A cognitive model of the same-different task based on the inhibition of "different" answers

Vincent LeBlanc

A thesis submitted in partial fulfillment of the requirements for the
Master's degree in Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

Table of Contents

Acknowledgments

I would like to thank the Natural Sciences and Engineering Research Council of Canada, the government of Ontario, the government of Québec and the Faculty of Social Sciences of the University of Ottawa for their financial support.

I am grateful for the many stimulating discussions I had with Denis Cousineau, Bradley Harding, Marc-André Goulet and Philippe Vincent-Lamarre, which were invaluable and greatly helped me progress.

Furthermore, I thank Claude Richard-Malenfant and Julien T. Groulx for collecting the data for my experiments and the quality of their work in the laboratory.

I send a special thanks to Étienne Dumesnil, as his remarks on the non-linear impact of matches on response times and the discussion that followed were the spark that led to the creation of my proto-model. This thesis would have taken a very different path without him.

Finally, I would not have gotten this far without the unconditional love and support from Henrietta Blinder, my anchor during the (too many) storms that graduate studies imposed on me.

Abstract

"[The] sense of sameness is the very keel and backbone of our thinking" (James, 1890). To make sense of the ever-shifting information in our environment, we constantly assess whether the world around us changes or not, if objects are the "same" or if they are "different". This basic decision-making process is found from the lowest level of cognition (e.g. when contrasts are encoded by the retina), to the highest (e.g. when comparing concepts), and anywhere in between. In an experimental context, this process is studied with the "same-different" task, where subjects are asked if two stimuli presented sequentially are strictly identical or not. This experiment has been documented since the 1960s and its results have been replicated with diverse stimuli types (letters, shapes, faces, words, etc.). However, every attempt to model the subjects' accuracy and response times on correct and incorrect answers simultaneously was unsuccessful so far. Part of the challenge in explaining this task is that "same" answers are faster than expected compared to "different" answers, a phenomenon called the "fast-same effect".

This thesis aims to assess whether a formal model based on the inhibition of "different" answers is plausible, effectively changing the problem from "fast-same" to "slow-different". In the first chapter, I review the previous theories and models of the same-different task to learn why they failed. By elimination process, I identify the only cognitive architecture that seems congruent with the data. I then propose a model prototype based on the inhibition of "different" answers that implements this architecture. In the second chapter, I test this prototype with an experimental paradigm designed specifically to assess its plausibility. I conclude that resources should be spent in developing a formal model based on the inhibition of "different" answers, as the prototype's qualitative predictions are confirmed by both the typical same-different data and the newly acquired data.

The Same-Different Task

"[The] sense of sameness is the very keel and backbone of our thinking" (James, 1890). To make sense of the ever-shifting and noisy information we receive at each instant, we must be able to know what changes and what does not. This "sameness" is highly subjective and depends on the context; is this the same leaf even though it moves with the wind? What if it witters away from the cold? This basic decision-making process is found from the lowest level of cognition (e.g. when contrasts are encoded by the retina), to the highest (e.g. when comparing concepts), and anywhere in between.

To study the ability of identifying sameness or detecting differences experimentally, we need to measure participants. The simplest task that achieves that is called the "same-different" task, as first coined by Nickerson (1965). By showing two stimuli and asking participants to answer by pressing one button for "same" and another for "different", we obtain the purest behavioral measure of human performance on identity relationship decision-making. Our laboratory has been mostly studying Bamber's same-different task (1969), in which participants must judge as quickly and accurately as possible whether two strings of letters presented sequentially (S_1 then S_2) are the same or if they differ on any of their letters. S_1 and S_2 have an equal number of letters (L), letters that match between S_1 and S_2 do not change positions and a letter can only be present in one position per trial. The strings can contain 1 to 4 letters and can have any number of mismatches (or differences, D). Consequently, the number of matches, M, is the number of letters that are not mismatches, such that $M + D = L$. The participant answers "same" when there are no mismatches ($D = 0$ or $M = L$) and "different" when there is at least one mismatch ($D > 0$ or $M < L$). The task is balanced to have an equal number of "same" and "different" trials to reduce response bias. There is also an equal number of 1 letter ($L1$) trials, 2

letters ($L2$) trials, $L3$ trials and $L4$ trials. Additionally, for a given L, there is an equal number of trials for all numbers of mismatches. Table 1 summarizes the distribution of trial types. This design allows to study the impact of manipulating only L or D in a somewhat continuous fashion.

This task has been studied for more than 50 years and has produced reliable results with all sorts of stimuli, such as letters, shapes and faces to name a few (Bindra, Donderi, & Nishisato, 1968; Decker, 1974; Egeth, 1966; Eriksen & O'Hara, 1982; Hock, 1973; Nickerson, 1965, 1967; Nishisato & Wise, 1967; Proctor, Rao, & Hurst, 1984; Taylor, 1976). Yet, as of today, no one has been able to adequately model and explain how the response times (RT) and accuracies vary with L and D in this task. I propose to contribute to the current body of knowledge in two ways. In my first chapter, I review the theories that attempted to explain the same-different task and present aggregated data from my laboratory to falsify or validate their claims. These insights lead me to propose a proto-model that could account for all the richness of the aggregated data. In the second chapter, I first present the effects of the Delayed Presentation protocol and of redundant targets on the RT of decision-making tasks. I then define the predictions of the proto-model in the light of these two theoretical contexts. Finally, I use the Delayed Presentation protocol to induce early termination in participants and show how the proto-model seems to explain the data of the same-different task adequately.

1 Literature and Data Review

In the first part of this chapter, I review the theories that were suggested to explain parts of the same-different task data, highlighting where they failed and what can be salvaged from them. In the second part, I detail my first project, in which I aggregated the data of eight experiments ran in Denis Cousineau's laboratory. This data reveals additional restrictions that models of the same-different task must respect, leaving very few likely model options. This allows me to propose a proto-model that could theoretically fit the typical RT and accuracy data. I then present this proto-model's qualitative predictions on the speed of errors, on which most previous theories are silent, and show how it fits this new aspect of the data.

1.1 Literature Review

In this section, I review Bamber's *Identity Reporter* (1969), Krueger's *Noisy Operator Theory* (1978), Proctor's *Encoding Facilitation Theory* (1981), Ratcliff's *Diffusion Model* (1978) and Nickerson's reflections on a bias for positive answers (1965) and on self-terminating "same" answers (1967). Each of these theories and commentaries explain in their own way some parts of the RT or accuracy data and offer a different perspective on how we make decisions.

To clarify the results and predictions, I break down the "different" trials ($0 < D \leq L$), into "some-different" trials ($0 < D < L$) and "all-different" trials ($D = L$). This does not change the response given by the participant ("different") but helps to differentiate the conditions in which there is at least one match ("some-different") from those which exclusively contain mismatches ("all-different").

1.1.1 The "fast-same effect" and the Identity Reporter

The most well-known RT result of the same-different task is that correct "same" answers are faster than correct "different" answers (Egeth, 1966; Nickerson, 1965). The reason this result

is discussed in almost every paper on the same-different task is that it goes against our naïve intuition on how we compare objects. A "same" judgment should be exhaustive (i.e. all letters should be tested) because we cannot know if all the letters match between two stimuli before testing each of them. On the other hand, a "different" judgment can be self-terminating, as we know two stimuli are different as soon as one mismatch is found. Hence, "different" judgments should be at least as fast as "same" trials, and the fact that they are not (most of the time) has been the main concern of most models of that task. Figure 1 shows typical RT results for the same-different task.

1.1.1.1 The "fast-same effect".

In his first article on the same-different task, Bamber (1969) summarizes what we naïvely expect of the RT on the same-different task using these equations:

$$RT_s = a_s + M \times b_M \qquad \text{Equation 1}$$
$$RT_d = a_d + \left(\frac{M}{D+1}\right) \times b_M + b_D,$$

where the s and d indices refer to "same" and "different" trials, M and D refer to matches and mismatches, a is the non-decision time (i.e., the encoding and response-production stages) and b is the time needed to test a letter (i.e., b_M is the slope of RT when M increases).

In these equations, the RT of a "same" judgment depends solely on M, increasing by b_M for each match in the stimulus. A "different" judgment, however, is self-terminating and does not require that every match be tested. Assuming participants look at letters in a random order, they scan $\frac{M}{D+1}$ matches on average before finding the first mismatch, at which point they answer. Hence, for each match in the stimulus, the RT of *D1* trials increases by $0.5 \times b_M$, the RT of *D2* trials by $0.\overline{3} \times b_M$, and *D3* trials, by $0.25 \times b_M$. In other words, the slope of *D0* trials should be twice that of *D1* trials, three times that of *D2* trials and 4 times that of *D3* trials.

Instead, the RT slope of *D1* trials is slightly greater than the slope of *D0* trials, the slope of *D2* trials is about equal to that of *D0* trials, and the *D3* slope is smaller. Furthermore, the *L1D0* answers are faster than *L1D1* answers by approximately 50 ms (depending on the task, the stimuli, the interstimulus interval, etc.). The combination of these two results, the smaller slope and faster *L1* trials for "same" trials, has been coined the "fast-same effect" by Bamber (1972).

1.1.1.2 The Identity Reporter theory.

In his 1969 article, Bamber proposes a two-process model, illustrated in Figure 2, to explain the speed of "same" answers. The first process, the "Identity Reporter", can only answer "same", is exhaustive and scans the stimuli in parallel, which makes it faster than a serial process. The second process, the "serial processor", is self-terminating but serial, and thus generally slower than the Identity Reporter, but can answer both "same" and "different". Because the two processes are racing, the Identity Reporter will be giving the "same" answers most (but not all) of the time, leading to a faster average "same" RT.

An assumption of Bamber's model is that the Identity Reporter can only operate on the physical properties of the stimulus, which was invalidated by the results published in his second article (1972). In this new task, participants had to answer "same" if the letters had the same name, regardless of their physical appearance (e.g. "A" and "**a**" are "same"). Not only was the fast-same effect still present, but the RT of all conditions were slower than on his first task (Bamber, 1972, Figure 3). This suggests that the task is typically done by comparing visual features, otherwise the RT would not be slower on this new task, but that the fast-same effect cannot be attributed strictly to a visual Identity Reporter.

This was not the only reason that led Bamber to reject his model. The Identity Reporter also predicted equal RT for *L1D0* and *L1D1* trials; if the Identity Reporter is only faster than the

serial processor because it processes letters in parallel, the RT of single letter "same" and "different" trials should be equal. Hence, his model is unable to explain that *L1D0* trials are faster than *L1D1* trials (a result he obtained in both his experiments).

Another reason for rejecting his theory was that the slope of the "same" RT in his results had an upward concave shape. Bamber proved that parallel exhaustive models can only make two qualitatively different predictions on the slope. If we assume that testing a single letter always takes the same time without any variability within letters (i.e., testing two "I" always takes the same time) and between letters (i.e., testing two "I" takes the same time as testing two "Q"), then a parallel exhaustive model would predict a linear increase in RT. If we allow for any combination of variability within and between letters, then this model would predict a downward concave shape. The mathematical proofs for this reasoning are found in Bamber's article (1972, Appendix) as well as in Sternberg's review of the different modeling attempts of this task (Sternberg, 1998, p. 425). Hence, a parallel exhaustive process cannot be responsible for the upward concave slope of RT for "same" answer, which led him to reject all theories based on this architecture.

