A Modified Replication of the Semantic Priming Stroop Task: is efficient mind reading possible?

Kira C. Choinski

Bachelor Thesis

Supervisor: Dr. Martin Schmettow

Second Supervisor: Simone Borsci

Faculty of Behavioral, Management and Social Sciences (BMS)

University of Twente.

Abstract

In the present study, the Semantic Priming Stroop Task (SPST; a variation of the original Stroop task), was put to a test because it could potentially pose a useful implicit measurement tool for research as well as diagnostic purposes. It found application only two times before and therefore needs verification. Two conditions were created to show whether the SPST works under optimal conditions or not. In a repeated-measures within-subject design, participants were therefore exposed to a picture, followed by a word. The picture-word pairs were designed to be either semantically unrelated or related, provided the participant possessed the demanded knowledge and thus made the demanded association. Participants were requested to respond to the ink color of the word as quick as possible, like in the original Stroop task. It was assumed that longer response times indicate that the demanded association was made. The validity of the conditions was tested by means of a questionnaire and the conditions were reclassified accordingly for the statistical analysis. Deviating from the prior studies, no effect was found. It was concluded that the SPST does not work in general. However, one of the prior studies observed the effect with a slightly different but more limited study design. Thus, further research could show that the SPST works under different, more limited conditions, as it presumably did before.

Keywords: Stroop task, semantic priming, implicit methods, response times

1. Motivation Behind the Study

As the study of the mind, Psychology tries to understand human thought more than any other science, often in form of attitudes or opinion. To measure these, psychologists have the possibility to utilize explicit measures (i.e. self-reports), which aim to determine thoughts by asking questions. Thus, research questions of explicit measures are often transparent to the assessed individuals. This way, responses go through a conscious validation process in which motivations like social desirability considerations potentially bias outcomes of self-reports.

To avoid this problem, scientists can employ implicit methods which provide results not by asking questions directly, but by measuring the cognitive processes in question through tasks. Like this, implicit methods generally do not expose their research question to participants and thus circumvent the conscious assessment process by participants, thereby avoiding bias.

This advantage is also assumed to be present in the paradigm called Semantic Priming Stroop Task (SPST). It is a variation of the original Stroop task and involves apart from color-naming a semantic priming aspect. Before the color-naming task, participants are exposed to a priming stimulus. In the modified Stroop task that is implemented afterwards, the color-words are replaced by target words and unrelated words. The target words are semantically related with the prime, provided the participant knows the aspect in question. Just as in the original Stroop task, participants are then asked to name the ink color of the words as quick as possible. A longer response time in the condition with prime and target word in comparison to unrelated words is then assumed to indicate that the particular participant has the association/ knows about the aspect in question. Essentially, one could say that with the SPST researchers can ‘read minds’ because on the basis of the response times it could be said whether the specific participant had the association in question or not (i.e. if the participant has the specific knowledge or not), without asking them directly.

The SPST is a rather novel paradigm, only applied two times to this point, introduced by Sparrow Liu and Wegner (2011). Since the publication of the Open Science study by AArts et al. in 2015, paradigms with a low replication rate should be treated with even more caution. With the large-scale replication project, it was unfortunately found out that many previous studies must have made mistakes, considering that the majority of the phenomena did not replicate. Among these problematic cases included also many priming paradigms. As these results further fueled the replication crisis in social sciences, there is a heated debate going on about the factors that possibly contributed to its origin, best summarized as ‘researcher’s degrees of freedom’ or ‘the garden of forking paths’. Also, the study by Sparrow

et al. exhibit some weaknesses or at least opacities that leave room for speculation about its strength

With respect to the threatening replication crisis of also many priming paradigms and the low number of applications of the SPST, a replication of the proposed paradigm seems necessary. Therefore, the present study aims to investigate the paradigm of the SPST by putting it to a test, employing two optimal conditions that should show a delay in response time in an association condition compared to a non-association condition, provided the SPST works. The following section will give further insight into the issue as well as the goals and approach of the present study.

1.1 Implicit Methods

A possible solution to the above explained problems with explicit methods are so-called implicit methods that aim to assess thoughts by circumventing validation processes and hence bias. To achieve this, the research question in implicit methods is nontransparent but assessed through tasks that stimulate and measure the activation of specific mental constructs (associations, attitudes, beliefs etc.). There is still some disagreement between scientists about whether implicit methods measure unconscious mental constructs but the fact that implicit methods avoid a validation process that is present in explicit measures is largely uncontroversial (Gawronski, LeBel, & Peters, 2007). Nevertheless, this advantage is the reason why implicit methods are currently used, to achieve unbiased results.

1.1.1 Semantic Priming Stroop Task. One such implicit method, as introduced by Sparrow et al. (2011), is the Semantic Priming Stroop Task (SPST). In this task, participants first get provided with some sort of prime. In the thereupon following modified Stroop task, the color-words of the original Stroop are replaced by a proportion of target words and a proportion of unrelated words. The target words are semantically related with the prime, for participants that have the demanded association/ know about the construct in question. As in the classical version of the Stroop task, the subjects are then asked to name the ink color of the words as quick as possible. It is assumed that when a word strongly captures the attention of the participant (presumably due to exposure to the preceding prime), the response time is longer. A delay in response time is thus understood as indicating a strong association between the prime and the word. With the use of two classes of words (one class= target words; other class= unrelated words), the researcher could this way infer whether someone possesses a

certain association/ has a certain knowledge or not. In the specific case of the paper by Sparrow et al., the prime consisted of trivia-questions.

Under this assumption, Sparrow et al. conclude that longer response times of the participants in the study represent stronger associations between trivia-questions and words in the subject. This way, Sparrow et. al employed the SPST to test the hypothesis that people tend to think about technology when they need information. This assumption is accounted for by the idea that our society largely uses the internet as a sort of extended memory, i.e. looks up information in the internet rather than having all the information available in their minds.

Another application of the SPST was published by Schmettow, Noordzij and Mundt (2013). Here, the task was implemented utilizing picture-word pairs as associations. In order to understand what individuals associate with a picture of a computing device, words of three different categories were presented to the participants together with the pictures. The results were as expected, i.e. longer response times were observed for participants that were predicted to have a certain association when viewing a computing device.

So, it can be summarized that the SPST has been applied only two times at this point.

1.1.2 Theoretical background. To understand the theoretical background behind the SPST, its components will be shortly explained here.

1.1.2.1 Classical Stroop task. The SPST is a variation of the well-known Stroop task. In that task, subjects are asked to name the ink color of color-words, not the words themselves. Thereby, the word and its color are either compatible (‘red’ in red ink) or incompatible (‘red’ in blue ink). In the incompatible condition, people usually experience a disruption/ interference in naming the color instead of reading the word, presumably because the latter process is inevitably automatized (Monahan, 2001). This phenomenon is called the Stroop effect. The effect can be observed in a delay of response time. This way, the Stroop task found application in experimental research as a measure of attention for example (MacLeod, 1992), as well as in diagnostics (Sergeant, Geurts & Oosterlaan, 2002). The paradigm of the original Stroop task, as just described, has been replicated hundreds of times, as can be seen in a literature review of MacLeod for instance (1991).

