' and 'School Shootings'. These were chosen because these are stories that are nearly common knowledge. Other than the associations being known, the words and pictures individually needed to be abstract. Only together should it convey the association.

In order to make the stimuli while following these guidelines, both researchers came up with around 25 to 30 or so associations made up of a picture and a word. These were then tested on the other researcher by means of letting them figure out what the association was that belonged to the word picture pair that the first researcher came up with. If the answer matched the association, the word picture pair would be approved. This kept going on until there was a total of 50 valid word picture pairs that indicated a specific association. After this was finished, 50 non-associative word picture pairs were created through getting random pictures and words to match together. These were quickly checked by the researcher to make sure they did not implicate any association.

After all the stimuli were completed, they were duplicated once to achieve the 200 stimuli necessary for the experiment. This would also make it possible to look at consistency over the same stimulus.

Design

The experiment consisted out of 200 trials for the participant to complete. These trials were divided in two conditions; Pairs that indicated an association (A) and pairs that did not indicate an association (N). The 200 trials consisted out of 50 different 'A' trials and 50 different 'N' trials which were randomly distributed twice along the experiment. The trials are verified after the experiment in order to verify the associative power they were meant to convey. This repeated measures design was necessary to make sure that when a delay occurs, it should occur again in the repeated trial. This also gave extra data per participant which was useful as the aimed sample size was not very big at 40. After every 25 trials there was a break that the participant could use for however long they preferred.

Procedure

The experiment started out by having the participant reading and signing an informed consent which stated that all the data was anonymous and the data would be kept confidential but available for future research. The participant gave their age and gender, afterwards the experimenters gave a short introduction and then the actual experiment started. The programme then gave them further instructions in the native language of the participant.

During the experiment, participants had to complete trials. During every trial the participant got shown a picture and afterwards a coloured word and needed to press the correct arrow key that corresponded with the right colour of the word. These were Red, Green or Blue which corresponded with left (←), down (↓) and right (→) respectively. The words had a variety of topics but did not include colours themselves. Examples are words like: Pen, Grandma, Wolf etc. The participants first got a couple of practice trials to ensure they were aware of how the experiment exactly worked. Every trial that was completed by the participant was evaluated by looking at the reaction time needed to complete it and whether the pressed key was congruent with colour of the word that was shown. The reaction time was measured in milliseconds.

After all 200 trials were finished, a questionnaire was conducted. This questionnaire concerned all the Associative trials that were shown during the experiment. The participants were questioned as to what they were thinking about when they encountered a specific word-picture pair. It was noted whether the association the participant thought of was in line with the aimed result. The Non Associative trials were not tested whether an association was found with the participant or not.

Programme

The main part of the experiment was done on the computer, where a specific program was located. The programme in question was made with PsychoPy, an Open Source software used for making experiments.

Participants

A total of 40 participants were recruited. Participants were recruited by either SONA-systems (a framework set up by the University of Twente to aid Psychology students in finding participants) or by the researchers through convenience sampling in their social circles..Ages ranged from 20 to 63, but most of the participants were around the age of 22. There were a total of 19 women and 21 men making it an almost equal amount of representation. Due to the convenience sampling, most of the participants were either Psychology students at the University of Twente or various family members of one of the two researchers. All of the participants were Dutch, German or English native speakers due to the requirement of it to make the associations easier to understand.

Data analysis

In contrast to most experiments within Psychology, this experiment does not use the *P*-value in order to accept or reject the null-hypothesis. Instead we estimate the difference in group means of reaction times (Schmettow, 2018) and derive level of certainty from Bayesian credibility intervals.

The within-subjects design that this study applied is put into a multilevel model. This means that every participant is tested on two separate conditions and the difference between the conditions gets estimated in this model. The conditions that are used to test the data with are based on the outcome of the aforementioned questionnaire. The overall response times (RT) that are found per trial per condition are then compared and estimates will be formed through utilizing an exgaussian distribution. An exgaussian distribution combines both a normally distribution and exponential distribution to estimate what future data will look like if tested on the same parameters. The estimates formed by this process can be viewed in a Comparison of group means (CGM). The coefficients and credibility intervals that stem from these CGM are used for the inference. A last analysis will be done to figure out whether a different effect could be happening with making mistakes instead of delayed responses. This will be done through a logistic regression. This will also be done through a CGM; the main result will give an indication of how large the odds are the response will be wrong in the default condition which is the non-associative stimuli. The difference with the second condition is visualized as a factor of the odds the default condition displayed.