Unfortunately for Bamber, I have shown in two ways that this upward concave slope is the result of a sampling problem. The first way was found when thinking about the design and analysis method applied to the data. Because participants have very few trials in some conditions, the mean might not be an adequate measure of central tendency. This problem is made worse by the fact that most past experiments had a small number of participants (often smaller than 10) and that the distribution of RT is known to be highly non-symmetrical. Hence, I argued that this upward concave slope stemmed from using the mean RT per condition and per participant, and proposed to replace it with the median. As I suspected, using their median

resulted in linear slopes for the "same" RT in most experiments. The second way that I showed that the upward concave slope was a sampling problem was during my data review (section 1.2), where the large number of participants resulted in a linear slope even when using their mean RT.

Bamber's contribution to the understanding of the same-different task was important. Although he only worked on it for a brief period, his articles brought the issue to the attention of many researchers and are the most cited on the subject. His equations have summarized our intuitions and suggest that the RT of both "same" and "different" answers should be mainly dictated by the number of matches. Also, his rebuttal of parallel exhaustive processes is incomplete (as I will discuss in Chapter 2) but still helped to reduce the number of alternatives available to explain the data on this task.

1.1.2 Internal noise and the Noisy Operator theory

In his Noisy Operator theory (1978), Krueger explains that noise has a greater impact on our sense of sameness than on detection of change. He argues that because our senses are imperfect, we need to filter out both perceptual and internal noise to perceive static objects as such, otherwise we would constantly be alerted by tiny irrelevant changes. These noise-triggered signals are not specific to the same-different task and apply to everyday situations, which makes it likely that evolution could have given us a way of dealing with it. I will first present his intuitions and then discuss his model.

1.1.2.1 Differential impact of noise on "same" and "different" judgments.

The rationale behind Krueger's model is based on the impact of noise on the encoding, decision-making (or testing) and response-production stages of cognition. During the encoding process, the physiological aspects of vision makes it obvious that a matching stimulus would sometimes be encoded as mismatching. The information we gather on the world is ambiguous

and noisy, but we can work through this uncertainty if we are given enough time and context. In an experimental task, we have none of those: we require speeded decisions from the participant and decontextualize the stimuli as much as possible. Thus, it is possible that some elements will be mis-encoded, if only for a brief moment, and using this incorrectly encoded information to make a decision can then lead to errors.

When mis-encoding a matching letter, it is guaranteed that the incorrectly encoded letter will not match anymore. For example, when presented with "B", we might encode "β", "P", "8" or something completely unintelligible. For a mismatching letter, however, it is very unlikely that noise will result in encoding a matching letter. This is because there is an infinite number of ways to mis-encode "B" into something that does not match, and only one way of mis-encoding it into the matching letter. For example, if the letter in S_1 is "L" and the corresponding letter in S_2 is "B", the odds of mis-encoding the "B" into a "L" are very low. Hence, matching letters are far more sensitive to noise than mismatching letters during the encoding stage. Consequently, noise in this stage of cognition will lead to an error on a "same" trial because it generates a mismatch, which requires a "different" answer. Noise on a "different" trial, on the other hand will not result in a less different stimulus. In fact, a "some-different" trial benefits from noise if a match is mis-encoded. This differential advantage of "different" over "same" trials increases with *L*.

Once the letters are encoded, they are tested by the decision-making process, where the encoded versions of the letters in S_2 are compared to S_1's and transformed into matches and mismatches. If noise intervenes in the process, matching letters could be interpreted as mismatches and mismatching letters as matches. On a "same" trial, a single mistake in the decision-making process would result in a "different" trial (because there would be at least one mismatch). However, there must be a testing mistake on every mismatching letters as well as no

mistakes on all matching letters to turn a “different” trial into a “same” trial. Hence, just like for the encoding stage, “same” trials are sensitive to noise during the decision-making stage, and more so as *L* increases. On the other hand, “different” trials are less prone to errors, becoming more resistant as *L* increases, and even more resistant if *D* increases.

Finally, noise can affect the response-production stage, leading to a “same” answer even though the decision-making stage found a mismatch or a “different” answer even if only matches were found. This is often reported by participants, who claim they knew the answer but were not able to stop themselves from pushing the wrong button. At a first glance, there are no reasons to believe that this happens more often for “same” or “different” trials, as experiments are balanced to have an equal number of trials of each and to have an identical physical response (counterbalanced across participants). However, there have been some arguments in favor of a bias for “same” judgments (discussed in section 1.1.4), which could be implemented as a tendency to make one response more than the other, regardless of the result of the encoding or decision-making stages.

1.1.2.2 The Noisy Operator theory.

Krueger’s model is based on these intuitions regarding the differential effect of noise. In this model, there is a process that counts the number of mismatches between the two stimuli. For each count, there is an associated probability that the stimuli are truly the same or truly different, which defines a (discrete) probability density function (PDF) of counts for both answers. These two functions overlap to some extent, such that a low enough count can only be generated by a “same” trial and a high enough count can only be generated by a “different” trial. Krueger defined these threshold values as the points where the likelihood ratio of the two probabilities is 25 times more in favour of one answer. If the mismatch count falls between the two thresholds, a

new “pass” begins: no decision is made, the PDFs and likelihood ratios are recalculated based on the previous mismatch count, and the number of mismatches is recounted.

To implement noise, Krueger defined his stimuli as matrices of one hundred 1’s and 0’s. The encoding noise is simply a chance of randomly flipping each of these values. Because the noise during response production can be considered a response bias, this noise is reflected in the mean of the “same” and “different” PDFs. Similarly, the noise in the testing stage is implemented as the variability around the means of the PDFs.

If Krueger’s intuitions are sensible and clever, the results in his 1978 paper were not convincing. First, he made many arbitrary decisions regarding his model, which were only justified by their fits with the data. Such choices include the duration of a pass, fixed at 200 ms, the likelihood ratio thresholds to answer “same” or “different”, fixed at 25:1, and the number of bits per letter, fixed at 100 (Krueger, 1978, pp. 285–286). Second, some parameters that should be fixed per participant were instead estimated for each experiment. For example, the non-decision time and the probability of flipping a bit changed across experiments (Krueger, 1978, Table 1). Third, the RT fits were decent for his single letter experiments, keeping in mind that each pass lasts 200 ms. However, the data he observed was unusual. On two out of three experiments (labeled “Case 1” and “Case 3”), the standard deviations were very high, between 186 ms and 330 ms, and oddly close to being half of the mean RT. In his experiment labeled “Case 2”, the RTs were closer to what would be expected, but the task was not directly comparable to Bamber’s because S_1 and S_2 were geometrical patterns presented simultaneously (Krueger, 1978, Table 3). Fourth, he used summary values of other experiments to test his model on stimuli of more than 1 letter, which obviously led to poor fits (Krueger, 1978, Table 6). Finally, his model is inconsistent when estimating accuracies (Krueger, 1978, Tables 2 and 5)

and produces poor fits for incorrect RT (Krueger, 1978, Tables 4 and 7), which suggests that the method for updating the distribution after each pass is inadequate.

Given the theoretical focus on noise and error, the model's poor performance on predicting accuracy did not convince many to further explore rechecking as a model of RT. Even if its predictions are lacking in many ways, Krueger's model was the first to model both RT and accuracy. More importantly, his intuitions regarding noise still hold and can be used to understand the accuracy results, a method I will apply in subsection 1.2.3.

1.1.3 Encoding Facilitation theory

The Encoding Facilitation theory of Proctor was the first to suggest that the difference between the "same" and "different" RT was not caused by the comparison process, but by the encoding process (1981). This theory is based on a simple residual activation principle, which is that a concept that was activated recently will be temporarily easier to reactivate than any other concept. It explains the data observed in Bamber's nominal task (1972) by stating that we manipulate concepts as well as visual objects. For example, both the "a" and "A" objects would activate the concept of the letter "a". If S_1 and S_2 are physically identical, the *visual* residual activation will help us identify S_2 faster, thus beginning the comparison process earlier than if they were physically different. However, the *conceptual* residual activation would still result in faster RT than if the two letters had different names. Thus, in a "same" trial, the encoding of S_1 would facilitate the encoding of S_2, whereas it would not in a "different" trial.

Although this theory elegantly addresses the RT difference between *L1D0* and *L1D1* trials, it also has weaknesses. One problem stems from a crucial aspect of his model, which is that encoding S_1 inhibits the encoding of things that are not S_1. Proctor says that it is the difference in the name of the two stimuli that creates encoding inhibition. However, this

inhibition cannot rely on knowing the name of S_2, otherwise S_2 would need to be encoded and identified before the encoding inhibition can take place (a chicken and egg problem). Thus, to inhibit the encoding of S_2, S_1 must inhibit every other concept, even those that are not relevant to the task (e.g., seeing "A" inhibits "red space-suit"). This would be impossible to apply to our daily life and shows that the theory was not developed for stimuli of more than one letter.

Another problem with Proctor's theory is that he failed to address how the nominal inhibition is done, a matter with a large impact on trials containing matches and mismatches. There are three reasonable implementations of this inhibition: by letter name, by letter name at a specific location, or by stimulus name. In the inhibition is simply by letter name, the name of the first letter will inhibit the name of all other letters in any location of the stimulus, even if these letters are matches. For example, if S_1 and S_2 are "ABCD", the "A" of S_1 will inhibit the "B", "C" and "D" of S_2, but facilitate its "A". Thus, both matches and mismatches are inhibited by the three other letters, but matches are also facilitated by one letter. Because all letters are equally inhibited, this is equivalent to saying that there is no inhibition and that matches are facilitated. Hence, it is certain that this form of inhibition is not what Proctor intended.

If the name inhibition is limited to the location of the letter, then the "A" of S_1 would only inhibit (and facilitate?) the letter in the first position of S_2. In this situation, all matches are facilitated by one letter and all mismatches are inhibited by one letter. Let us compare the predicted RT of *L4D1* trials, where 3 letters are facilitated by the residual activation and one is inhibited, to *L4D3* trials, where these proportions are reversed. If encoding is done in serial, the average number of matches to be tested before finding the mismatch should be 1.5 for *L4D1* trials and 0.25 in *L4D3* trials (according to Equation 1). A look in Table 2, which contains the numerical values of Figure 1, shows that these 1.25 extra tests result in an extra 83 ms. Hence,

this explanation is not plausible; in order to be congruent with the slope of *D0* trials, the RT difference between *L4D1* and *L4D3* trials should be closer to 30 ms. If encoding is done in parallel instead, then the *L4D3* trials should also be faster, as the answer depends on a race between three mismatches instead of only one (a phenomenon called statistical facilitation, see Raab, 1962). Just like for serial encoding, however, it is not plausible that this facilitation would result in a RT difference of 83 ms. Thus, it is also unlikely that Proctor intended that inhibition was done depending on the location of the letter.

If the two types of letter inhibitions are not adequate, it means that the nominal inhibition Proctor was referring to is produced by the name of the stimulus, which is the last major problem of his theory. The Encoding Facilitation theory was set to explain the results from Bamber's second task, where the physical appearance of the letters was irrelevant to the sameness judgment (e.g., "ABCD", "abcd" and "ABCD" all matched). If the inhibition is done at the stimulus level, then the effect of inhibition on "ABCD" should be identical for "EFGH" and "efgh". Yet, Bamber found that the RT of "different" trials vary between uppercase and lowercase letters. Furthermore, the RT of "all-different" trials increased with *L*, which means that the inhibition seems to be dependent on the elements of the stimulus and not only the name of the global stimulus. Thus, there are no ways to make Proctor's inhibition idea fit with the data.