1.1.2.2 Semantic priming. Priming describes the influence of the provision of a cue on the subsequent processing of another stimulus, in a task that entails some sort of cognitive performance. This influential effect consists either in facilitation or inhibition of a response (Tulving, Schacter & Stark ,1982). Semantic priming in turn is a specific form of priming (i.e.

associative priming) that depends on semantic relation, as in for example: prime= bread, target= butter. With such a prime, reduced response times are usually observed for primes that facilitate a response to a stimulus because it is semantically related to it. In contrast to that, with primes that inhibit a response to a stimulus because it is semantically unrelated with it, increased response times are commonly observed (Heyman, van Rensbergen & Storms, 2015). Semantic priming (facilitative as well as inhibitory) is what is also suggested to occur in the SPST.

1.2 Replication Crisis of Social Sciences

In 2011, a series of unpleasant events, including instances of research fraud, started to lay doubt upon psychological research results (Pashler & Wagenmakers, 2012). In a following extensive joint work of the Open Science Collaboration from 2015, a team of 270 international researchers reconducted 100 psychological studies that had been published in journals in 2008, of which only 39% could be replicated (Aarts et al.).

1.2.1 Priming paradigms that could not be replicated. Taking a closer look at the Open Science study, it attracts attention that also many priming studies did not replicate. For instance, the study by Schnall, Benton & Harvey (2008) which presumably found evidence for the hypothesis that priming people with cleanliness would decrease the severity of their moral judgements, could not be replicated by the researchers (Johnson, Cheung, Donnellan, 2014). Also, for the prior found positive effect of spatial distance primes on feelings of affect and emotional closeness by Williams and Bargh (2008), the Open Science researchers could not find this effect in a replication of the study (Lynott et al., 2014). The hypothesis here was that people primed with close distances would have stronger emotional closeness to their kin and hometown than when primed with long distances.

Moreover, a study that addressed semantic interference effects, similar to the one that is suggested to be at work in the SPST, also failed to replicate (Galak, 2015). This study's hypothesis was that the response time to a naming-task would be slower for semantically related picture-word pairs than for semantically unrelated ones (Janssen, Schirm, Mahon, & Caramazza, 2008).

Since the Open Science study is highly alarming information for at least most social scientists, it broke loose the replication crisis and with it hot debates about possible triggers and contributing factors that enabled the crisis in the first place.

1.2.2 Contributing factors of the crisis. The first question arising in regards to the crisis is why it happened or what factors enabled it. Several aspects were identified as possible contributors of the crisis that will be addressed in turn here.

To start with, the value and meaning of replication appears to be commonly misunderstood. In practice unfortunately researchers gain greater economic advantages from providing new results than replicating 'already known effects', as journals will rather report on new insights than presenting an 'already known phenomenon' (Aarts et al., 2015; Bakker, van Dijk & Wichters, 2012). This is a bad state of affairs because the probability that results are true-positive depends among many other aspects on the number of replications. As a consequence, replications are of great value but are not rewarded as such.

Perhaps partly due to this pressure on practitioners to find hitherto unknown effects to get published, the indications amass that a big part of psychological research does not provide new insights but possibly often produce false-positive results (Ioannidis, 2019). Reasons and means by which this gets possible exist in the often-called researcher's degrees of freedom.

1.2.2.1 Researcher's degrees of freedom. When scientists come to the point to analyze their collected data, they have a certain not insignificant amount of freedom in how to approach and conduct it. They choose which data to define as 'outlier' and exclude, they decide whether more data shall be collected, they decide which control variables to consider, which conditions are combined and which compared, and perhaps most strikingly, they decide on the analytical method (Simmons, Nelson, & Simonsohn, 2011). This freedom of choice in cleaning and analyzing data is what Gelman and Loken refer to as 'The Garden of Forking Paths' (2013). They state that researchers have a nearly infinite number of 'paths' they can follow when they did not specify a concrete data analysis plan beforehand and instead fit the analytical method to the observed data. Consequently, they urge the importance of a prior established analysis plan. In contrast to that, it unfortunately seems to be common practice for scientists to probe many analytical alternatives and report only on those that yielded statistical significance (Simmons et al., 2011).

The issue becomes clearer when taking a closer look on p-values as they are typically consulted as indicators for the confidence of data. Relying on p-values means that results are considered significant when statistical models provide a 95% chance that the outcome occurred not at random (i.e. the 5% hurdle is generally recognized as sufficient). But regarding the shocking outcome of the replication study by Aarts et al. (2015), this approach seems to be inadequate. In plain terms, p-values merely provide evidence against the null

hypothesis but allow researchers to state that this means strong evidence for the alternative hypothesis.

Overall, with this information it becomes clear that a practitioner can easily arrive at the required 5% with data that is just noise, as long as they 'fish' their data extensively enough using different statistical tests until they come across some pattern. The claim that such questionable research practice takes place is also supported by a survey conducted in 2016. Nearly 50% of the asked members of the German Psychological Association stated to have employed HARKing (=hypothesizing after the results are known) at least once in their career (Fiedler & Schwarz, 2016).

However, this does not necessarily mean to accuse all researchers for 'fishing' and ascribe bad intentions to them, but to emphasize the importance of understanding the ambiguity that the garden of forking paths inevitably results in if researchers do not understand their methods well enough. Moreover, this emphasizes again the urgency for a greater appreciation of the value of replications, since replications would quickly reveal such mistakes.

1.3 Doubts Concerning the SPST

As already explained above, many psychological studies cannot be replicated, presumably due to several mistakes in data analysis made by researchers in the past. Further, the replication crisis lays doubt on many priming paradigms, as well as semantic interference effects as they are suggested to be involved in the Stroop task for example. This leads back to the SPST, which found application only two times at this point.

The study by Sparrow et al. (2011) raises the question, whether a stronger association actually results in a delayed response in the color naming task. Since the classical Stroop effect seems to be widely replicable, such a hypothesis appears intriguing. Nevertheless, the study design as implemented by Sparrow et al. was rather weak. To facilitate understanding, the employed procedure is explained here in short: Each participant was made to answer an easy question block and a hard question block, each of 16 items. After each block, the modified Stroop task with 24 words was provided (8 target words and 16 unrelated words), thereby having to hold a 6-digit number in mind.

Coming back to the weaknesses of the study design, first of all, the small amount of only 24 repetitions makes the design appear less strong. Compared to other studies that measured response time in relation to the Stroop task, where numbers like 480 (Mewhort, Braun & Heathcote, 1992), 96 (Pothos & Tapper, 2010) or 288 (Szűcs, Soltész, Bryce &

Whitebread, 2009) trials are employed, 24 seems a very small number of observations. Also, there were no practice trials, at least not mentioned. Further, no validity test was conducted that could assure that the questions were truly perceived as strong or easy respectively for the individuals. Moreover, the study introduced a memory task but did not test for recall later. Additionally, the response times were analyzed assuming a normal distribution, despite the fact that response times are not normally distributed. Another point that draws attention is the non-transparency of the stimuli development, as it is not explained how they were created or selected. But most strikingly is the fact that employing this method, that has never been used before, the researchers assumed to find the appealing association between searching information and thinking about computers.

Taken together, as useful as the SPST could be, to this point, there is only very little evidence of its functionality. Valid replications need to be done in order to proof whether such an effect exists or not.

1.4 The Aim of this Study

In light of the current serious replication crisis and considering the weak evidence for the potential effect, the present study aims to replicate the SPST effect in an 'optimal condition approach' to counteract the crisis in this particular case. In specific, *the present study investigates whether the suggested effect of the Semantic Priming Stroop Task works under optimal conditions, observable by means of a delayed response time in a semantic priming condition, as compared to a control condition.*

The 'optimal conditions' consist of one condition that employs stimuli that presumably trigger associations in the participants in comparison to another condition where the stimuli will likely not trigger associations. Further, an independent check for validity will find out whether the participants had the associations or not, which ensures a correct classification of the stimuli into the conditions per participant for the data analysis (explained in detail in 'Methods').