Results

The experiment was finished with the intent of testing the Semantic Priming Stroop task. We tested the result both on the population level and the participant level, seeing if the test works on average and if not, whether it works for some people. If the priming effect exists then we would expect there to be a difference in reaction time between the Associative Stimuli and the non-associative Stimuli.

We looked for the difference between the two conditions of the response time for associative stimuli (A) and non-associative stimuli (N) without looking at the post-test questionnaire results. In Figure 1 the distribution of the raw data is shown. The distribution in the condition which indicates the A stimuli (true) seems to have a slightly wider distribution than the condition which indicates the N stimuli (false). However, most of the reaction times seem to gather around the same value for both conditions. This can be seen in table 1, where the exact difference can be found.

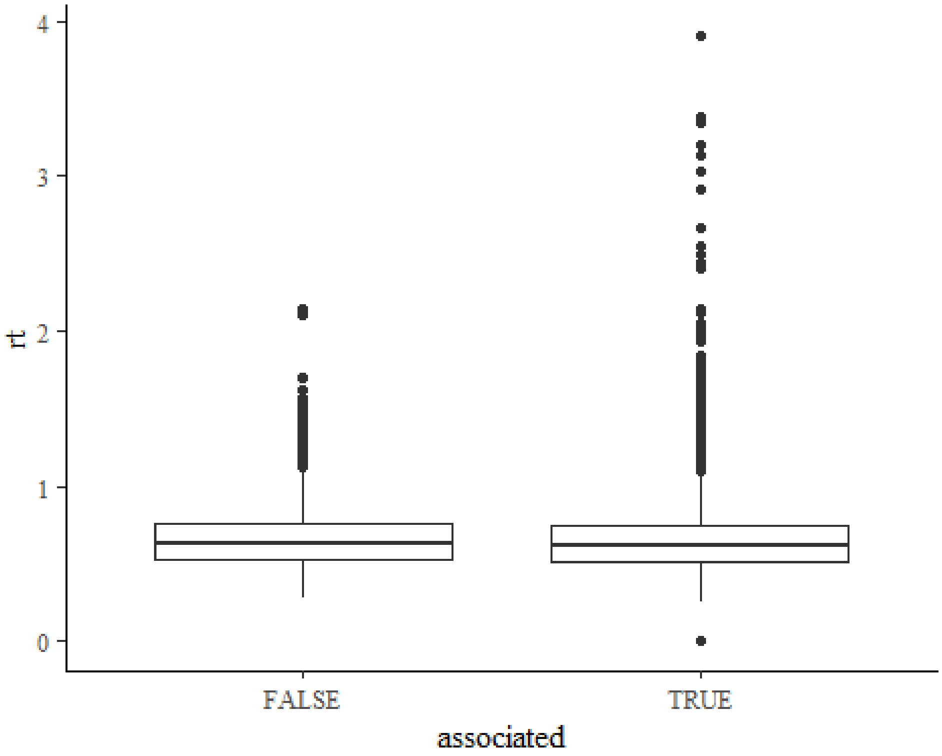


Figure 1: Data exploration boxplot

Table 1: Population level effects estimates without post-test classification

fixef	Median	2.5%	97.5%
Intercept	0.654	0.637	0.670
associatedTRUE	-0.00754	-0.0263	0.0124

From table 1 it can be seen that the A stimuli response times are estimated to be 0.007 seconds faster than the N stimuli. This value however falls within a credibility interval between 0.02 seconds faster and 0.01 seconds slower indicating that the value is in between these values with a 95% certainty.

The population level effects can be found in table 2. In table 2 the intercept states the default condition, which is the population estimates response time for the N stimuli. The center value states that the estimated response time in the default condition is around 683ms, with the credibility interval stating that if it that value is not correct, it should be between 655ms and 710ms. Furthermore, the difference with the A stimuli response times, is estimated to be around -3ms, with the credibility interval falling between -14ms and 9ms.

Table 2: Population level effects estimates with post-test classification

fixef	Median	2.5%	97.5%
Intercept	0.684	0.656	0.710
associated	-0.00314	-0.0149	0.00947

The participant level was analysed as well in order to check whether an effect can be found per individual instead of overall. Figure 2 and 3 show the difference in response times between the two conditions. Figure 2 shows the spread of the estimate center differences without the upper and lower bounds of the credibility interval. Figure 3 also incorporates the upper and lower bounds of the credibility interval per participants.