Like the previous models, this model inspired me despite of its shortcomings. The encoding facilitation idea makes sense and can elegantly explain at least a part of the fast-same effect. Also, other authors (Krueger, 1978) have reported that increasing the interstimulus interval between S_1 and S_2 reduces the accuracy of "same" trials, which is congruent with this idea of residual activation.

1.1.4 Response bias towards "same" answers

There was another theory to explain the advantage of "same" RT in one letter stimuli: a bias to answer "same". Ratcliff's Diffusion Model (1978) is a model of two-choice alternative tasks that accumulates evidence in a random-walk. He proposed that by making the random-walk start closer to the "same" threshold, his model implemented this bias and adequately described both the RT and accuracy distributions on the same-different task (for a few experiments). Ratcliff's model has many flaws, both in its conceptualization and its interpretation. First, fitting his model is computationally demanding, to the point that today's computers can take up to a day to fit a single participant. Back in 1978, when the model was elaborated, there was practically no one in the world who had the infrastructure required to test or verify his model, including himself. Thus, he manually picked values that provided a good fit for his free parameters. A second problem is the large number of free parameters, which can be much larger than the number of conditions fitted depending on the implementation. For example, Ratcliff (1985) had 23 parameters to fit 4 conditions and Gomez, Pera and Ratcliff (2007) had 21 parameters to fit 8 conditions. Not only does it make it easy to fit to any data, but it also leads to a third problem, which is to give a meaningful interpretation to these parameters. In other words, it is a model that is inaccessible, lacks parsimony and is so opaque that it does not increase the understanding of the phenomenon it is modelling (Proctor, 1986; Proctor & Rao, 1982; Proctor et al., 1984). Finally, both Farell (1985) and Sternberg (1998) made extensive reviews of the same-different task, and each considered Ratcliff's theory to be inadequate to explain how we do this task, even if it can provide good fits (for people with the resources to compute it). It is worth noting that although Ratcliff located this bias in the testing stage, it could also be implemented in the encoding stage, and thus his model shares a rationale with the Encoding Facilitation theory.

Raymond Nickerson, the first to identify and discuss the fast-same effect (1965), also argued in favor of a response bias towards "same" by invoking the seemingly innate preference for certain judgments, including for sameness. He mentioned a great deal of experiments that show how humans prefer and are typically faster to make positive judgments, such as "up", "top", "true" and "same" (Nickerson, 1967, 1978). Interestingly, these words are often said before their alternatives ("true or false", "same or different"). It does not seem far-fetched that our cognition would be built in such a way that these positive judgments are easier, faster and considered more important than their alternatives. This idea of preference for sameness is also supported by the priming literature: after being subliminally exposed to a stimulus, people tend to prefer stimuli that are identical or related to the first stimulus. Finally, William James' famous quote, which I chose as the first sentence of this thesis, shows that the father of American scientific psychology also believed that sameness was more important than change detection.

It is unfortunate that Ratcliff's model is good at fitting the data yet not viable to explain the phenomenon behind it, because Nickerson's explanation would give it a good theoretical support. Despite Ratcliff's claims, it is difficult to see how this bias could explain any other result than the faster L1D0 trials: this theory is silent on the slope. The response bias towards "same" answers did not have a lot of influence on the work I will present in the following sections and chapter. However, the debate on the origin of the fast-same effect (response bias, encoding facilitation or a better testing process) inspired me and helped me think through the results to come up with my own ideas.

1.1.5 Self-terminating "same" answers

Nickerson (1967) also made a remark on the design of the same-different task and on how it could lead to answering "same" in a non-exhaustive fashion, a comment that was echoed

by Hawkins (1969) and Farell (1985). Importantly, this phenomenon is not equivalent to the response bias towards "same", which assumes that "same" answers are exhaustive.

By definition, a mismatch is always indicative of the correct answer ("different"), but a match is not, because the "some-different" trials contain matches. This is not a problem by itself if the proportion of matching and mismatching letters is balanced. However, this proportion is not balanced in most experiments, starting with Bamber's. Counting the number of matches and mismatches in Table 1 reveals that there are 84 mismatches and 156 matches; of the 156 matches, 36 (about 25%) are seen in a "different" trial. This can be problematic in two different ways. The first way is if participants treat matches probabilistically, in which case they would sometimes count matches as evidence towards a "different" answer. This would lead to errors in "same" and "some-different" trials. The second way is if participants sometimes answer "same" without seeing all letters (i.e. if "same" answers can be given in a self-terminating fashion), in which case they would sometimes make mistakes in "some-different" trials.

The predictions of these two scenarios differ slightly, but they both predict a perfect accuracy for "all-different" trials, which probably explains why they were not formally implemented in any model or theory on the same-different task. Still, I will test whether self-termination can be responsible for "same" answers in Chapter 2 and will also provide an alternative to self-terminating or exhaustive processing in subsection 1.2.4.

1.1.6 Conclusion

I have used 5 different historic perspectives to explain the RT and accuracy data of the same-different task and have shown how each of them is either falsified or insufficient to account for the richness of the data. In the next section, I will detail said data while keeping in mind the strengths and weaknesses of the theories I just presented.

1.2 Data Review

In this section, I elaborate on the data that was used to argue against the theories in the literature review. Although the same-different task was replicated many times over the years, most studies had between 4 and 8 participants (two notable exceptions being Farell, 1977; and Ratcliff, 1985 with 16 participants each), and had no standards regarding the duration of presentation of S_1 and S_2, the interstimulus interval or the positioning of the stimuli on the screen. Fortunately, Denis Cousineau's laboratory has been collecting data on this task for the past 5 years, and many experiments are similar enough that their data can be agglomerated.

1.2.1 Filtering experiments, conditions, participants and trials

Combining these data together required that I filter out experiments, conditions, participants and trials. I found 8 experiments that I deemed comparable. The selection criteria were that they had to follow Bamber's design (shown in Table 1) and use the same 12 consonants ("B", "C", "D", "F", "J", "K", "L", "N", "S", "T", "V" and "Z"). Next, I filtered conditions out of the experiments such that they only contained trials that were comparable across experiments. For example, one experiment studied the impact of having S_1 and S_2 either in different colors or in the same color, while asking participants to make a same-different judgment on the letters. In this experiment, only the "same" trials where the color did not change and the "different" trials where the colors changed were kept. To filter out participants, I used very selective criteria on accuracy and RT. On tasks that are similar to Bamber's, the only article I know that reports an average accuracy lower than 95% is Ratcliff and Hacker (1981), in which they manipulated the participants' level of cautiousness. Hence, I considered safe to set the accuracy cut-off at 85%, which eliminated 5 participants out of 151. To set the RT cut-off, I first computed the mean RT of all 14 conditions per condition. Then, for each participant, I counted

the number of conditions where their mean was more than 2 standard deviations away from the condition's mean. Participants that were too fast or slow in at least 6 conditions and whose grand mean RT was also more than 2 standard deviations away from the experiment's grand mean were excluded. This second criterion identified 6 participants, 2 of which were already identified by the accuracy cut-off. In total, 9 participants out of 151 were filtered out. Finally, my selection criterion for trials was that RT had to be between 200 ms and 1000 ms because it is unlikely that answers given outside of these RT values were generated by the processes I am trying to study. To verify that the selection of trials did not significantly alter the results, I ran all analyses a second time using trials with a RT between 100 ms and 2000 ms, and although the patterns were slightly less clear, the general conclusions and ordering of conditions did not change.

I will now show and explain the results that our laboratory obtained. I have separated the results in three subsections: RT of correct answers, accuracy and RT of incorrect answers.

1.2.2 RT of correct answers

In Figure 1, I presented the classical RT results on the same-different task, which were obtained during this data review process. I now analyze this figure as I discuss the three factors that impact the RT, elaborate on their theoretical implications and indicate whether they support or invalidate the theories mentioned in the first section. The two first factors are the RT of *L1* trials and the impact of increasing *M*, which form the fast-same effect that was previously discussed (section 1.1.1.1). The third factor is the impact of increasing *D*, which received much less attention from the scientific community and gives a different perspective on the RT results. As I present the results, I will show how they invalidate certain cognitive model architectures.

1.2.2.1 The fast-same intercept.

The first factor is what I call the intercept part of the fast-same effect: single letter "same" trials are faster than single letter "different" trials (correct answers in *L1D0* trials are faster than in *L1D1* trials). For example, given a "same" trial where S_1 is "A" and S_2 is "A", and a "different" trial where S_1 is "B" and S_2 is "A", the "same" trial will be faster by about 50 ms. This RT advantage cannot be caused by the visual properties of S_2, as it is "A" in both trials. It is also unlikely that the movement to answer "same" benefits of some advantage over the movement to answer "different". Clearly, this RT advantage has to do with the way we process S_2 depending on its relationship to S_1, and not to S_2 *per se*.

Would this RT advantage still be present on a trial containing zero letters? If "same" trials were still faster in this theoretical (and impossible) scenario, this would indicate that there is an advantage for "same" answers that is independent of the number of letters being tested. On the contrary, if "same" and "different" trials were as fast, then the RT advantage of "same" trials would be explained by the slope (which will be discussed in the next subsubsection, 1.2.2.2). Table 3 puts these two scenarios mathematically.

Because we are not capable of measuring the RT of these imaginary scenarios, theories have been developed for both. Having an intercept advantage indicates some sort of facilitation that is not related to the decision-making process, as suggested by the encoding facilitation and response bias towards "same" answers. The Noisy Operator theory, on the other hand, posits that there is no intercept advantage and that the RT advantage for 1 letter stimuli resides in the slope.

The fast-same intercept by itself does not support any of the three theories in the above paragraph but invalidates the Identity Reporter (as previously mentioned). Both scenarios are compatible with the self-terminating "same" judgments because it is silent on that aspect.

1.2.2.2 The fast-same slope.

I now discuss the impact of increasing *M* on RT, the second part of the fast-same effect. According to Bamber's equations (Equation 1), the RT of *D1* conditions should increase by 0.5 tests for each increase in *M*, while the RT of *D0* conditions should increase by 1 test for each *M*. Yet the *D0* slope is smaller than the slope of *D1*, which has been interpreted as a "fast-same" effect. However, it is not obvious if we are dealing with a "fast-same" slope rather than a "slow-different" slope. The reason why the name "fast-same" was chosen is probably because if the fast-same intercept. Another plausible reason is that the RT of "different" trials can be modelled by our intuitions of a serial self-terminating process, while the "same" RT are more elusive. By assuming the intuitions on "different" trials are correct, this makes the "same" trials seem fast.

In Figure 1, the slope of *D0* trials is smaller than that of *D1*, about equal to that of *D2* trials, and greater than that of *D3* trials. In other words, the slope is only "fast" compared to *D1* trials. Furthermore, "same" trials are not always faster than "different". *L4D0* trials seem to be as fast as *L4D3* and *L4D4* trials, and the slopes hint that *L5D0* trials would be slower than *L5D3*, *L5D4* and *L5D5* trials (this is an extrapolation that requires proper testing).