Regarding the lack of transparency in the Sparrow et al. (2011) study, it is difficult to conduct an exact replication. Thus, the design and procedure of the present study are slightly modified in that pictures instead of questions are employed as semantic primes, no additional cognitive load is introduced and a subsequent validity check is added, which disqualifies it to be declared a literal replication. Consequently, the current study is a modified replication of the mentioned study, according to the definition of the American Psychological Association (n.d.).

Since the present study is not an exact replication of the Sparrow et al. study, the features are contrasted in Table 1 for clarification. The differences further include that the analysis is done utilizing parameter estimates and credibility limits (95%). The response times are measured by 200 trials to be able to filter for noise. Lastly, an appropriate statistical model for response times (i.e. ExGaussian) is employed for the analysis of the response time measures.

If the results show the effect as assumed by the study of Sparrow et al. (i.e. strong association= delay in response time), there might be more certainty about the effect of the SPST. If in this optimal condition, no such results can be observed, this should cast serious doubt on the functionality of this novel paradigm.

Table 1

Comparison between the study by Sparrow et al. (2011) and the present study.

Feature	Sparrow et al. (2011)	Present study
Stimuli	Questions (easy/hard)	Pictures (semantically related/ unrelated to word)
Trials	24	200
Addition	Add. cognitive load	-
Validity check	-	Post-Questionnaire
Analysis	Assuming n-distribution	Assuming ExGaussian distribution
	Employing p-values	Employing parameter estimates and Cl (95%)

2. Methods

To facilitate understanding on the following explanations regarding the two conditions of this study, this is a short description on how they need to be understood. In paragraph 2.1.2, the development of the conditions of *the intended design* are explained. These are the items the researchers assume to induce associations and non-associations respectively, referred to as pre-questionnaire-classification. In important contrast to that, in paragraph 2.1.4 the establishment of the *'true conditions'* through the post-questionnaire is pointed out. These conditions entail how participants truly perceived the items and are referred to as post-

questionnaire-classification. To ensure validity, the data analysis is based on the individual post-questionnaire-classification conditions.

2.1 Stimuli

As explained above, for the present study picture-word pairs were utilized as stimuli to observe the suggested effect of the SPST, i.e. a disruption indicated by a delay of response time. Therefore, two conditions were created. For the purpose of correctly understanding the conditions, first the meaning of association in this context is briefly explained.

2.1.1 Forms of association. An association can be of two kinds in this case. One is that the picture and the word are associated with each other, implying a simple connection between the words. For example, as with the picture 'Yogurt' and the word 'Spoon'. The association is that one usually eats a yoghurt with a spoon. Another kind of association is that the picture in combination with the word evokes a third construct. For instance, seeing a picture of a wolf and the word 'grandma' would evoke the construct of 'Red Riding Hood' for participants who know the fairytale.

Regarding that the latter form of association is more complex than the former, for this study the complex form of association was employed. Considering that the purpose is to observe an effect that is suggested to be indicated by response time, this form of association should be observed more easily than a less complex association (provided such an effect exists). To clarify: since 'wolf' and 'grandma' presumably evoke rather a complex narrative, namely the fairy tale 'Red Riding Hood', it should be more attention-grabbing than a simple association like 'eating a yoghurt with a spoon'.

2.1.2 Intended design (pre-questionnaire-classification). One intended condition consisted of association picture-word pairs (AC), meaning that the combination of the picture and the word would induce an association. The other intended condition involved non-association picture-word pairs (NC), namely the combination of the picture and the word would not induce an association. As described before, this study is aimed at the observation of the purposive effect of the SPST.

2.1.3 Stimuli development. To reduce the probability that an association is invalid and not induced for the participants to a minimum, the stimuli were carefully developed and tested as will be described in detail now.

The pairs were first developed without pictures, as word-word pairs, to provide a better overview of the pairs and their associations in a table. Therefore, the researchers first

came up with topics and the desired associations that are probably known by almost every person, including a great age-range and a great international range. The latter was achieved through the collaboration between the German and the Dutch researcher. To achieve a great age-range, thereby attending to various individual-differences, the design of the stimuli concentrated on widely known fairy tales, movies, events and persons.

Considering that response times usually provide very noisy data, the stimuli development was already aimed at creating enough pairs for both conditions to be later able to clean the data for noise.

First, both researchers developed 30 association pairs and 25 non-association pairs separately, then exchanged the stimuli and like this tested them on each other. This way, the check for cross-cultural validity, as well as a first check for individual differences was given. From this task, 50 association word pairs and 50 non-association word pairs emerged. Next, the pictures were added to create the complete stimuli as intended as picture-word pairs. Afterwards, the stimuli were given to the third researcher who also eliminated some invalid or weak items. A discussion within the context of all three researchers brought to light some rules for the stimuli, as described in the following criteria section. Thus, some associations or their picture-word pairs had to be substituted by some that adhere to the criteria. Lastly, the set of stimuli was tested on three external persons that were presented with the pictures and the words and requested to answer as quick as possible what strikes their minds at that point. This way, two non-working stimuli were removed and like this, the final set of 50 association stimuli and 50 non-association stimuli was obtained (Appendix 1).

Eventually, the stimuli were duplicated, providing the study with 100 AC and 100 NC stimuli (plus 15 NC for a practical trial prior to the actual study trial). This was done to acquire a greater number of stimuli to be able to reach a greater reliability. Additionally, this way the study simultaneously tests whether the effect still holds (if it holds) when stimuli are presented twice. In the end, the items were translated into German and Dutch by native speakers because the test presumably may not work when words are not read and understood automatically.

2.1.3.5 Criteria for the stimuli. The words that were chosen to evoke a certain association in combination with the picture, had to fulfill certain criteria to be accepted as stimuli. First, only single words (no compound words) were employed, as there are single words utilized in the original Stroop task, i.e. colors.

Second, the words could not be color-words because it could thoroughly disturb the data, since this would probably yield the original Stroop effect which is already known to exist. In this study, the original Stroop effect would bias the priming effect that is aimed to be observed, i.e. a disruption due to the evoked association. For this reason, the same no-color rule holds for the pictures. The pictures were still presented in color, but it was refrained from pictures where the color was explicitly decisive. For example, a green hat could not be employed to (in combination with the word 'Ireland') evoke the association 'Leprechaun' because the color of the head in particular induced the association. In contrast, a wolf can still be grey to (in combination with 'grandmother') evoke the association 'Red riding hood' because the color of the wolf is not crucial to the association. Also, the association was tried to be designed far from any color at best, to rule out any possible unwanted disturbances with the color-naming task.

A third important criterion for the stimuli was that both, the word and the picture, would not evoke the association on their own. Only in interaction they can induced the third construct. For instance, a picture of 'Santa Clause' already induces the association Christmas without the word 'winter'. For the same reason, the words were not allowed to include names, considering that those often already are too narrative by themselves and would thus pose the danger to induce an association on their own where it is not intended. In contrast to that, the associations deliberately needed to be narrative as in e.g. 'Harry Potter' or 'Unicorn'. Associations like 'School' were excluded due to the fact that they may not need notably more cognitive effort.

Additionally, the pictures were kept as clear as possible and not too rich in content, showing only the object in question. In sum, all these measures were taken to ensure a clear difference between NC and AC items, so it is nearly ruled out that a potential delay in response time in the AC was due to inconsistent choice of stimuli and thereupon resulting noisy data.