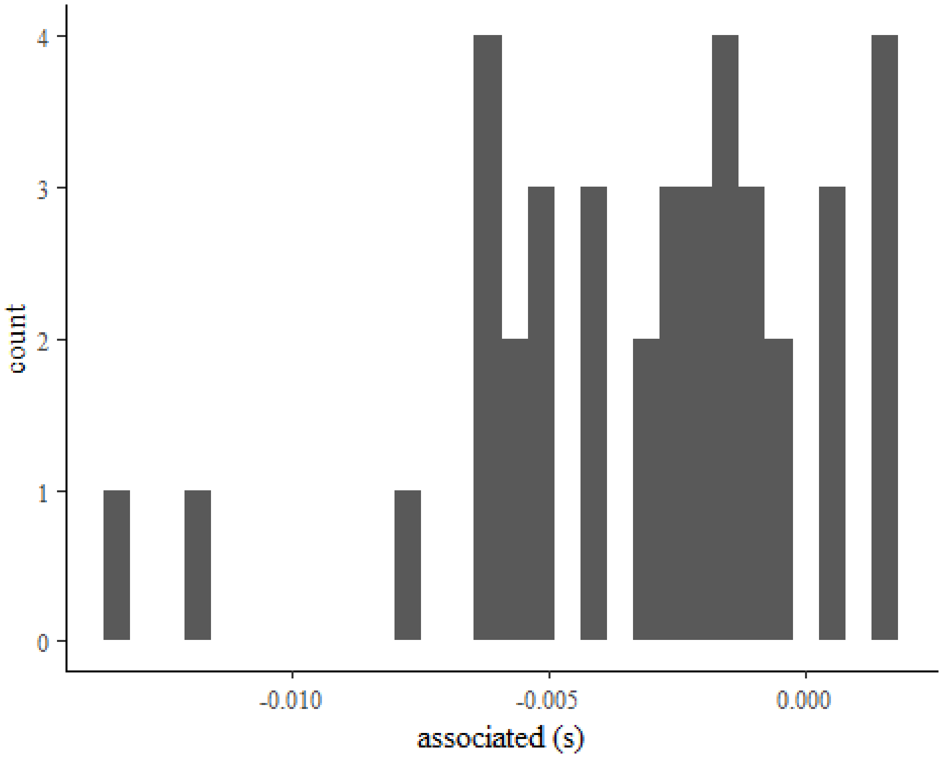


Figure 2: Barchart estimated difference in response time per participant

Looking at figure 2, most of the estimated response time differences per participant are around 0 to 5ms. 7 participants had an estimate of a difference higher than 0ms slower than the default condition; all the other participants were faster according to the results.

In figure 3 it shows mainly the same effect for most participants, with the center estimate to be roughly around the same value, most of them being under the 0ms difference.

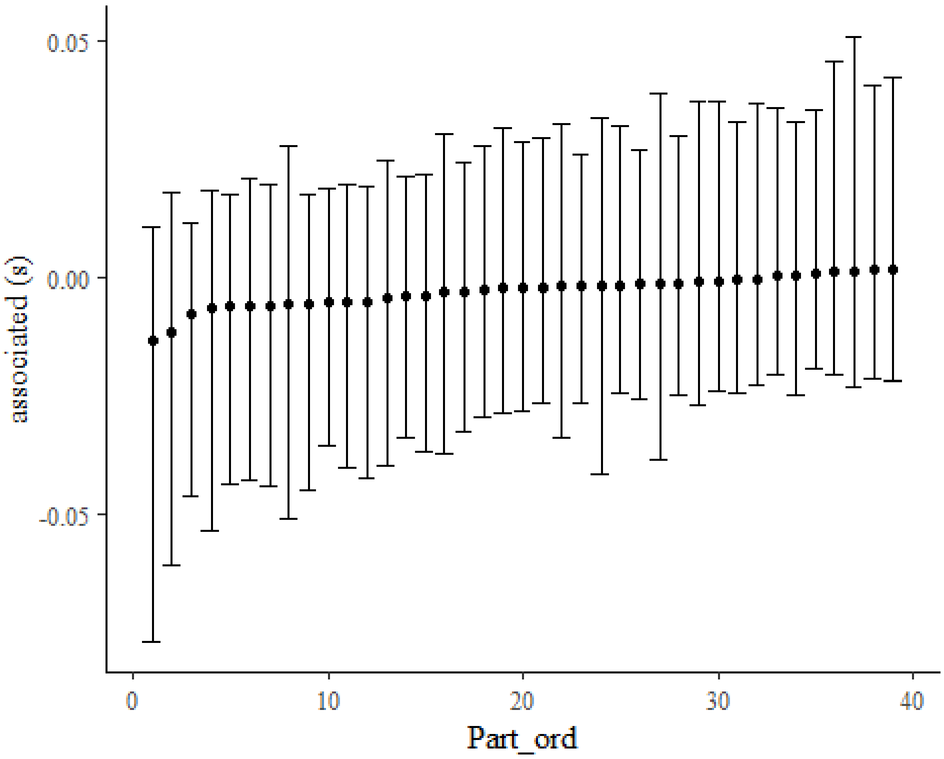


Figure 3: Estimated difference in response time and credibility interval per participant