To generate a fast-same slope, we would need a process that does fewer tests than there are matches, such that the "same" slope is less than the time needed to make a single test. A first way to achieve this is to test the letters in parallel, but, as discussed in section 1.1.1, Bamber showed that the only way a parallel exhaustive model can predict a linear RT slope is when assuming that testing a letter always takes the same time without variability, which is not plausible. However, another property of models was discovered after Bamber's time, the capacity, which allows for parallel exhaustive models to produce a linear RT slope. I will test and show how this explanation is not likely in subsubsection 2.6.2.4. A second way would be to

have self-terminating "same" answers (section 1.1.5). Assuming self-termination is equiprobable after testing any number of letters, the number of tests would increase by 0.5 tests instead of 1 for each M, making the slope theoretically equal to that of the $D1$ trials. This overly simplistic model comes close to the expected slope, and so it seems likely that a more developed early termination model could fit the "same" slope. As a counter-argument, I will later present a model that accounts for the RT without a bias towards "same" (subsection 1.2.4) and show that self-termination is not viable to explain the "same" RT (subsection 2.6.2.2).

There are also theories that generate a "slow-different" slope, although no articles mention it. In this scenario, the slope for "same" trials is exactly the time needed to make a single test, and the slopes of "different" trials are the ones that vary. This assumes that "same" judgments are serial exhaustive, the second most logical architecture given that parallel exhaustive models were ruled out. The first theory that assumes a "slow-different" slope is the Encoding Facilitation theory (section 1.1.3), which would simultaneously explain the fast-same intercept. However, as was discussed when presenting that theory, it would require a proper way of applying inhibition. A second theory that predicts a "slow-different" is the noisy operator theory thanks to its rechecking property (section 1.1.2), but this model would also need to be upgraded to fix its problems.

It should be clear by now that no single theory can account for the RT of correct answers. So far, whether the "same" slope is fast or the "different" slope is slow is still unknown.

1.2.2.3 Impact of mismatches.

I now present a compelling argument in favor of the "slow-different" slope through the impact of increasing D on RT. Figure 3 is a transposed version of Figure 1 that shows the RT slopes as functions of D (instead of M). These slopes are much harder to describe in words

because they are not linear and thus the effect of the absolute change of D on RT is not clear. For example, increasing D by 1 has a different effect from $D1$ to $D2$ versus from $D2$ to $D3$.

After studying this figure, I realized that the beneficial effect of D on RT increases with L. My first attempt to apply this observation was to describe the slopes as functions of the change in L/D. Let us compare the slope between $L2D1$ and $L2D2$ to the slope between $L4D2$ and $L4D4$. In both cases, the L/D proportion is divided by 2, and the slopes are similar. This was still not quite adequate, as can be seen by looking at the slope between $L4D1$ and $L4D2$, where the proportion is also divided by 2, but the slope is not equal to the other two. It seems that not only the change in proportion, but also the values of the proportions are important.

It is this reflection that brought my compelling argument in favor of the "slow-different" slope. Looking at the following equivalence for L/D, we see that changing M has a linear effect for a fixed D, but that changing D results in an asymptotic effect for a fixed M:

$$\frac{L}{D} = \frac{D + M}{D} = 1 + \frac{M}{D}. \qquad \text{Equation 2}$$

In a serial self-terminating process, the smaller the proportion of matches over mismatches is, the more likely it is that a mismatch will be seen first. Thus, as M gets closer to zero, the RT will approach its optimal value. This is exactly what we observe in Figure 4: the RT of "all-different" trials ($L1D1$, $L2D2$, $L3D3$ and $L4D4$) are very similar and the slopes seem asymptotic as D increases. In other words, the impact of D on RT does not seem to be relevant. Rather, changing D seems to affect RT by decreasing M, which I interpret as the matches having a negative impact on the production of a "different" response. As far as I know, no theory discusses this relationship between M and D or talks of a "slow-different" effect.

1.2.2.4 Discussion.

I consider that there is some credible support for both a "fast-same" and a "slow-different" slope. However, I believe that a "slow-different" slope is more likely, based on the following reasoning. If "same" judgments are exhaustive (and they should be most of the time), then the process must be serial (it cannot be parallel as per Bamber's proof). The important point here is that a serial process predicts that the slope will be directly proportional to the number of tests to be made. Hence, because the slope of "same" RT is the time needed to test a single match and because this slope is smaller than that of "some-different" RT, then the "some-different" judgments are slower than expected.

This brings me to propose the following interpretation of the RT data: testing matches makes it harder to reach a "different" answer. This can explain both the interaction effect mentioned above and the fact that the RT of "all-different" trials is independent of M and D. The idea that inhibition was involved in this task was suggested by Proctor in his Encoding Facilitation theory, and I was very quick to argue against it. My proposition differs from his both on the source and the target of this inhibition. In Proctor's model, it is the content of S_1 that causes the inhibition, and it does so by slowing down the encoding of anything that is not a part of S_1. My suggestion is instead that it is the matches in S_2 that cause the inhibition, and that they do so by making it harder to conclude that the stimuli are different (i.e. during the decision-making). On the other hand, I am convinced that Proctor's view on encoding facilitation is valid, and I think it explains the fast-same intercept in a simple and elegant way.

In conclusion, the RT data leads me to believe that matches benefit from encoding facilitation and that the more matches are present, the stronger they inhibit the "different" response. Let us now see if this view is congruent with the accuracy data.

1.2.3 Accuracy

The accuracy in the same-different task has not been studied extensively for two reasons: the accuracy is typically very high (as stated previously) and the number of trials per condition is often low. The only theory that discussed accuracy thoroughly was the Noisy Operator, although the remark on the potential of self-terminating "same" judgments also provides some intuitions on the source of errors. Thanks to the many experiments that were ran in our laboratory, I had access to one of the largest datasets of errors on this task, which is summarized in Figure 4.

Let us describe the accuracy results using M and D. First, looking only at "same" trials, we see that increasing M leads to a small linear decrease in accuracy. In "some-different" trials, however, the relationship between accuracy and M seems to be quadratic instead of linear. Finally, on "all-different" trials, the accuracy is increasing as M gets closer to 0: *L1D1* trials are significantly less accurate than *L2D2* and *L3D3* trials, which are significantly less accurate than *L4D4* trials.

These observations are almost identical to those I made on RT: increasing M leads to a linear decrease in performance for "same" and to an increasingly large decrease in performance for "some-different", and "all-different" trials are barely affected by M or D. It seems like the accuracy results can also be explained by an interaction between M and D, which is not surprising as any theory that aims to model the RT of a task should also be able to model the accuracy using the same parameters.

Some authors (e.g. Bamber, 1969) have reported that the quantity of errors for "same" trials was greater than expected considering that there is no misleading information in a "same" trial (only matches are present) versus "some-different" trials (where the match(es) could lead the participant to answer "same" incorrectly). I would argue that this remark is subjective and

was probably based on too little data points (similarly to the upward concave shape of RT that he observed). In Figure 4, we see that only two of the "some-different" conditions are less accurate than all "same" conditions (*L3D1* and *L4D1*), whereas two are about as accurate (*L2D1* and *L4D2*) and two are more accurate (*L3D2* and *L4D3*).

I will briefly present the qualitative predictions of the two theories that touched on accuracy, and then combine and integrate them with my proposition on the RT data.

1.2.3.1 Predictions from self-terminating "same" answers.

The idea behind having some self-terminating "same" answers is that matches are treated probabilistically, such that there is a chance they will trigger a "different" response. This means that the risk of answering "different" increases with the number of matches that were tested. There are many ways a participant could (unconsciously) determine the risks linked to answering after seeing a certain number of matches. It could be done experiment wide, meaning that participants have a rough idea of the proportion of matches that are in a "different" trial over the entire experiment. This does not seem realistic, especially for participants who have not done the task before. Another way would be to consider a match as uninformative, treating it as a mismatch 50% of the time. Again, it would make little sense to do so, and would result in a very low accuracy for "same" trials. I would argue that if participants treat matches probabilistically, it is most likely that it is by considering the number of untested letters in the trial, e.g., given that there are two letters left to test, how likely is it that there will be a mismatch?

This simple idea predicts a 100% accuracy both on "all-different" trials, as it does not discuss the way mismatches are treated, and on *L1* trials, as there is no ambiguity after testing a single letter. However, it correctly orders the accuracy of "same" and "some-different" trials. Intuitively, answering "same" in a self-terminating fashion sounds like a typical speed-accuracy

trade-off scenario, where making more tests before answering would increase the accuracy. Additionally, because the number of "some-different" trials increases with *L* (6, 8 and 9 trials out of 24 in the *L2*, *L3* and *L4* condition), answering "same" after testing a single match is much riskier in a *L4* trial than in a *L2* trial. Hence, self-terminating "same" judgments predict more errors in "same" trials at *M* increases (as is observed in the data). To mitigate this risk, participants could sometimes answer "different" when seeing a match, which is riskier than answering "same", but could still lead to a correct "different" answer with a faster RT. Depending on how it is implemented, this would predict more errors as *M* increases and a "some-different" slope that is smaller than predicted by Bamber's equations.

This theory has probably not been used in any model because it predicts a 100% accuracy on 5 conditions out of 14. Furthermore, it would be required to implement an early termination model to verify how well it can fit the data on the 9 other conditions, which is beyond the scope of this work. However, I will propose an alternative to self-termination in subsubsection 1.2.3.3 that will be tested in subsection 2.6.2.

1.2.3.2 Predictions from the Noisy Operator theory.

Krueger's Noisy Operator makes sense on a biological level and could theoretically produce the correct accuracy for both "same" and "different" trials, and yet its accuracy fits were not interesting (section 1.1.2). I will show why it is so with a simplified version of the model to predict the accuracy of *L4D1* trials.

First, let us assume there is no noise during the response-production stage. Then, the only way to make an error on a *L4D1* trial is to incorrectly test the mismatch and correctly encode and test all matches. The *L1D1* trials indicate that mismatches are incorrectly tested approximately 2.9% of the time and the *L1D0* trials show that matches are correctly encoded and tested about

97% of the time. Thus, because "same" judgments are exhaustive, the expected accuracy on a *L4D1* trial is $100\% - (2.9\% \times 97\%^3) = 97.35\%$, which is far from the observed 78.5%.

Second, let us add the response-production noise into the equation. The probability of noise leading to a "same" answer regardless of the stimulus cannot be larger than the error rate of the most accurate "different" trials, *L4D4*, which is 1.7%. Similarly, the probability of the response-production stage making a noisy "different" answer cannot be greater than 3%, the error rate of the most accurate "same" answer. Because the response-production stage cannot be responsible for all the errors (otherwise all "same" trials would have a 97% accuracy and "all-different" trials, 98.3%), let us assume that noise during the response-production stage will lead to a "same" answer 1% of the time, to a "different" answer 1% of the time, and to the answer given by the decision-making stage 98% of the time. Then, the accuracy of *L4D1* trials become $100\% - (1.9\% \times 98\%^3) = 98.21\%$, an even worse prediction than before.

Although this is a simplified version of Krueger's model, it reveals one of its weaknesses and shows that it cannot account for the accuracy of most conditions. However, there are some compelling reasons not to discard it, which I will now present.