2.1.4 Establishing the 'true conditions' (post-questionnaire-classification). The intended design as explained above was checked by a post-questionnaire to find out how the stimuli were truly perceived by the participants. This way, the 'true conditions' for the statistical model could be established.

The validity test was administered through the use of a multiple-choice questionnaire that involved all 50 AC items. Therefore, the AC items were presented with four answer options, of which the fourth was always 'none of the above' to give the participant the chance

to indicate if the association was not known at all. The two wrong answer options were chosen in a way that if one does not know about the association, the other options could be a logical choice, too, as with e.g. ‘skyscraper’ and ‘airplane’ the option ‘holidays’ was provided. In the end, also the questionnaires were translated into German and Dutch by native speakers.

Most importantly, the thereupon resulting reclassified conditions are much stronger than those based on the pre-questionnaire-classification. This is the case because the questionnaire can ensure the validity of the two conditions individually per participant. That is why the post-classification conditions were used as the basis on which the data analysis would rest.

2.1.5 Apparatus. For the implementation of the experiment, the open source program ‘PsychoPy2’ was utilized.

2.2 Procedure

Participants could sign up via the SONA System’s test subject pool of the University of Twente or were recruited via the researchers’ wider social network.

1. Participants received a printed information sheet that roughly explained what activities they would perform in the study (a color-naming task and a questionnaire) and got a short overview about the theoretical background of the original Stroop task. Information about the two conditions and a potential association between the pictures and the words was withheld from the participants to avoid bias.
2. Participants gave their informed consent to take part in the study by signing the consent form.
3. Age and gender of the participant were recorded and entered into a table.
4. The software ‘PsychoPy’, entailing the items of both conditions in a random order, was run in the participant’s native language. Instructions for the task were provided by a text on the screen, telling participants to name the color of the word as quick as possible by pressing the arrow keys to indicate the color (left=red, down=green, right=blue). The instruction was supported by a picture of the arrow keys and fingers on it, because the finger position was standardized for all participants.
5. A practice trial of 15 NC items was provided first, with the purpose to make the participant learn the keys. This way disruption due to learning curves that could distort the sensible response time measures was reduced.

6. The actual trial of 200 items (100 AC and 100 NC) was provided in eight blocks of 25 items each. Every participant was exposed to both conditions which were mixed randomly by the program to avoid patterns that could possibly give respondents an idea about the study's assumption. For each trial, the response time of the respondents to each item was measured by the program.
7. Participants filled out the printed post questionnaire, to test the validity of the employed stimuli. Here, the respondents were asked to indicate what they think about, when seeing picture and word together.

2.3 Participants

The research involved a convenience sample of 40 people, consisting of undergraduate students from the University of Twente and people from the researchers wider social network (M_{age}= 32.33, SD_{age}= 13.97; 47.5% female, 52.5% male). Exclusion criteria were minors and people who are not German, Dutch or English native speakers. The age range spanned from 20 to 63 years. The student participants were offered SONA Credits for their participation. All respondents took part in the study voluntarily, which was ensured by an informed consent. The Ethics Committee BMS approved this research as ethically justifiable.

2.4 Design and Data Analysis

The present study employed the SPST in a repeated measures within-subject design with two conditions. Accordingly, each participant was exposed to the intended association condition as well as to the intended non-association condition (pre-questionnaire-classification), which are explained above in detail. Through the subsequent validity check, the 'true conditions' on which the data analysis rests were established (post-questionnaire-classification).

Consequently, the independent variable in the present experiment was the *Pairing Condition reclassified by the questionnaire* with two levels (AC; NC). The dependent variable that was measured was the *response time* (Figure 1). The conditions were presented in random order, where one picture-word pair constituted one stimulus. So, the complete pairs were randomly ordered for every participant, not single pictures or words.

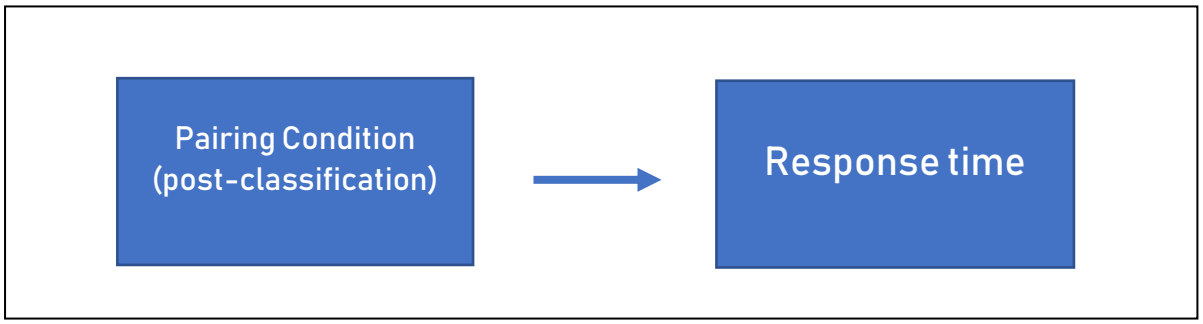


Figure 1: The present study investigated the Semantic Priming Stroop Task by observing the response times to the two pairing conditions in the post-questionnaire-classification.

As the first step of the data analysis, the collected response time measures were screened to determine the final dataset. Incorrect answers that were given by participants in the SPST were eliminated, so only correct answers would count into the data.

2.4.1 Multilevel model. The employed within-subject, repeated measures design is translated into a multilevel model. Since each participant was exposed to both conditions, it is not only possible to estimate the difference in mean RT on a population level but also on the participant level. The resulting provision with a population estimate as well as one submodel per participant, enables us to gather further insight into details of the collected data such as potential effects in single individuals, that might be missed when regarding only the general result on the population level. Moreover, the fact that subjects act as their own control provides a way of reducing the amount of error arising from natural variance between individuals.

2.4.2 ExGaussian distribution. Considering that the present study measures the dependent variable on the basis of response times, it is pointless to use models that assume a normal-distribution (e.g. t-tests) because response times are not normally distributed (Schmettow, 2018). This is the case because any activity takes at least a certain minimum amount of time to execute, so that response times can logically not be negative or close to zero, which results in a strong skew.

But not only the Gaussian distribution is inappropriate to plot response time measures. Since response times cannot be close to zero, also a Gamma distribution does not work well to plot this kind of data, because it assumes that response times of zero are possible. This way, moving right in direction, the gamma distribution is more stretched/ there is a reduction in

skewness that makes the graph more symmetrical, often overdoing the left tail (Schmettow, 2018). In other words, response time data is generalized by Gaussian or Gamma distributions because their assumptions divert from the characteristics of this kind of data, which can consequentially lead to the introduction of serious bias (Marmolejo-Ramos, Cousineau, Benites & Maehara, 2015).

What better presents the strong left offset of response time measures is the ExGaussian distribution. This is the case because it is composed of a normal distribution and an exponential distribution. The normal distribution describes the many factors that make up the response selection and motor processes that partly comprise response time, whereas the decision mechanisms are (according to neurological studies) best described by an exponential distribution. The sum of both of these time sets thus has a distribution best described by a combination of a Gaussian distribution (n-distribution) and an exponential distribution, which is called the ExGaussian distribution (Marmolejo-Ramos et al., 2015). Like this, the ExGaussian distribution shows realistically the response times with a steep left climb at an arbitrary point different from zero and a long right tail. Hence, for this study's data analysis, an ExGaussian distribution is assumed, therefore an ExGaussian regression equation is employed.