Another analysis was done to figure out whether or not the A stimuli did in fact cause more incorrect responses instead of making the response time higher. Logistic regression was done for this and is shown in Table 3. It discards the usage of the response times and only focuses on the rate of which participants performed correctly for both conditions and estimated the average difference between those conditions.

Table 3: Correct responses analysis

fixef	Median	2.5%	97.5%
Intercept	3.375	2.961	3.904
associatedTRUE	-0.00168	-0.524	0.536

While the default condition of non-associative stimuli had an estimated center of 3.375 which means the odds the participant had a wrong response were 1 to 3.375, the associative stimuli had the odds decreased with a factor of 0.001684. The odds the participant got the associative stimuli wrong are then 1 on 3.369. The confidence intervals in the logistic regression are wider than the found intervals at the other estimate scores.

Discussion

The experiment of Sparrow and colleagues (2011) was originally performed with the Semantic Priming Stroop Task. This study achieved a high reputation due to the impressive results they reported and the acceptance in Science Magazine. Sparrow and colleagues reported a change in cognitive processes due to the internet, which was considered a big discovery. This is just one of the examples of the usefulness that the Semantic Priming Stroop task could offer. Reading someone's mind by using a simple implicit methods test is a step forward in the field of Psychology and makes way for a large amount of research that has yet to be done. However, this task has not seen a lot of replication studies. The current study therefore aimed to replicate this Semantic Priming Stroop task in order to seek the validity of the original research.

Comparing the studies

This replication has been conducted in a best case scenario in order to see if the Semantic Priming Stroop task works at all. However, when looking at the estimates in our results, it shows very little to no effect between the two conditions that we set up. With an estimated average difference between conditions of around 3ms that is not even going into the right direction, it is very hard to draw the conclusion that there is an effect like it was previously claimed (Sparrow et al., 2011). While the original idea was that you could read minds by seeing that someone has an association based on the difference in reaction time, there seems to be nothing conclusive of the sort in our replication. If this is the actual effect that the Semantic Priming Stroop Task offers, then you would need countless trials to make sure someone actually thinks about something specific and thus read their mind. This would make it an experiment that takes very long to complete and thus unrealistic to complete for every participant. For the purpose of this task, 3ms of difference is too small.

It could be argued that the reason that no effect is found in our replication study is that our study is not an exact replica of the original experiment. We took a best-case scenario to test the paradigm, but it is possible that it simply does not work according to our parameters, but instead only under specific circumstances. In the Sparrow paper, these circumstances could have been the yes/no trivia questions that were either hard or easy. In their experiment the questions were connected to a couple of tasks that were performed prior to the actual experiment. In contrast, the associations in the current study that were used to prime the participants with were presented only milliseconds prior to the Stroop task itself. However, the associations in this experiment were always only shown partly; first the picture, then the word that did not mean anything when taken separately, but when together conveyed the association. It is possible though, with the parts of the stimuli being shown apart from each other, that the association did not actually internalize with the participants.

It is clear that at least some of the associative stimuli that were used in the current study internalized with participants. Some participants mentioned something about a number of specific associations. An example is the ‘Airplane & Towers’ which implied ‘9/11’ and ‘Plantation & Black people’ which implied ‘Slavery’. They mentioned for example how it was “entertaining” or that it went a bit “too far”. This could explain some of the outliers in the exploration dataset as shown in figure 1 as they could have had an increased connection to these associations compared to the other ones. Nevertheless, the impact that some of the trials could have had were filtered out due to the repeated measures design and the post-test questionnaire that verified associations and non-associations.

It could be possible that even our conceived ‘best case scenario’ was not in fact a ‘best case scenario’ as not all stimuli had the same effect. It could therefore be argued that the stimuli had to be refined further to achieve the same effect as the aforementioned stimuli. However this would mean that this version of the Stroop task only works in very niche situations (if it works at all), such as the trivia questions from Sparrow and colleagues (2011) and is not widely applicable. This makes this ‘new’ implicit method not as groundbreaking as initially thought. If this is indeed the case, this paradigm of the Stroop task only works under very specific circumstances and would thus not very useful as an implicit method that is used more often.