1.2.3.3 Combining the two theories.

Clearly, both the self-terminating "same" answers and the Noisy Operator theories are inadequate to explain the accuracy of the "same-different" task. Unfortunately, they are the only two propositions that were given to account for accuracy. What has not been considered yet is if they could be combined, and what would the predictions of that combination be.

In their simplest forms, these two models assume that each letter is looked at once (at most) before making a decision (in the case of the Noisy Operator, the decision can be to recheck). However, nothing guarantees that this is what happens. Micro-saccades and top-down

processes (such as the reading effect) lead to unpredictable eye movements, which ultimately determine what will enter the brain and the decision-making process. The order in which letters are looked at and tested is likely not fixed nor exactly random, and some letters might be tested multiple times while others might not get tested at all.

By allowing letters to be tested multiple times, we create a new alternative to self-terminating and exhaustive processes. Suppose that an exhaustive process requires four matches to answer "same". A model that tests letters multiple times could also require four matches, yet reach this number by testing a single letter four times. If we allow for this variability in the Noisy Operator theory, it can make predictions that are similar to those of the self-terminating "same" answers theory.

Supposing that it has the threshold of an exhaustive process, a Noisy Operator that has a chance of never testing a letter would see its accuracy decrease in "some-different" trials as the proportion of matches over mismatches increases. This property is akin to what has been described in section 1.2.2.4 about how matches negatively impact "different" answers.

On "all-different" trials, we also obtain better qualitative predictions if we allow for this kind of process. Under the regular Noisy Operator, mistakes in an "all-different" trial require that all letters be incorrectly tested, for which the probability quickly tends towards 0. Assuming letters are incorrectly tested 1.2% of the time (an overly simplistic estimate obtained by subtracting the error rates of *L1D1* and *L4D4* trials), this should happen only on 0.0144% of the *L2D2* trials, and virtually never on *L3D3* and *L4D4* trials. If instead the model is equally likely to sample all letters, then the probability that it samples a letter twice in a *L2D2* trial is of 25%, which brings the error rate up to 0.3% instead of 0.0144%. While this is still far from the

observed 2.5%, it is still an improvement. Similarly, applying this reasoning to "same" trials gives qualitative predictions of accuracy that are close to those of the normal Noisy Operator.

1.2.3.4 Discussion.

In conclusion, the Noisy Operator model can be improved by integrating random scanning and testing of the stimuli's letters, making it more biologically sound and giving a model that is similar to what would be obtained if we tried to merge the two previously proposed sources of error together.

Although I did not provide any formal quantitative predictions of this type of model, the fact that it can generate errors on all 14 conditions of the same-different task is an improvement from all previous theories. Most importantly, these errors result from the same phenomenon invoked to explain the RT, which is the ratio of matches over mismatches.

1.2.4 A proto-model of the "same-different" task

In this section, I present a proto-model that integrates the observations and suggestions I made in subsections 1.2.2 and 1.2.3. Then, I use this model to make predictions on the speed of errors, which will be tested in the next subsection.

The proto-model makes three assumptions that are plausible but not readily testable: matches are faster to encode than mismatches (Encoding Facilitation theory), noise interferes during the encoding, testing and response-production stages as explained in subsection 1.1.2 (Noisy Operator theory) and there is no response bias towards "same", i.e. the noise in the response-production stage does not favour a "same" response.

With these assumptions in mind, the proto-model behaves in the following fashion. First, letters enter the brain in a serial random order dictated by top-down (reading effect) and bottom-up (saccades) effects. Second, the order in which letters enter the brain also dictates the order in

which they are encoded, and thus tested. Third, letters can be encoded (and tested) any number of times, including zero. Fourth, when the testing process identifies a match, it inhibits the "different" response. Fifth, "different" answers are self-terminating in the sense that only one mismatch is required to generate the answer. Finally, "same" answers have an exhaustive threshold, but only require that the correct number of matches be identified, and not that every letter in S_2 be a match.

What remains to be seen now is whether this model can predict the speed of errors. Let us begin with the "same" trials. Errors in these trials require that at least one match is encoded or tested incorrectly, which means that increasing L (and thus M) increases the risks of making at least one mistake. Because noise has a low probability of resulting in an error, it is very unlikely that more than two such errors would happen in a single trial. Hence, I am expecting the slope of these errors, which are "different" answers, to be between that of correct *D1* and *D2* trials, as the correctly tested matches should inhibit the error (leading to slower errors). Most importantly, the RT intercept of these errors should be equal to that of correct "same" answers. This is because all letters of a "same" trial are matches and thus benefit from encoding facilitation. In short, incorrect "different" and correct "same" answers on *L1D0* trials should have similar RTs, and the slope of errors should be comparable to that of correct answers. This is a prediction that no other theory makes.

For "different" trials, errors (incorrect "same" answers") happen if enough matches are tested before the first mismatch. Note that unlike matches, mismatches do not cause inhibition and thus race against the "same" answer. Because there is no way to slow down a "same" answer, the incorrect "same" answers on "different" trials should be as fast as correct "same" answers. This is not to say that there will be no slow errors on "different" trials. Noise during the

response-production stage should produce incorrect “same” and “different” answers in equal proportions. How noise affects this stage is not clear, but it could be argued that the confusion or hesitation stemming from conflictual messages (noise versus the decision-making stage) would lead to a slowed down response.

Let us now verify if these predictions of identical correct and incorrect answers on “same” trials and of fast incorrect answers on “different” trials hold.

1.2.5 RT of incorrect answers

There is little literature on the speed of errors (or RT of incorrect answers) for the same-different task. As previously mentioned, authors frequently reported accuracies of 95% or more. With such a small number of data points, it is almost impossible to define a RT distribution, let alone analyse it. Once again, the size of the dataset I used helped me to circumvent this issue.

In this subsection, I present the claims of Krueger and Nickerson regarding the speed of errors, then test the data to verify them. I conclude that the proto-model makes accurate predictions on the speed of errors.

1.2.5.1 Previous claims on the speed of errors.

In his article on the Noisy Operator, Krueger reports that there are more errors on “same” trials, that errors are slower than good answers in general, and that errors on “same” trials (incorrect “different” answers) are slower than errors on “different” trials (incorrect “same” answers). As discussed in the previous subsection, the belief that there are more errors on “same” trials than on “different” trials is subjective. Regarding errors being slower than correct answers, Krueger cited no sources and his own data did not show that errors were systematically slower ((Krueger, 1978, Tables 3 and 4). His results did suggest that incorrect “different” answers were

slower than incorrect "same" answers, but he did not clearly state whether this difference was significant.

Similarly to Krueger, Nickerson (1967) predicted that errors on "some-different" trials should get faster as the proportion of *M* increased when he commented on the possibility of self-terminating "same" answers. Once again, this prediction was not verified by data.

1.2.5.2 Testing the speed of errors.

To assess the speed of errors, I compared the RT of correct and incorrect "same" answers together and similarly for "different" answers. Figure 5 shows the RT of "same" answers at the top and of "different" answers at the bottom. Participants with a perfect accuracy in a condition were excluded from that condition, and the median RT of each participant was used instead of the mean. The error bars show the 95% Confidence Interval (CI) around the mean of medians.

For "same" answers, it is hard to draw any conclusions, mostly because of the very small number of incorrect "same" answers, which results in large CIs. A rough look at the figure suggests that there are only 2 conditions, *L1D1* and *L4D2*, that seem to generate incorrect "same" answers that are slower than correct "same" answers, which supports my predictions. Still, I do not consider these results to be safe to use.

For "different" answers, however, the RT of incorrect answers (in blue) are more reliable and are very close to what I predicted. The slope of the errors is 19.44 ms, which is statistically indistinguishable from the slope of *D2* trials (20.03 ms) and smaller than that of *D1* trials (29.84 ms). The RT of *L1D1* is also as predicted, with the RT of errors (465.45 ms) being equal to the RT of correct answers (464.80 ms). Given that no other theory makes this prediction, it seems likely that the phenomena I invoke in the proto-model are involved in the same-different task.

1.3 Conclusion

In this first chapter, I reviewed the same-different literature and showed how each previous theory can only cover some aspects of the data. I then described the typical RT and accuracy data on this task and provided intuitions on how to explain both. Next, I presented a proto-model based on these intuitions and used it to make predictions on the speed of errors. Finally, I verified the speed of errors and confirmed that the proto-model is a valid candidate to explain how we do the same-different task.

The main problem with this proto-model is that it can only make qualitative predictions. Without a formal implementation, this idea will remain as it is. Fortunately, there are experiments that can be made to test whether it is worth investing the time in developing such a model, which is the subject of my second chapter.

2 Applying Delayed Presentation to the Same-Different Task

In this chapter, I put the proto-model proposed in the first chapter to the test. To do so, I will apply the Delayed Presentation protocol (section 2.1) to the same-different task. I then explain how this protocol can help diagnose the type of model involved in the task (subsection 2.2.1), and present the redundancy effect and how detecting it can (in)validate some assumptions of the proto-model (subsection 2.2.3). This allows me to formulate hypotheses in support of the proto-model (section 2.3) and to create an experimental design to test them (section 2.4). Finally, I report the data obtained from said design (section 2.5) and show how they warrant that additional resources be invested in developing a formal model based on the proto-model (section 2.6).

2.1 Delayed Presentation Protocol

The Delayed Presentation protocol was proposed to me by Denis Cousineau as a method to study whether participants base their answer on the detection of matches or of mismatches when doing the same-different task. In this protocol, participants are briefly shown a part of S_2, called the probe (*P*), before seeing S_2 entirely (see Figure 6 for the detailed timing and steps of a single trial). By manipulating the content of the probe, this method can, among other things, detect whether an answer is given in a self-terminating or exhaustive fashion (subsection 2.2.1).

I picked the duration of the probe to be 23.53 ms (rounded to 24 ms in this text) based on informal discussions with my supervisor and colleagues and on our monitors' refresh rate of 85 Hz (2 frames at 85 Hz last 23.53 ms). I then ran a pilot study to test whether this duration was small enough to not be noticeable by the participants and large enough to have an effect. This delay went entirely unnoticed by most participants (a handful of participants thought their eyes had slight troubles at times, but did not report feeling distracted by it), which meant it fit my first

criterion. To check whether it had the expected effect, I compared the RT of trials where the probe contained nothing to those where the probe contained S_2. Normally, this should be equivalent to comparing trials where S_2 is shown either 400 ms or 424 ms after S_1, and in which the timer starts after 400 ms; the RT of trials that are shown later should be 24 ms slower. This is indeed what was observed, which ruled out the possibility that having different probes in a single experiment could have an unforeseen effect on RT. Hence, it seems like 24 ms is an appropriate duration for the probe.

The pilot did not provide any other interesting results given its lack of statistical power. This problem stems from the fact that Bamber's design has very few "some-different" trials, which are critical trials for my purpose. There was one encouraging result however: the *L2D1* condition, the "some-different" condition with the most trials, behaved as the proto-model expected regarding both the RT and the accuracy. This made me confident that the Delayed Presentation protocol was appropriate, and I created a new design with a larger number of "some-different" trials.

Before presenting this design (section 2.4), I will show how the Delayed Presentation protocol can identify the architecture of a mental process and put the proto-model to the test.