For the data analysis the statistics software 'R' was used. For the results, parameter estimates and credibility limits (95%) were utilized to observe the effect on the response time directly.

3. Results

The present study investigated whether the suggested effect of the SPST (i.e. that true associations cause a delay in response time in comparison to no associations) works under optimal conditions. If the SPST works, thus a positive difference in response times from the non-association condition to the association condition is expected. First, the population-level was regarded.

3.1 Population-level Analysis

To get a first impression of the difference in response times (RT) between the association condition and the non-association condition, a boxplot of the data was compiled, based on the original classification of the stimuli (pre-questionnaire-classification). As can be seen, the distribution of the RT in the AC is scattered wider than those in the NC, which again supports the choice of the ExGaussian distribution; i.e. a steep left climb and a long right tail clearly

emerge (Figure 2). Nevertheless, the central value of RT appears to show nearly no difference between the two conditions.

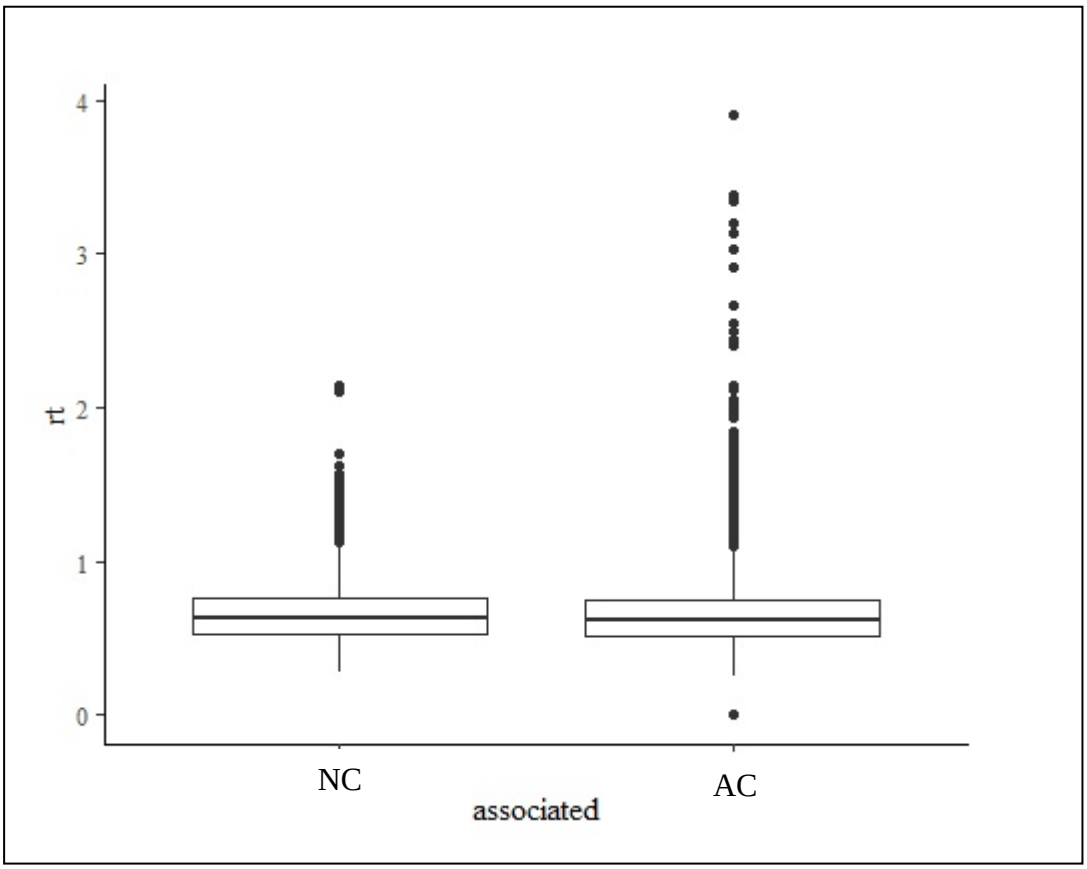


Figure 2: RT distributions in the association condition and in the non-association condition (pre-questionnaire-classification)

Next, a comparison of groups model (CGM) was conducted. The following values are based on the reclassification of the AC stimuli as identified by the questionnaire (post-questionnaire-classification). As can be seen in the table, the NC has an average RT of 0.684 seconds (Table 2). Since the respective credibility limits between 0.656 and 0.71 are tight around the center value, one can be rather sure that the group mean quite accurately describes the true average RT in the NC in the population.

Further, the coefficients reveal that on average in the population, the difference in RT between the two conditions amounts to -0.003 seconds, which is a value in the opposite direction (negative) than hypothesized. Thus, participants took on average 0.003 seconds *less* time in the AC than in the NC. However, it is still practically zero. The credibility limits here are considerably tight around zero, so that one can be quite sure that the true difference in RT is practically zero.

Table 2

Average response time to SPST on population level in the non-association condition (NC) and the difference in RT between the two conditions (post-questionnaire-classification; invalid AC items taken out)

Parameter	Center	(CI 95%) lower	(CI 95%) upper
Intercept	0.6836951	0.6556274	0.7100678
associated	-0.0031483	-0.0148912	0.0094687

Deviating from what was expected, a positive difference in RT from NC to AC could not be observed, but a minimal negative one. It can be concluded that the difference in RT between the conditions is practically zero. This supports the conclusion that there is no effect.

However, granted that the upper credibility limit of about 0.012 seconds would be the true value for the difference in RT between the conditions and one would claim that the SPST works (even though the value is extremely small), the test is just impractical. When the difference value is so very small, one would need to make endless trials to come to a useful result.

3.2 Participant-level Analysis

Second, it was zoomed in on the participant-level to see whether the SPST works for single individuals. This was done by means of error bars that take into account the credibility limits (Figure 3). As visible here, with credibility limits tight around zero, the center estimates show a difference between both conditions close to zero. For no single participant it can be said with a 95% certainty that there was an effect.