Sparrow’s Results

The Sparrow study (2011) however did find significant results with the Semantic Priming Stroop task supporting their hypotheses. The question is how they arrived there. There is a multitude of things that could have happened during the experiment of Sparrow and colleagues (2011). Unspecified confounding variables like different sources of mental load could be a reason. One example of this is that the participants had to remember certain pieces of information that would be relevant later. This could be enough of a distraction to increase the delay the participants responded with in the hard trials. Furthermore when looking at the trivia questions that Sparrow and colleagues (2011) used, it is possible that participants had completely different thoughts when confronted with the questions than was mentioned in the conclusions that were drawn. The easy questions were common knowledge like “Does a triangle have 3 sides?” and it was proven that most of these easy questions were answered correctly. However, for the hard questions, the participants typically guessed the answer (Sparrow et al., 2011), which may imply that they had very different thoughts about the questions. Possible thoughts could be about not knowing the answer or trying to guess the answer themselves, or possibly trying to figure out what is meant with the question. But as is tradition with NHST experiments, this does not matter as it does not contribute to whether the hypothesis is correct or not, it only cares about whether there is an actual significant difference in results, not why it happened.

Even if the Semantic Priming Stroop task was accurate in testing for the interference in the case of the Sparrow study due to specific conditions and/or variables that they applied, the study still makes use of the yes-or-no mentality that the *p*-value brings with it as can be seen in their conclusions. After reporting a significant *p*-value that was under the 0.05 threshold, Sparrow and colleagues (2011) report that the participants thought of looking the question up with a search engine. The delay that was found significant could mean a lot of things, but the conclusion was drawn in this way because this was what the confirmed hypothesis said.

Another thing to note is that throughout the published summary of the paper, the sample size is never mentioned which raises questions if it is even up to par. After checking the complete methods on the website of Science magazine, it can be found that the Modified Stroop task experiment had a sample size of 46 people and that both of the two conditions consisted out of 16 trials. This amount is not large for a test using an implicit method they themselves came up with. Comparing this to the current study, where the only effect we found was of around 0.007 seconds of difference in a similar sample size but with 200 trials instead of 16, it could be said that the current study has a more reliable dataset.

It is very possible that the research of Sparrow and colleagues (2011) was weakened by various widely accepted strategies to increase the statistical significance that researchers tend to strive for to acquire the publication (Simmons et al., 2011; Giner-Sorolla, 2012 & Gelman, 2018). While they did not make use of the practice of making a difference between genders (Sparrow et al., 2011), the other practices are not as easy to find as these are practices that are not mandatory to state in the final report (Simmons et al., 2011). However there is no evidence to support the statement that this was the case in the Sparrow study (2011) and this is speculation at best.

Publication Magazines

Because of the reported results in the Sparrow paper, it has seen a lot of citation due to its publication in the Science Magazine, but should it have been accepted in the first place? Glancing at the paper raises questions, primarily in the first experiment where they use an implicit method that has not been tested before. Now that it is confirmed that the Semantic

Priming Stroop task is a niche method at best, it is debatable whether the paper should have been accepted at all. It is also interesting that the sample size and amount of trials is only mentioned when looking deeper into the study, effectively hiding the small amounts of participants and trials.

Looking at Psychology magazines in general, there seems to be a trend to mostly accept "sexy" results in order to make reading the papers more appealing (Asendorpf et al., 2013). A natural response would be to misuse the degrees of freedom more often in order to achieve a publication. The null hypothesis significance testing (NHST), is bad for Psychology in general because of the many legal ways to get false-positive results (Gelman, 2018; Asendorpf et al., 2013; Simmons et al., 2011). Having researchers switch over to confidence limit testing instead of the *p*-values with the NHST solves this problem by making research focus more on gradations of results instead of answering yes or no to good sounding results.

Reflection

A big part of the replication crisis is seemingly the failure of the *p*-value in particular and the tradition of Null hypothesis significance testing. In a previous paper by D. Szucs & J.P.A. Ioannidis (2017) they talked about an all-or-nothing mentality where researchers try everything to achieve the 'significance level' required to get the yes or no answer they were looking for. The amount of false positives that are being uncovered in this replication shows that there are too many ways to mess up the *P-value* that compromises the validity of research in general. This added to the small amount of Psychology publication outlets and relatively large amounts of papers being produced, all researchers try to get the most groundbreaking results, even if it means editing their research in ways they did not realize was bad for the science of Psychology in general, as it is not forbidden to use these ways to increase the significance level (Simmons et al., 2011).