2.2 Testing the Effects of the Probe

In this section, I show how changing the probe contents can affect performance on the same-different task. I first give the interpretations obtained from mean RT changes depending on the content of the probe (subsection 2.2.1). I then briefly touch on the importance of making interpretations based on the RT distributions of correct and incorrect answers as well as the accuracies rather than simply using the mean (subsection 2.2.2). Finally, I present the redundancy effect, which detects if seeing more than one critical attribute simultaneously leads

to correct RTs that are better than predicted by a race model (i.e. a parallel self-terminating process), and detail the usage of two tests that take advantage of it (subsection 2.2.3).

To keep the text simple and to be coherent with the experiment detailed in section 2.4, the only trials that will be considered are *L4D0* for "same", *L4D2* for "some-different" and *L4D4* for "all-different", making the "*L4*" part irrelevant. Trials will now be referred to simply by their number of letters and the type of probe. The content of the probe is denoted as follows: P^* means that it contains all letters of S_2, P^M means it contains two matches, P^D, two mismatches, and $P^{\pm}$, a match and a mismatch (this is illustrated in Figure 6). Hence, $D2P^*$ refers to "some-different" trials with 4 letters and 2 mismatches in which the probe contains all 4 letters, $D4P^D$ refers to trials with 4 letters and 4 mismatches in which the probe contains 2 of the mismatches, etc.

2.2.1 Example of interpretation of mean RT differences

Presenting parts of S_2 earlier can impact the participants' performance, which then helps diagnose how they processed that information. In this subsection, I interpret the simple differences between the mean RT of conditions to draw conclusions on the architecture of the mental process that gave a "same" answer. The goal of this subsection is to provide the reader with intuitions on how the Delayed Presentation protocol can be used in our context, and not to elaborate on every possible diagnostic.

There are two possible probe contents in "same" trials: either it contains all the letters of S_2 ($D0P^*$) or only two matching letters ($D0P^M$). The P^* trials are the reference trials, being identical to a typical trial in a same-different task. If the $D0P^M$ trials are faster than the $D0P^*$ trials, the decision is taken before seeing all four letters, and the "same" answer is given by a self-terminating process. Alternatively, it is also possible that the process is overwhelmed when it receives too much information simultaneously, a characteristic called limited capacity

(Townsend & Nozawa, 1995). In this case, both self-terminating and exhaustive processes would result in faster *D0P^{M}* trials. There are only two letters to process during the first 24 ms, which means they are processed faster than in *D0P** trials. The self-terminating process can then answer right away, and the exhaustive process has already tested one or two letters when the rest of the stimulus appears, keeping the workload lower through the trial.

If the RT of *D0P** and *D0P^{M}* trials are equal, the answer could be given by a (serial or parallel) self-terminating process with an unlimited capacity, i.e., it is not speeded or slowed down by the quantity of information it receives (Townsend & Nozawa, 1995). If the process is exhaustive, then it must be serial with an unlimited capacity and testing the two matches must take more than 24 ms. In this scenario, the delay plays absolutely no role. An unlikely alternative process would be that of a limited capacity process that happens to slow down enough to compensate precisely for the arbitrarily chosen delay of 24 ms. The explanation from the proto-model would be that the process will do all its tests on the two matches that are available, such that the presence of two or four matches does not affect RT.

At last, if the *D0P^{M}* trials are slower, the interpretation depends on the size of the difference. A difference of less than 24 ms means that testing the first two matches takes less than 24 ms and that the process does not test the matches twice. A difference of more than 24 ms could be explained by an attention capture effect, which would result in a slow-down when the last two matches are presented in the *D0P^{M}* trials. This is equivalent to having the encoding and testing steps sharing their resources (limited capacity), Alternatively, it could mean that "same" answers are generated by a super-capacity process, i.e., the process gets faster when more information is available (Townsend & Nozawa, 1995), which would make the *D0P** trials faster. However, this is not likely given the data in Figure 1: if "same" answers depended on a super-

capacity process, their RT would get faster with *L* instead of slowing down. Finally, if the *D0P*^{M} trials are exactly 24 ms slower than the *D0P** trials, it would indicate that the process is parallel and exhaustive. Remember that I argued against this type of process (subsubsection 1.1.1.2) because the slope of the "same" RT would need to have a downward concave shape. However, if the process has limited capacity, the RT would also be slowed down by increasing *L*. The resulting combination of these predicted upward and downward concave slopes could be the observed straight line. Although this explanation seems farfetched, it is the only one that would predict a slowdown of exactly 24 ms.

2.2.2 Using RT distributions and accuracy

The interpretations of RT differences above only take into consideration the mean RT of correct answers, as stated at the start of this section. To understand properly how we do the same-different task, we need to use the full array of information generated by participants, including the RT distributions of both correct and incorrect answers and the accuracies. It would be extremely tedious to list all possible combinations of results and their interpretations as I did for the mean RT and doing so would not contribute to improving the understanding or intuitions of the reader. Still, I want to point out how this additional data is crucial to correctly draw conclusions on the architecture of the process.

For example, I mentioned that if *D0P*^{M} trials are exactly 24 ms slower than the *D0P**, we should investigate the possibility of a parallel exhaustive process. While the difference in mean RT could be 24 ms, it is possible that the PDF of the two conditions differ in shape such that the mean difference is irrelevant (which will turn out to be the case). Furthermore, a parallel exhaustive process would predict a lower accuracy for the *D0P** trials, because there are more

matches present and thus more chances for noise to intervene. We will instead see later that it is the $D0P^M$ trials that are less accurate.

This brief argument shows how crucial it is to take all the data into account before dismissing a result with a simplistic explanation. Nevertheless, it should now be clear that the Delayed Presentation protocol is adequate to test the proto-model's assumptions.

2.2.3 The redundancy effect

Suppose a target-detection task where participants must decide as quickly as possible if the attributes "A" or "B" are present (for example, the color red or a square shape). This task is a typical self-terminating task, where an answer can be given as soon as one of the attributes is encoded and tested. These tasks have been conceptualized as a race between two processes, such that the faster of the two is the one generating the answer.

The redundancy effect is observed when presenting the two attributes simultaneously leads to better RT than predicted by a race model. Intuitively, if it takes exactly 100 ms to detect red and exactly 120 ms to detect a square, presenting the two attributes together should always take 100 ms. If the detection times vary, say 100 ±10 ms and 120 ±40 ms, it then becomes possible that the square will be detected faster than the red. This phenomenon is the "statistical facilitation" I referred to earlier. This statistical facilitation has its limits, however, and if there is more increase in performance than what would be predicted by statistical facilitation, we suppose there is an interaction between the processes detecting the attributes, which was coined the "redundancy effect".

At a first glance, this redundancy effect is of limited interest for the same-different task. First, we already know that we are not in a race-like situation on "same" trials. As can be seen in Figure 1, the "same" RT increases with L, which does not fit the predictions of a self-terminating

process (race) or of the redundancy effect. Second, although "different" trials seem to be self-terminating upon detecting a mismatch, the same-different task differs from a typical target-detection task because participants do not have a list of targets to look for in advance. The attribute "mismatch" is a relational attribute that is not a property of the stimuli *per se*, but of the relationship between two stimuli. Thus, these trials are not exactly comparable to those in a visual search task.

This does not mean that studying the presence of the redundancy effect in "different" trials is irrelevant. In fact, the presence of the redundancy effect is more informative in the same-different task concerning the locus of the gain in speed. In a visual search, the increased performance can come from the encoding stage if detecting an attribute facilitates the detection of the second, e.g., if attributes "A" and "B" are "kitten" and "cat", or if the attributes share the same physical location, e.g., when showing a red square. It can also happen during the decision-making stage; because there are two sources of evidence towards the "target-present" decision, the threshold can be reached faster. In the same-different task, however, the mismatches do not share properties during the encoding stage, as previously stated, and their physical location is randomized and controlled. Hence, if there is a redundancy effect in the same-different task, it is located in the testing stage.

2.2.3.1 The Cumulative Distribution Function test.

I now introduce the first formal test for the redundancy effect, developed by Miller (1982) and known under many names: the Miller Inequality, the race-model inequality or the cumulative distribution functions (CDF) test. This last name will be used in the rest of this paper. The CDF test has been used to rule-out the usage of race-models in many cognitive processes (Ulrich & Miller, 1997, p. 367).

Let us describe the relationship between the CDFs of the RT of correct answers on trials where only one of the attributes are present, $F_A(t)$ and $F_B(t)$, and the CDF of the RT of correct answers on trials where both attributes are present, $F_{AB}(t)$. If there is no overlap between $F_A(t)$ and $F_B(t)$ (i.e., if one reaches 1 while the other is still at 0), then $F_{AB}(t) = \min\big(F_A(t), F_B(t)\big)$. If there is at least some overlap between the two CDFs, then we expect to observe statistical facilitation. However, if the sum of $F_A(t)$ and $F_B(t)$ is below $F_{AB}(t)$, then we are not in the presence of simple statistical facilitation, but of the redundancy effect instead:

$$F_{AB}(t) \leq F_A(t) + F_B(t)$$
$$0 \leq F_A(t) + F_B(t) - F_{AB}(t) \qquad \text{Equation 3}$$

where $F(t)$ is the CDF value after t ms elapsed, and the indices A and B indicate which critical attribute(s) is/are present. A simple way to understand this CDF test is that if the fastest observed RT for "A" and "B" trials is T ms, then an "AB" trial should not result in RTs faster than T, or $F_{AB}(T - 1\ ms) = 0$.

2.2.3.2 The Stimulus Onset Asynchrony test.

The attentive reader might remember that Figure 3 shows an asymptotic improvement of RT as D increases, which led me to argue that increasing D does not improve RT, but rather that increasing M worsens it (subsubsection 1.2.2.3). Why then bother with a test to determine whether the gains from adding mismatches are greater than those expected by a race?

Almost 20 years after the CDF test, Ulrich and Miller presented a complementary test based on stimulus onset asynchrony (SOA, 1997). They developed this new test because the CDF test did not have enough discrimination power: regardless of if data always passes, always violates or occasionally violates the CDF test, there are two possible model candidates to choose from. Figure 7 shows these six models and how using the CDF and SOA tests together can

discriminate them. In this figure, we see that the race-model (with unlimited capacity) and the limited capacity model (which is also a race model) both always pass the CDF test, but can be told apart by applying the SOA test. This SOA test uses a logic similar to the one applied in the CDF test, but requires a protocol where there is a delay between the presentations of the two critical features. The Delayed Presentation protocol is hence not only useful by itself (as seen in section 2.2.1) but it can also leverage this new test to determine the capacity of the process responsible for the "different" answers.

The SOA test focuses on the presentation delay d between the two critical features, where a positive d means that "A" is presented d ms before "B". Each feature has a "detection time" distribution, T_A and T_B, and the trial also has a "motor time" distribution, M. Although Ulrich and Miller consider the encoding and testing stages to fall under the "detection" part of the equation, it is also possible to consider T_A and T_B as the distributions of the time needed for testing, and M as the distribution of the time needed for encoding and response-production.

There are only two assumptions for the SOA test, grouped under the term "SOA independence". The first is that the expected value of M must be independent of d, which Miller and Ulrich claim is likely (Mordkoff, Miller, & Roch, 1996). The second is that the joint CDF of T_A and T_B must also be independent of d. Because the testing time of "A" (or "B") should not depend on whether it was presented before or after "B" (or "A") or on the time elapsed between the presentation of the two letters (within reason), the joint CDF should also not be affected by these factors. If applying the SOA test to the data gives invalid results, it probably means that the SOA independence was not respected, which is also a diagnostic on its own. Note that there are no constraints on the distributions T_A, T_B or M, as long as they have a computable CDF.