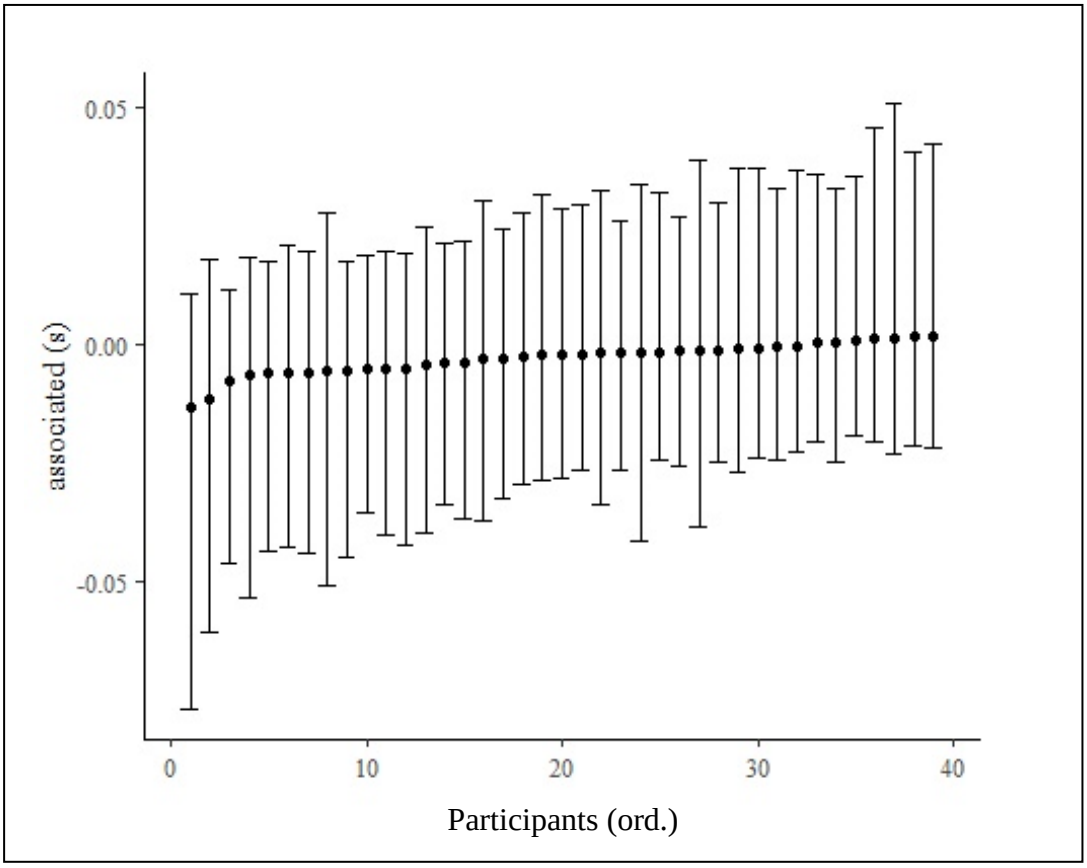


Figure 3: Center estimates and credibility limits of the difference in RT between the conditions per participant.

Nevertheless, even when assuming that the SPST works for single individuals, the test again becomes clearly impractical. Researchers who want to employ this test would first have to find out whether their participant is suitable for the test or not.

In sum, the results show that there is no population-level effect, as well as no participant-level effect.

4. Discussion

As already noted at the beginning of this paper, implicit methods such as the SPST can help researchers avoid bias in regards to a transparent research question. With the SPST in specific, one could potentially assess participants knowledge about a specific topic or even simply whether someone possesses a specific association (e.g. opinions, values), maybe even for diagnostic purposes. It can be concluded that the SPST would without question be a useful utility for many study designs as well as for diagnostic purposes.

However, in contrast to what was expected, no semantic priming Stroop effect was found, neither on population-level nor on individual-level. These results cast very critical light

on the paradigm. In other words, the question is why Sparrow et al. (2011) found an effect and we could not find one.

An issue that strikes attention is their use of t-tests that assume normal distributed data, when the data consists of response times (which are not normally distributed as explained above), as well as their use of p-values. Consequentially, Sparrow et al. only found strong evidence against the null-hypothesis with a p-value of $p < .001$. As explained before and supported by the paper of Gelman and Loken (2013), P-values and the assumption of a 0.05 alpha-level are problematic. When considering the response time group means, there is a difference of 0.12 seconds between the conditions. This is not a very strong effect regarding that the original Stroop effect was observed to have a value of 0.52 seconds, 0.46 seconds or 0.22 seconds in prior studies (Hatukai & Algorn, 2017; Linnman, Carlbring, Åhman, Andersson & Andersson, 2006). However, the effect is still much stronger than what was observed for the present study. Hence, P-values stay a problematic object but this is not an explanation for not finding an effect of the SPST at all in our study.

Of course, since the experiment has not been replicated until now, there is always a small chance that the results of Sparrow et al. consist of random noise, but this is rather unlikely. Therefore, we will first take a look at possible limitations of the present study.

4.1 Limitations

As mentioned before, the current study is not an exact but a modified replication of the study by Sparrow et al. (2011). One could thus criticize that our study may not be as optimal a replication of the SPST paradigm as it was proposed in the Sparrow et al. paper. This is especially the case because we used pictures instead of the trivia-question blocks.

To make the most crucial difference of the study designs clear again: Sparrow et al. made use of trivia-questions (easy and hard) that would potentially cause the participants to think about computers for help, resulting in a longer response time in the thereupon following modified Stroop task when the word was a computer word in comparison to when it was another word. In contrast, the present study utilized pictures that would potentially prime participants to think about the respective association, resulting in a longer response time in the modified Stroop task when the word induced an association in comparison to a word that did not. As can be seen clearly now, there is a notable difference in how the priming was initiated, as well as a clear similarity in the total process.

So, the underlying assumption stays the same in both studies. Namely, that a semantic priming process would disrupt the other cognitive process of color naming in the modified

Stroop task, visible in an increased RT. Nevertheless, what constitutes one important difference in the Sparrow et al. study is that questions in contrast to pictures are a task. It is common sense that looking at a picture will set a different mental process in motion than being requested to do something (answering a question). So, the first critique is that the present study failed to employ a task for the semantic priming as Sparrow et al did, which might be necessary for the SPST to work.

However, if the task only works with a task or more specifically questions, it puts severe limits upon its usability. Not everything you want to find out with help of the SPST might be possible to prime semantically in the form of questions. Following the assumption of Sparrow et al., questions are a reasonable measure for the purpose of priming some sort of information finding process. But it is questionable if you can prime a lot of other potential research topics with questions like fairy tales or personal traits.

Another aspect that was credited only as a side note in the study, is the fact that the experiment by Sparrow et al. (2011) built in an additional cognitive workload for the participants. After the priming process with the trivia-questions, the participants were to perform the modified Stroop task, thereby holding a 6-digit number in mind. The reason for this is not further explained but it could be a possible explanation why the SPST effect did not work in the present experiment, assuming that this extra amount of cognitive load plays a major role for the working of the task. Tracking down the argument for this additional treatment, we found that prior studies investigating the workings of the Stroop task build in an extra cognitive load.

In a study that investigated the effect of working memory load on selective attention by Gao, Chen and Russell (2007), a similar study design concerning the additional cognitive load was employed. Participants were exposed to either a single number or a 6-digit number before conducting a Stroop task consisting of congruent and incongruent color-words in regards to meaning and color. After the Stroop, a digit was shown and the participant had to decide whether the digit present or absent in the number they had seen before the Stroop. They found a reliable Stroop interference effect, but the effects magnitude did not change as a function of the additional cognitive load. They suggest that working memory load plays a negligible role for the outcome of the Stroop effect and concluded that working memory load showed no relevant effect on the degree of Stroop interference.

On the other hand, in an experiment by Kim, Kim and Chun (2005) observed that the cognitive loads influence on the Stroop task interference depends on the type of cognitive load. They found an increase in Stroop interference for when the type of cognitive load and

the type of required information for the target overlapped. When for instance the participant retains a set of nonsense-syllables (verbal) in mind and then implements a Stroop trial that involves to decide whether the meaning of a color-word (verbal) is the same as an also provided color patch, cognitive load and target are from the same type. They explain it with a limited capacity mechanism that are available for verbal target processing. Additionally, the study observed that when the type of cognitive load overlapped with the distractor of the Stroop task, Stroop interference decreased.

Coming back to the study of Sparrow et al. (2011), they utilized digits as additional cognitive load, which is (=verbal) and the modified Stroop consisted of a colored word. The required information here was the color (=non-verbal) and the distractor was the word-meaning (=verbal). According to the results of Kim et al. (2005) we would assume that the Stroop interference was decreased because the cognitive load was of the same type as the distractor (i.e. verbal). What this means for the Sparrow et al. study is that the RT might have been measured shorter in both conditions equally because the target was always the ink color of the words and the cognitive load remained the 6-digit number for hard, as well as for easy trivia-questions. Consequently, the additional cognitive load should not have interfered here in general but the reason for utilizing this extra task here stays unclear.