Assuming that our experiment was a success in our aims to check whether this specific Stroop task paradigm works or not, it is important to find out why this replication has not been done before. The replication crisis has reclaimed many generally accepted theories as mentioned by Giner-Sorolla (2012).

Conclusion

In short, looking at how the Semantic Priming Stroop task has been used in the past with Sparrow (2011) and Schmettow and colleagues (2013) makes it sound like it is a new implicit method that works. However, after thoroughly having tested this specific paradigm of the Stroop task, it has become clear that it is unlikely to work in most cases. The very basic version of this task, performed with a scenario that had optimal conditions in both the stimuli and the Design, shows practically no effect. This casts some doubt on the previous studies that made use of this Stroop task. Even when the results from these previous studies show that it worked in their case, it is evident that the results of this task at its core do not replicate under other circumstances. Thus, it is safe to say that the Semantic Priming Stroop task does not work in general.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on psychological science*, *9*(1), 40-48.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PloS one*, *7*(1), e29081.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16-23.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562-571.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. *The Sage handbook of methods in social psychology*, 195-212.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology*.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Journal of personality and social psychology, 46(3), 598.

Schmettow, M., Noordzij, M. L., & Mundt, M. (2013, April). An implicit test of UX: individuals differ in what they associate with computers. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2039-2048). ACM.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *science, 333*(6043), 776-778.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, *11*, 390.

Appendix A: R Output

DAP Mindreading 18

M Schmettow

15/05/2019

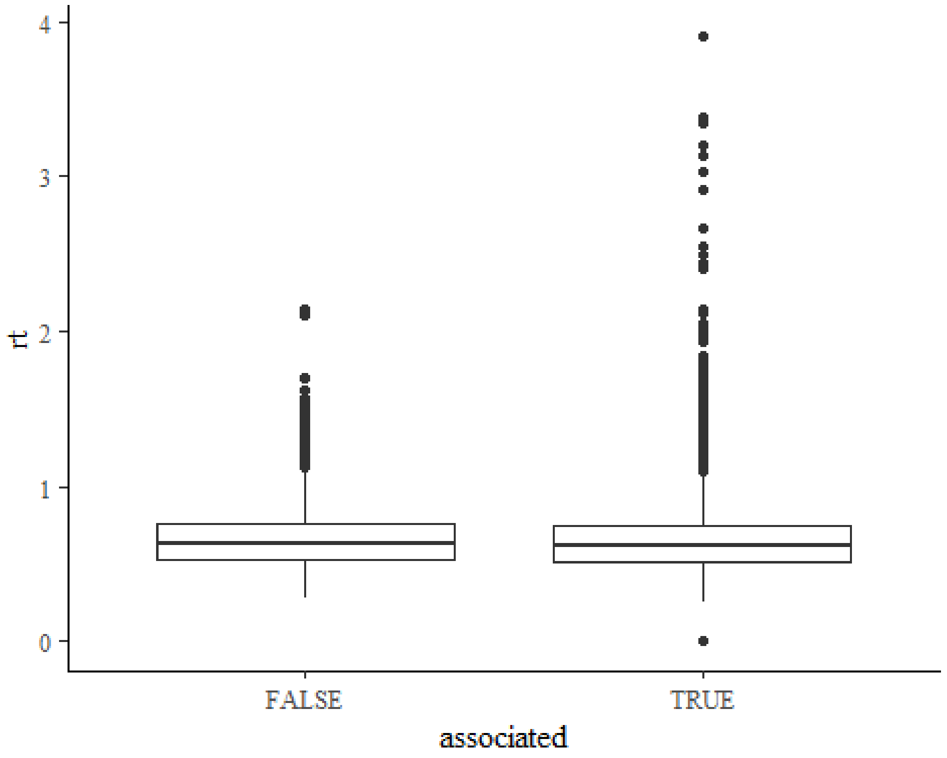
Reading Data

```
raw_files <- dir(path = "rawdata", pattern = "csv$", full.names = T)

read_raw <- function(f){
  read_csv(f, ) %>%
  slice(-1:-17) %>%
  select(Part = participant, Stim, Word = text_EN,
         Block, correct = resp.corr, rt = resp.rt) %>%
  mutate(Part = str_extract(Part, "[[:digit:]]+$"),
         correct = as.logical(correct),
         trial = row_number())
}
#read_raw(f)



Stim <- readxl::read_excel("Stimuli.xlsx")

Quest <-
  readxl::read_excel("Quest.xlsx") %>%
  select(-demo) %>%
  gather(Part, associated, -Stim) %>%
  mutate(Part = str_remove(Part, "part.")) %>%
  right_join(Stim, by = "Stim") %>%
  mutate(associated = as.logical(if_else(Cond == "N", 0, associated)))

# summary(Quest)
#
```

```
# Quest %>%
#   filter(is.na(associated))



MR18_all <-
  map_df(raw_files, read_raw) %>%
  left_join(Quest, by = c("Part","Stim")) %>%
  filter(Cond != "T") %>%
  select(Part, Block, trial, Cond, Stim, Word, associated, Image = primeImage, correct, rt)
%>%
  as_tbl_obs()

MR18 <- filter(MR18, correct)

save(MR18, MR18_all, file = "MR18.Rda")
write_csv(MR18, "MR18.csv")

load("MR18.Rda")

MR18

MR18 %>%
  group_by(Part) %>%
  summarize(n())

## # A tibble: 39 x 2
##    Part  `n()`
##    <chr> <int>
##  1 1        99
##  2 10       92
##  3 11       94
##  4 12       90
##  5 13       95
##  6 14       88
##  7 15       97
##  8 16       89
##  9 17       95
## 10 18       90
## # ... with 29 more rows
```