If this assumption holds, then the SOA test predicts that the expected RT of the race model is a function of d that will vary depending on the capacity of the process (see Figure 8):

$$E[RT(d)] = \begin{cases} E[\min(T_A, T_B + d)] + E[M], when\ d \geq 0 \\ E[\min(T_A - d, T_B)] + E[M], when\ d < 0' \end{cases} \quad \text{Equation 4}$$

where $E[X]$ is the expected value of expression X, $min(X)$ is the minimum value of expression X, T is the distribution of detection times of an attribute, M is the distribution of the time taken by the encoding and response-production stages, the A and B indices refer to their respective critical attributes and d is the delay between the presentation of attributes A and B, such that a positive d means that A is presented d ms before B.

2.2.3.3 Methodological considerations.

There are methodological details to consider before applying these tests to the same-different task in the Delayed Presentation protocol. Specifically, one must think of the way to obtain the three required types of trials: those where only "A" is present, where only "B" is present and where both "A" and "B" are present.

As previously mentioned, being a mismatch is not an intrinsic attribute. Thus, unlike in the visual search paradigm, the distributions T_A, T_B and their CDFs cannot be generated according to attributes of the letters. An alternative method would be to use the location of the mismatch as its attribute. A first solution would be to compare *L2D1* to *L2D2* trials. If x is a match, we can control its position such that trials Ax and xB are compared to AB. Unfortunately, it is not possible to have trials with a $P^{\pm}$ probe with *L2* trials. Comparing *L3D1* to *L3D2* or *L4D1* to *L4D2* allows the use of $P^{\pm}$ probes but poses the problem that there are many possible *D2* conditions. Consequently, each stimulus would be tested more than once: the trials Axx and xBx should both be compared to ABx, the trials Axx and xxB should be compared to AxB, etc. Not

only does this result in complicated interactions, but it would also require an extremely large number of trials.

The best alternative then is to compare *L4D2* to *L4D4* trials, which is why this chapter only bothered with *L4* trials. There exist 6 types of *L4D2* trials which form 3 complementary pairs: AAxx and xxBB; AxAx and xBxB; and AxxA and xBBx. Any member of these pairs can be considered a condition where only one attribute is present, and the *L4D4* trials make the condition where both attributes are present. Unfortunately, having enough trials to obtain a robust CDF of each type of trial for each participant is not realistic. At first, this problem seems to be solved by simply grouping three types of trials together and comparing the two groups of three. Doing so is not adequate however, because choosing any three stimuli always results in an unequal distribution of the location of the mismatches. Thus, comparing the pairs would also include an effect of the location of the mismatches.

To work around this issue, I propose to run both the CDF and SOA tests by using every possible groups of three types of trials and their complement, thus controlling for both the type of trial and the location of the mismatches (see Appendix A for more details). Although computationally intensive, this solution is also the most exact given the context.

2.3 Research Objectives and Hypotheses

I have presented three tools that can be applied on the data gathered from the Delayed Presentation protocol: comparing the RT distributions and accuracy between conditions, the CDF test and the SOA test. Together, these tools can test all assumptions of the proto-model developed in Chapter 1. My objective is now to develop an experiment that would verify the assumptions of the proto-model. I first define my theoretical framework, that is the assumptions

that will not be tested. Next, I list every assumption that the experiment should test and how to invalidate them.

2.3.1 Theoretical framework

There are three assumptions that will not be tested by the experiment. The first is that the encoding of matches is faster than the encoding of mismatches, as proposed by Proctor (see subsection 1.1.3). This is the simplest explanation for the intercept of the "fast-same" effect and makes sense on a physiological and cognitive point of view.

The second is that noise can intervene during the encoding, decision-making and response-production stages and does so following Krueger's intuitions. This assumption is a clever approach to explaining how errors can be generated in "same" trials and allows the model to generate fast "different" answers. Most importantly, it predicts that matches are more sensitive to noise than mismatches.

Finally, the probability that noise during the response-production stage will lead to an error is independent of the trial or of the decision-making process. In other words, the probability of pressing the wrong key on the keyboard is not biased towards one answer.

2.3.2 Serial random testing process

The first hypothesis is that there is a process that tests encoded letters in a serial, random order, and that it can test them any number of times, including zero.

A simple way of rejecting this assumption is if there is evidence of parallel processing, as it would imply that the order is not random (there is no order) and that all letters are tested an equal number of times. For "same" answers, the process involved in testing matches probably does so in parallel if the RT of $D0P^{M}$ trials are slower than the $D0P^{*}$ trials by exactly 24 ms.

Regarding "different" answers, we can verify if the RTs of "all-different" answers are generated by a parallel self-terminating process via the CDF and SOA tests.

2.3.3 One testing process for both answers

The second assumption is that there is only one process responsible for the testing stage, which generates the evidence for both "same" and "different" responses. Whether this assumption holds depends on if both responses can be generated by a serial or by a parallel process. If they are, the ideal theory would assume that this testing process is unique.

An interesting corollary of this hypothesis is that the capacity of the testing process should affect both "same" and "different" answers. If the type of capacity involved in "different" answers does not account for the "same" data, then the hypothesis of a single testing process must be rejected.

It is unknown whether "same" and "different" answers are given by a single process or two processes (an issue that was central to the same-different literature), and it is beyond the scope of this project to solve this question. Similarly, regardless of whether there is a single testing process, it would be difficult to determine if that process also takes care of the encoding and/or response-production stages.

2.3.4 Self-terminating "different" answers

This hypothesis states that testing a single mismatch is sufficient to generate a "different" response. This idea was not strongly debated in past research, but the Delayed Presentation protocol provides a good opportunity to test it.

If the $D4P^{D}$ trials are faster than the $D4P^{*}$ trials, it is possible that more than one mismatch must be found before answering, rejecting the hypothesis of self-termination.

However, it could also mean that the process has limited capacity, which could be tested by using the CDF and SOA tests.

2.3.5 Fixed threshold for "same" answers

This hypothesis states that the requirement to answer "same" does not vary drastically across trials. More formally, the distribution of threshold values for a "same" answer should not vary for a given *L*. Typically, the quantity of information that an exhaustive process requires to generate its answer is assumed to be defined by the task. For example, it would be set at 1 (arbitrary) unit of evidence towards "same" for each letter in S_1. Because *L* varies between trials in Bamber's design, it is likely that, on some trials, the threshold would be set incorrectly, e.g., 2 units for 3 letters, or be more variable, e.g., 1 ± 0.2 units of evidence per letter. With a fixed *L*, as will be the case in my experiment, there should be much less variability.

This means that RT of "same" answers should not depend on whether it was given in a "same" or a "different" trial, i.e. regardless of if it is a correct or incorrect answer. Hence, the errors on $D2P^M$ trials should be as fast as the correct answers made on $D0P^M$ trials. Furthermore, if the capacity of the testing process is unlimited, then the correct answers on $D0P^*$ trials should also be as fast as the errors on $D2P^M$ trials. Obtaining this result would support that the "same" threshold is fixed, but also that letters can be tested multiple times and that there is only one testing process (because the capacity of "different" answers can be applied to "same" answers).

2.3.6 Inhibition of the "different" answer by matches

This assumption is the core of the proto-model and states that testing a match will make it harder for a "different" answer to be given. As stated in subsubsection 1.2.2.4, this inhibition is caused by the matches in the probe or in S_2 (not in S_1, like in the Encoding Facilitation theory) and affects the testing stage.

There are two ways to verify this hypothesis. The first is to compare the RT of $D2P^{D}$ and $D4P^{D}$ trials. If there is no inhibition by matches and assuming that "different" answers are self-terminating (which will be verified beforehand), these trials should have equal RTs. If $D2P^{D}$ trials take longer, then matches are probably causing some inhibition. The second way is to compare $D2P^{M}$ and *D4* trials. If inhibition is at work, then $D2P^{M}$ trials should be more than 24 ms slower than *D4* trials.

2.4 Methodology

Here I detail the experiment I used to verify my hypotheses, which implements the Delayed Presentation protocol. As was stated earlier, only *L4* stimuli were shown, with *D2* being the only "some-different" trials. This had two advantages over using Bamber's design, the first being that the lowest number of trials per cell was now 72 (Table 4). The second advantage was that I could control for the specific location of the matches and mismatches in the probe and S_2, which would have required an astronomical number of trials in Bamber's design.

2.4.1 Participants

20 participants were recruited by placing posters in the University of Ottawa's main campus and through word of mouth. They were paid 8$ for their participation. One of the participants did not complete the task and another had an accuracy of 73%. Hence, two new participants were recruited to replace them. After further analyses, the data of 19 participants out of 22 was used, using the same exclusion criteria as in the data review (section 1.2).

2.4.2 Apparatus

The stimuli were displayed on a *Sony CPD-G420S* CRT monitor with a diagonal of 45.72 cm (4:3 screen ratio) running at a resolution of 1024 by 768 pixels and a refresh rate of 85 Hz. The colors had a 32-bit depth and color accuracy was calibrated with an *X-Rite Monaco Optix*

XR colorimeter using the *DisplayCaL3* software and the *Argyll CMS* drivers. Participants were sitting on a chair with arm rests and their heads were approximately 60 cm from the monitor. They used a Dell keyboard *KB216t* to provide their answer.

2.4.3 Protocol

This experiment follows the Delayed Presentation protocol described in section 2.1 and represented in Figure 6. The pilot experiment validated that the 23.53 ms delay was both undetectable and efficient and thus this delay was used. Participants completed a total of 1152 trials (Table 4). The duration of a single trial is 1906 ms plus the RT, for a conservative estimation of 2.5 seconds per trial and a total duration of 48 minutes.

Participants were prompted to take a pause of the duration of their choice after completing 25%, 50% and 75% of the trials. They also received a warning whenever they made five errors in a row, or answered in less than 200 ms or in more than 1000 ms three times in a row. They could use this warning as a break of the duration of their choice. One participant received three such warnings for errors, but still maintained a high accuracy. Three other participants received a single warning concerning the speed of their answers.

2.4.4 Stimuli

The stimuli were composed of the consonants listed in Bamber's articles ("B", "C", "D", "F", "J", "K", "L", "N", "S", "T", "V" and "Z") and were typed in the Courier New font, size 18, in bold. Each letter occupied 0.5 cm by 0.5 cm on the screen, and there was a 0.2 cm vertical spacing between each row of letters. The letters of S_1 covered the vertical space from 0.95 cm to 0.45 cm above the centre of the fixation point, whereas the probe and S_2 covered the space from 0.45 cm to 0.95 cm below the fixation point. Their horizontal position was centred such that the letter positions were fixed regardless of the number of letters displayed (i.e., the letters in the

probe did not shift when the rest of S_2 appeared). Hence, each letter subtended a 0.477° by 0.477° visual angle, and the surface containing both S_1 and S_2 subtended a 1.814° vertical visual angle and a 1.910° horizontal visual angle.