However, the recall rates on these numbers are not tested in the Sparrow et al. paper (2011) or at least this is not reported. Thus, it might be that some participants did not actually retain the number when others did, which would obviously introduce a serious bias.

Another point that one could criticize about the present study is that we tested the associations for validity but missed out on checking the non-associations for validity, i.e. whether these items did unintentionally evoke an association when they should not. But considering that the difference in response time is so close to zero, a change in results due to this second validity check seems very unlikely and leaves little space for debate.

4.2 Publication Process

Disregarding the issue of whether the SPST effect works or not, it is surprising that a high-ranked journal like ‘Science’ published a method that has never been used before and is (at least partly) not designed in a transparent way. The manuscript selection as well as the reviewing process are described quite strict by ‘Science’ themselves (“Science: Information for authors”, n.d.). Notwithstanding, the study was published this way.

As already mentioned earlier, researchers experience a high pressure to find significant results because in the field of science unfortunately, you only ‘win’ by writing the most

interesting publications (Bakker et al., 2012). The reason might be ‘sensationalism’ which scientific journals are accessible and vulnerable to even more than it is the case in regular journalism (Nature, 2017). As reported in a study by Song et al. (2010) the so-called publication bias, that significant results are favored over non-significant ones, is very likely in the dissemination process of research findings. Other studies confirm the existence of publication bias (Dickersin, Chan, Chalmers, Sacks & Smith, 1987).

Howsoever the publication process was implemented with the study of Sparrow et al. (2011) and neglecting the justness of the results, it is inexplicable how the reviewers did not question the poor design (i.e. few trials, no validity check for the prime, no check of the memory task) and the unexplained aspects (i.e. additional cognitive load) of the study. This leads to another point, namely the obviousness with which the results are treated.

The main claim of the study by Sparrow et al. was the change in strategy of thinking. The idea seems plausible and very appealing: People use computers to search for information so very regularly, at any place and time that whenever we do not know something by ourselves immediately, we would obviously think about computers. But taking a step back and thinking about it from a scientific perspective, it becomes clear that it is not very likely that a very recent technology like the internet would really fundamentally change the way people approach problems. Consulting prior brain research elucidates that fundamental brain processes are often thousands if not millions of years old (Nairne & Pandeirada, 2008), which implies that a few decades old invention like the internet presumably does not show in human brains in such a fundamental way already.

The point is that we do not want to deny Sparrow et al. their conclusion; the SPST effect might exist. But regarding the small amount of evidence and most of all the appealing nature of the claim, which potentially makes a researcher be biased against the outcome from the very beginning of the study, it seems grossly negligent to treat the results with the obviousness as was done in this case, also by the publishers. Results, especially appealing ones, should thus always be regarded with caution.

4.3 Future Research

To further investigate the functionality of the SPST and to find out whether the use of a task in form of questions is necessary for the test to work, the present study should be replicated with pictures as a first condition and trivia questions as a second condition. Even though the reason for the additional cognitive load that was built in to the study by Sparrow et al. (2011) stays unclear, a future research could include this aspect as a separate condition. This way it

can be investigated whether the retaining of a number in working memory indeed yields the needed but still unknown component to make the SPST work.

4.4 Conclusion

In this study, we did not find an effect of the Semantic Priming Stroop Task. Regarding the legitimate point of criticism that the present study utilized a picture instead of a question as prime, as was done by Sparrow et al., it can be claimed that we might have observed an effect if we used questions. Similarly, perhaps the additional cognitive load is for a yet unknown reason an important factor to produce the SPST effect.

Nevertheless, if the SPST works only with questions as a prime (and the reason for adding cognitive load to the task remains unexplained), the paradigm is thoroughly limited in its area of application. In conclusion we can thus say: No, in general the SPST does not work and the ways in which it possibly works are rather impractical.

References

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., et al. (2015). On the Reproducibility of Psychological Science. *Science, 349*(6251). doi:10.1126/science.aac4716

American Psychological Association (n.d.). Retrieved from https://dictionary.apa.org/replication

Bakker, M., van Dijk, A., & Wicherts, J. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Dickersin, K., Chan, S., Chalmers, T., Sacks, H., & Smith, H. (1987). Publication Bias and Clinical Trials. *Controlled Clinical Trials, 8*(4), 343-353.

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science, 7*(1), 45-52. doi:10.1177/1948550615612150

Galak, J. (2015). *Replication of Study 1 by Janssen, Schirm, Mahon, & Caramazza (2008, JEP:LMC).* OSF. Retrieved from osf.io/uhpyr

Gao, Q., Chen, Z., & Russell, P. (2007). Working memory load and the stroop interference effect. *New Zealand Journal of Psychology, 36*(3), 146-153.

Gawronski, B., LeBel, E., & Peters, K. (2007). What Do Implicit Measures Tell Us? Scrutinizing the Validity of Three Common Assumptions. *Perspectives on Psychological Science, 2*(2), 181-193. doi: 10.1111/j.1745-6916.2007.00036.x

Gelman, A., &. Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition" or “p-hacking" and the research hypothesis was posited ahead of time.

Hatukai, T., & Algom, D. (2017). The Stroop Incongruity Effect: Congruity relationship reaches beyond the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1098-1114. doi:10.1037/xhp0000381

Heyman, T., Van Rensbergen, B., Storms, G., Hutchison, K., & De Deyne, S. (2015). The influence of working memory load on semantic priming. *Journal of Experimental Psychology: Learning Memory and Cognition, 41*(3), 911-920. doi:10.1037/xlm0000050

Ioannidis, J. (2019). Why Most Published Research Findings Are False. *CHANCE, 32*(1), 4-13. doi:10.1371/journal.pmed.0020124

Janssen, N., Schirm, W., Mahon, B., & Caramazza, A. (2008). Semantic Interference in a Delayed Naming Task: Evidence for the Response Exclusion Hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition, 34*(1), 249-256. doi:10.1037/0278-7393.34.1.249

Johnson, D., Cheung, F., & Donnellan, M. (2014). Does Cleanliness Influence Moral Judgments? A Direct Replication of Schnall, Benton, and Harvey (2008). *Social Psychology, 45*(3), 209-215. doi:10.1027/1864-9335/a000186

Kim, M.-S., Min, S.-j., Kim, K., & Won, B.-Y. (2010). Concurrent working memory load can reduce distraction: An fMRI study. *Journal of Vision, 6*(6), 125-125. doi:10.1073/pnas.0505454102

Linnman, C., Carlbring, P., Åhman, Å., Andersson, H., & Andersson, G. (2006). The Stroop effect on the internet. *Computers in Human Behavior, 22*(3), 448-455. doi:10.1016/j.chb.2004.09.010

Lynott, D., Corker, K., Wortman, J., Connell, L., Donnellan, M., Lucas, R., & O'Brien, K. (2014). Replication of "experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology, 45*(3), 216-222. doi:10.1027/1864-9335/a000187

Macleod, C.M. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin, 109*(2), 163-203. doi:10.1037/0033-2909.109.2.163

MacLeod, C.M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121*(1), 12-14. doi:10.1037/0096-3445.121.1.12