Data exploration

```
MR18 %>%
  ggplot(aes(x = associated, y = rt)) +
  geom_boxplot()
```



```
load("M_1.Rda")
fixef(M_1)
```

fixef	center	lower	upper
Intercept	0.6538788	0.6369763	0.6703012
associatedTRUE	-0.0075351	-0.0262636	0.0123886

Multi-level model (RT)

```
M_3 <- brm(rt ~ associated + (associated|Part) + (1|Stim),
           family = "exgaussian",
           data = MR18)
```

```
P_3 <-
  posterior(M_3) %>%
  mutate(fixef = str_remove(fixef, "TRUE"))


save(M_3, P_3, file = "M_3.Rda")

load("M_3.Rda")
```

Population-level effects

```
fixef(P_3)
```

```
fixef            center       lower       upper
Intercept     0.6836951   0.6556274   0.7100678
associated   -0.0031483  -0.0148912   0.0094687
```

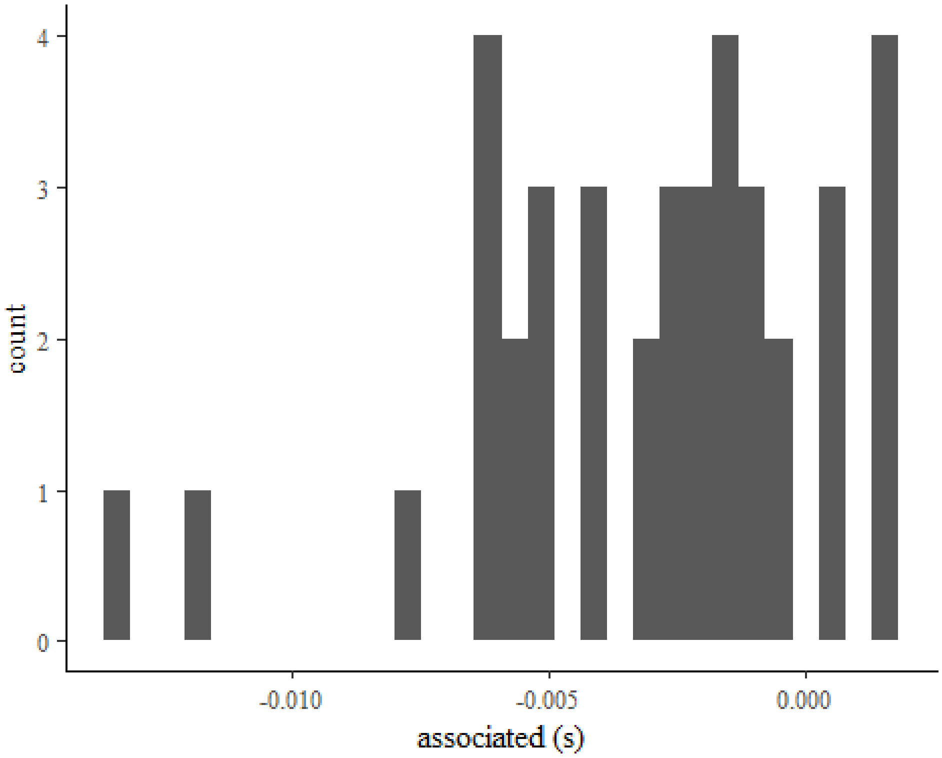
```
grpef(P_3)
```

```
fixef        re_factor      center       lower       upper
Intercept    Part        0.0800998   0.0621473   0.1057306
associated   Part        0.0111817   0.0005186   0.0343877
Intercept    Stim        0.0109727   0.0023315   0.0184289
```

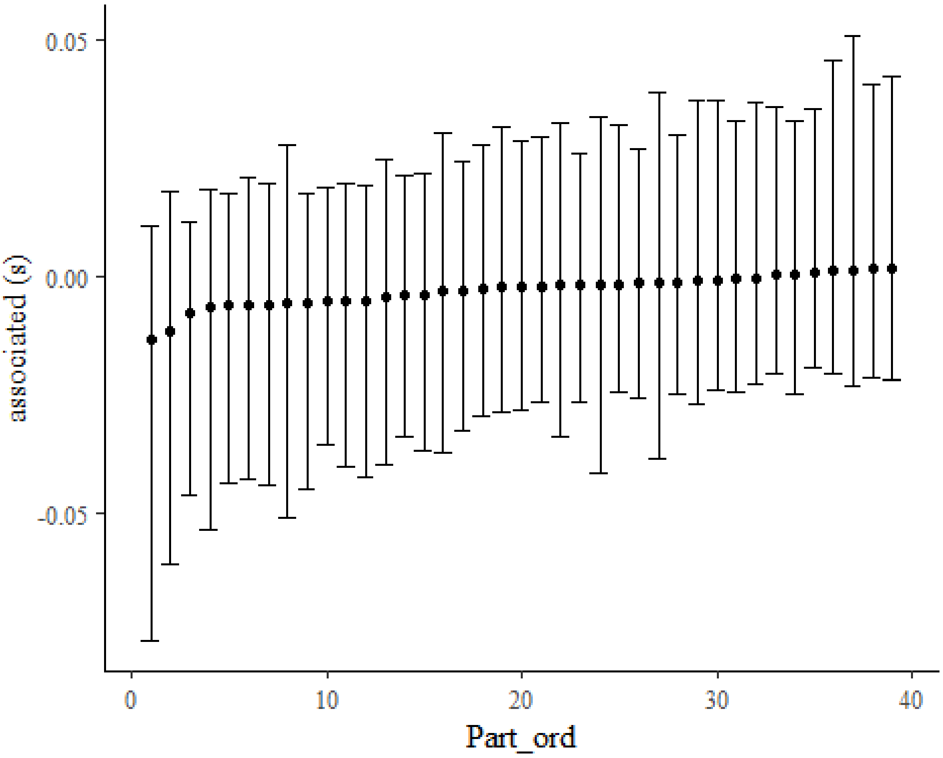
```
ranef(P_2) %>%
  select(re_entity, fixef, center) %>%
  spread(key = fixef, value = center) %>%
  ggplot(aes(x = Intercept, y = associated)) +
  geom_point()
```

Participant-level effects

```
P_scores <- P_3 %>%
  filter(fixef == "associated") %>%
  bayr::re_scores("Part")



T_scores <-
  P_scores %>%
  rename(Part = re_entity) %>%
  group_by(Part, fixef) %>%
  summarize(center = median(value),
            lower = quantile(value, .025),
            upper = quantile(value, .975)) %>%
  ungroup() %>%
  mutate(Part_ord = min_rank(center))

T_scores %>%
  ggplot(aes(x = center)) +
  geom_histogram() +
  xlab("associated (s)")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
T_scores %>%
  ggplot(aes(x = Part_ord,
             y = center, ymin = lower, ymax = upper)) +
  geom_point() +
  geom_errorbar() +
  ylab("associated (s)")
```



Logistic regression (correct)

Perhaps, associations effect not the reaction time, but causes more incorrect responses. We run the same model as above with correctness as outcome variable, using logistic regression.

```
M_4 <- MR18_all %>%
  mutate(correct = as.numeric(correct)) %>%
  brms::brm(correct ~ associated + (associated|Part) + (1|Stim),
          family = "bernoulli",
          data = .)


P_4 <- posterior(M_4)
PP_4 <- post_pred(M_4)
save(M_4, P_4, PP_4, file = "M_4.Rda")
```

Again, results are NULL.

```
load("M_4.Rda")
```

```r
fixef(P_4)
```

fixef	center	lower	upper
Intercept	3.3750809	2.9605760	3.9040504
associatedTRUE	-0.0016841	-0.5241573	0.5359343