Marmolejo-Ramos, F., Cousineau, D., Benites, L., & Maehara, R. (2015). On the efficacy of procedures to normalize Ex-Gaussian distributions. *Frontiers in Psychology, 5*, 1548. doi:10.3389/fpsyg.2014.01548

Mewhort, D.J.K., Braun, J.G., & Heathcote, A. (1992). Response Time Distributions and the Stroop Task: A Test of the Cohen, Dunbar, and McClelland (1990) Model. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 872-882, doi:0096-1523/92/S3.00

Monahan, J. (2001). Coloring single stroop elements: Reducing automaticity or slowing color processing? *Journal of General Psychology, 128*(1), 98-112. doi:10.1080/00221300109598901

Nairne, J., & Pandeirada, J. (2008). Adaptive Memory: Remembering with a Stone-Age Brain. *Current Directions in Psychological Science, 17*(4), 239–243. doi:10.1111/j.1467-8721.2008.00582.x

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science, 7*(6), 528-530. doi:10.1177/1745691612465253

Pothos, E.M., & Tapper, K. (2010). Inducing a Stroop Effect. *Applied Cognitive Psychology, 24*, 1021–1033. doi:10.1002/acp.1603

Schmettow, M. (2018). *New statistics for the design researcher A Bayesian course in tidy R.*

Schmettow, M., Noordzij, M., & Mundt, M. (2013). An Implicit Test of UX: Individuals Differ in What They Associate with Computers. *CHI Extended Abstracts.* Paris, FR: ACM

Schnall, S., Benton, J., & Harvey, S. (2008). With a Clean Conscience Cleanliness Reduces the Severity of Moral Judgments. *Psychological Science, 19*(2), 1219-1222. doi:10.1111/j.1467-9280.2008.02227.x.

Science: Information for authors (n.d.). Retrieved from https://www.sciencemag.org/authors/science-information-authors

Science journalism can be evidence-based, compelling - and wrong. (2017). Nature, 543(7644), 150-150. doi:10.1038/543150a Science journalism can be evidence-based, compelling - and wrong. (2017). *Nature, 543*(7644), 150-150. doi:10.1038/543150a

Sergeant, J., Geurts, H., & Oosterlaan, J. (2002). How specific is a deficit of executive functioning for Attention-Deficit/Hyperactivity Disorder? *Behavioural brain research, 130*(1-2), 3-28.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Song, F., Parekh, S., Hooper, L., Loke, Y., Ryder, J., Sutton, A., . . . Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1-220. doi:10.3310/hta14080

Sparrow B., Liu J., Wegner, D., M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333*(6043), 773-776. doi:10.1126/science.1207745

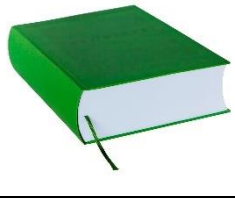
Szűcs, D., Soltész, F., Bryce, D., & Whitebread D. (2009). Real-time Tracking of Motor Response Activation and Response Competition in a Stroop Task in Young Children: A Lateralized Readiness Potential Study. *Journal of Cognitive Neuroscience, 21*(11), 2195-2206. doi:10.1162/jocn.2009.21220

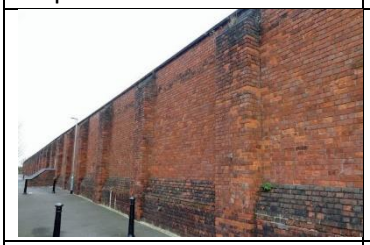
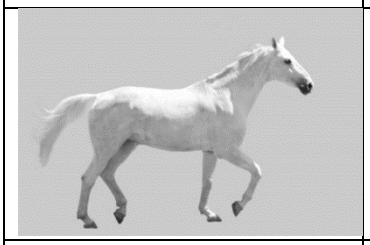
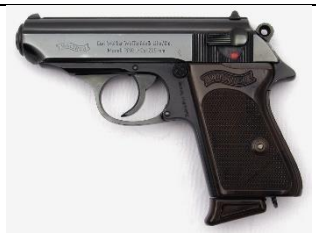
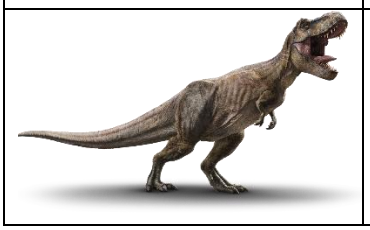
Tulving, E., Schacter, D., & Stark, H. (1982). Priming Effects in Word-Fragment Completion Are Independent of Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(4), 336-342. doi:10.1037/0278-7393.8.4.336

Williams, L., & Bargh, J. (2008). Keeping One's Distance: The Influence of Spatial Distance Cues on Affect and Evaluation. *Psychological Science, 19*(3), 302-308. doi:10.1111/j.1467-9280.2008.02084.x.

Appendix 1: Catalog of Stimuli

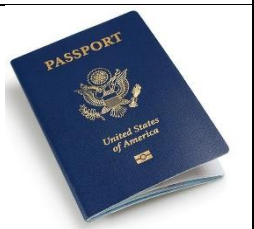
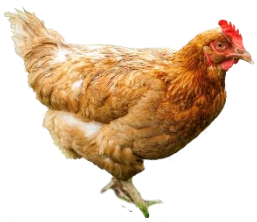
Association condition items (50)

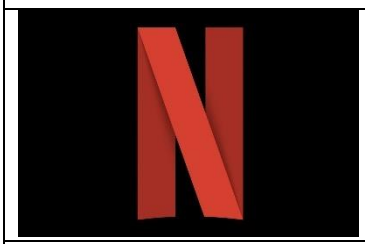
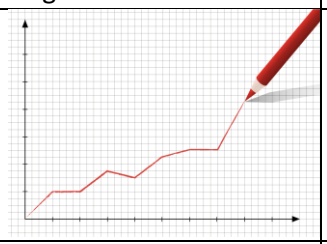
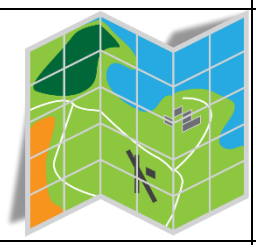
			
Disease	Physics	Trilogy	Brain
			
Romance	Thriller	Poisoned	Pretzels
			
Man	Tower	Apple	Astronaut
			
Plantation	Grandmother	Snake	Egg
			
Gingerbread	Fire	Records	Stepsisters
			
Pirate	Nephews	Billionaire	Ireland

			
Mouse	Crown	Greenhouse	President
			
Exit	Monarchy	Winter	Christmas
			
Ship	Chocolates	Meerkat	Wizard
			
USA	Germany	Sweden	Wolf
			
Rainbow	Martini	Man	Princess
			
Unwanted	Gun	Fairy	Stone
			

Park	Airplane

Non-association condition Items (50)

			
Trash	Alphabet	Percent	Clock
			
Backpack	Burger	Chocolate	Motorcycle
			
Silence	Street	Politics	Goodbye
			
Email	Forest	Lamp	Maths
			
Shop	Box	Poster	Advice
			
Bottle	Payment	Human	House

			
Police	Dog	Fish	Window
			
Car	Cup	Book	Pen
			
Letter	Dance	Switchboard	Doctor
			
Spotlight	Cap	Hospital	Alcohol
			
Floorboards	Weather	Statistics	Flag
			
Monkey	Fork	Office	Snake
			

Music	Bird	Cat	Bag
			
Zoo	Phone	Joke	Fruit
			
Smartphone	Song		Europe
			
Language	Hairdryer	Library	Parrot
			
Hammer			