As previously mentioned, the location of the mismatches in S_2, the location of the letters in the probe and the content of the probe were balanced. This is not difficult to obtain in *D0* or *D4* trials, because they only contain matches or mismatches. In *D2* trials however, 42 conditions must be balanced. Let us see how that translates in one of the six probe configurations, where the probe is the first two letters of S_2. To correctly balance all three probe conditions, the probe must contain two matches ($D2P^{M}$) on four trials, two mismatches on four trials ($D2P^{D}$), and a match and a mismatch on four trials ($D2P^{\pm}$). Because the position of matches and mismatches must be balanced in the $D2P^{\pm}$ trials, the four trials must be one of each of the following (where x are matches and A and B are mismatches): AxBx, AxxB, xABx and xAxB.

2.5 Results

In this section I report the data on RT, accuracy and speed of errors, and make simple comparisons between conditions. The discussion (section 2.6) will interpret these results and present additional analyses required to clarify them.

2.5.1 RT of correct answers

Figure 9 shows the mean RT and 95% CI of all 8 conditions. In the "same" conditions, the $D0P^{M}$ trials are 26.9 ms slower than the *D0P** trials. There is no difference in mean RT between the two "all-different" conditions. As for the "some-different", the *D2P** and $D2P^{D}$ trials have the same mean RT and are significantly faster than the $D2P^{M}$ trials by 11.8 and 13.3 ms, respectively. The $D2P^{\pm}$ trials, on the other hand, are not statistically different from the other probe conditions.

2.5.1.1 Comparing distributions of RT.

As discussed before, using solely the mean RT to infer the architecture of a cognitive process is overly simplistic. Comparing the RT distribution is more informative, but is more complicated. One approach to do so is to create a group distribution for each condition and then compare these distributions. A few methods have been proposed to improve on the naïve approach of treating all data as if it came from a single participant, and some fare much better than others (Cousineau, Thivierge, Harding, & Lacouture, 2016). Still, these methods rely on estimations. A second and simpler approach would be to apply the reasoning of repeated-measures to comparing distributions, by studying the difference between conditions for each participant. Unfortunately, because we are concerned with the RT of *correct* trials, the varying accuracy makes it rare that there will be an equal number of trials in two conditions.

To circumvent these problems, we can instead estimate participants' CDFs for each condition by using the RT percentiles. Because it relies on observed probabilities, this approach does not require any estimation and is not affected by the unequal number of correct trials between conditions. Furthermore, it has the advantage of comparing distributions while controlling for each participant's encoding and response-production time. In other words, the difference between quartiles reflects differences in decision-making time.

I opted to use the 5^{th}, 20^{th}, 35^{th}, 50^{th}, 65^{th}, 80^{th} and 95^{th} percentiles, which gives 6 intervals comprised of about 9 measures each in the conditions with the smallest number of correct trials. I chose to ignore the first and last 5% of all trials because the minimum and maximum RT were already arbitrarily constrained to be between 200 and 1000 ms by the experiment. To consider the lower and higher percentiles as the "true" fastest and slowest RTs would thus have been biased. To visualize how the CDFs vary between conditions, I subtracted the value of each

participant's percentiles across conditions, and then averaged the differences across participants. To obtain the PDFs, I subtracted each consecutive percentile within a condition for each participant (the equivalent of taking the derivative of a continuous distribution), then subtracted and combined them as I did for the CDFs. Figure 10 shows the differences between the CDFs and PDFs of *D0P** and $D0P^{M}$, *D2P** and $D2P^{M}$, *D2P** and $D2P^{D}$, and *D4P** and $D4P^{D}$ trials. I do not show the comparisons with the $D2P^{\pm}$ trials because they are very similar to the *D2P** trials.

The differences plotted on Figure 10 take the non-*P** condition's curves and subtract the *P** condition's curve. To interpret the CDF and PDF differences, let us imagine what would be the difference between the curves of trials with an empty probe (which are not in this experiment) and the curves of trials with a *P** probe. For the CDF, because the probe is empty, no actions can be taken until the rest of the stimulus appears. Hence, the difference between the CDFs should be exactly 24 ms (i.e., the RT of the empty probe is 24 ms longer). A value below 24 ms means that the non-*P** trials used the content of the probe or that the *P** trials have a disadvantage, while a value above 24 ms means that the non-*P** trials cannot make efficient use of the probe or that the *P** have an advantage. Regarding the PDF difference, the trials with an empty probe should have an identical PDF to the trials with a full probe, being simply shifted by 24 ms. However, if the x-axis is transformed into percentiles, the two PDFs should superpose perfectly and subtracting them should give 0 for all percentiles. In other words, if the PDF difference is not zero, the two PDFs do not have the same shape, being more stretched in one direction or the other in the percentile interval. Table 5 summarizes the interpretation of various values of the difference between the CDF and PDF. I recommend that the reader keep both Figure 10 and Table 5 on hand while reading the discussion section.

I now present the results from the comparison of RT distributions. For the "same" trials, the CDF of the *D0P** shows that $D0P^{M}$ trials have a few answers that are quicker than expected, but that the majority of answers are slower by more than 24 ms. The PDF difference shows that the $D0P^{M}$ RT has a longer right tail. For the "all-different" trials, the CDF different shows that $D4P^{D}$ trials are very fast, yet slightly slower than what is expected from a self-terminating process before the 65th percentile. On the other hand, the slower answers on $D4P^{D}$ trials are faster than those on *D4P** trials, and the PDF difference shows that the *D4P** are almost uniformly shifted to the right. Next, the $D2P^{M}$ are always slower than *D2P** trials, as can be seen from the CDF difference, but by less than 24 ms. The PDF difference of these conditions shows that they have very similar PDFs, excepted that the *D2P** PDF is more compressed to the left between the 50th and 80th percentiles. Finally, the CDF difference between *D2P** and $D2P^{D}$ trials remains below zero up to the 65th percentile, at which point the *D2P** seem to catch up and bring the difference close to 24 ms at the 95th percentile. The PDF shows that the slower *D2P** answers are compressed to the left, resulting in a much smaller right tail.

2.5.2 Accuracy

The accuracy results are much simpler to report (Figure 11). In the "same" conditions, the $D0P^{M}$ trials are less accurate than the *D0P** trials by about 1.62%. The "all-different" conditions show no difference in the accuracy between the two probe contents. In the "some-different" trials, the accuracies of *D2P** and $D2P^{M}$ trials are equal, and so are those of the $D2P^{D}$ and $D2P^{\pm}$ trials. The first pair is significantly less accurate than the second.

2.5.3 RT of incorrect answers

Initially, I had hoped that studying the RT of incorrect answers in this task would be simpler than in other experiments because of the larger number of trials per condition, which

should increase the number of errors. Thus, I first opted to exclude from this analysis the 6 participants that had a perfect accuracy in at least one condition, bringing down the sample to 13 participants. This method allowed me to treat the mean RTs as repeated measures, resulting in 16 measures per participant (correct and incorrect responses for the 8 conditions) and reducing the standard error. As can be seen by comparing Figure 9 (correct responses only) to Figure 12 (new figure), selecting only these 13 participants had an important effect on the mean RT of correct responses, reducing the means by between 20.6 and 29.9 ms depending on the condition.

Thus, the 6 participants I removed made a speed-accuracy trade-off and the sub-sample of 13 participants is not representative of the entire sample. Still, some observations can help guide future analyses. First, the incorrect "different" answers in *D0P** trials are faster than the correct "different" answers in all four *D2* conditions, just as I observed in the data review. The only other significant result is that incorrect "same" answers in *D2P** trials are significantly faster than correct "different" answers in *D2P**, $D2P^{M}$ and $D2P^{\pm}$ trials (and arguably significantly faster than correct "different" answers in $D2P^{D}$ as well).

Unfortunately, these results are not sufficient to properly test the hypotheses, and do not include all participants. I need ways to estimate the speed at which errors are made, especially those in *D0P**, $D0P^{M}$, *D2P** and $D2P^{M}$ trials if I wish to assess the proto-model's relevance. These extra analyses will be done as needed, as they depend on the question being asked.

2.6 Discussion

The results above are rich and must be combined together to properly test the hypotheses. In this section, I first go through the potential undesired effects of the Delayed Presentation protocol. Next, I test each hypothesis and determine if they are likely or not. Because it will be

shown that every hypothesis holds, I will then proceed to test whether the proto-model can account for all the aspects of the data.

2.6.1 Effects of the Delayed Presentation protocol

In this first subsection, I discuss potential or observed effects of the protocol on the results. This will prevent ambiguities when evaluating the hypotheses in the next subsection.

2.6.1.1 Lower answering thresholds.

The RTs in this experiment are faster than those observed in the data review (subsection 1.2.2), as can be seen by comparing the *D0P**, *D2P** and *D4P** trials of this experiment to the *L4D0*, *L4D2* and *L4D4* trials in the review (Figure 1 and Figure 9). Taken on its own, this result is not surprising: the mean RT of a condition varies across experiments because of their other uncontrolled variables. In this task, the fact that the number of *L4* trials each participant performed was 1152 instead of the usual 192 of most experiments in the review could explain these faster RTs by a simple training effect.

However, training would also predict a higher accuracy on the current experiment, but comparing Figure 4 and Figure 11 shows that it is lower by approximately 4% for the *D2* trials. A likely explanation would be that although the participants did not report consciously detecting the probe, they started "tuning in" with the task and lowered their threshold to answer "same". This would not affect the accuracy of *D0* and *D4* trials, but would make it riskier on *D2* trials. Furthermore, if matches are assumed to be encoded faster than mismatches, answering "same" after testing fewer matches would happen almost as often on *D2P** trials as it would on $D2P^{M}$ trials, whereas the $D2P^{D}$ and $D2P^{\pm}$ should not be affected. This is exactly the accuracy pattern we observe in Figure 11.

Note that the proto-model also makes the prediction that the accuracy of *D2P** trials should be equal to that of $D2P^{M}$ trials, but not that their accuracy would be lower than in other comparable tasks.

In conclusion, it is likely that training led the participants to make a speed-accuracy trade-off and take more risks by lowering their "same" thresholds. Fortunately, the proto-model can still be tested in these conditions because it simply assumes that a participant will keep its "same" threshold fixed within an experiment if *L* is fixed (after enough learning occurred).

2.6.1.2 Change detection effect.

Discussions with my colleagues brought up that it was conceivable that the Delayed Presentation protocol could provoke a "change detection" effect, where the apparition of the last two letters would be seen as a change in the stimulus. If this effect is "strong", then the apparition of the last two letters would trigger a "different" response, resulting in a lower accuracy with faster errors in $D0P^{M}$ trials and in a higher accuracy with faster correct answers in $D2P^{M}$ trials, when compared to their respective *P** conditions. On the other hand, if the change detection effect is "weak" it would only have a minimal impact on RT, slowing down the "same" answers and improving the "different" answers by initiating the response without the need to test and encode a mismatch.

The first step to study this question is to find if participants make faster or slower errors in P^{M} trials compared to *P** trials. As stated in the results section, I needed to come up with an appropriate analysis. I opted to define what is "fast" and "slow" for each participant, then to look at the proportion of errors that could be classified as such. To do so, I took the 20^{th} and 80^{th} percentiles of RT for the combined correct and incorrect answers for each condition, per participant, and then counted the number of incorrect answers that were faster or slower than

these threshold values. Finally, I summed the fast and slow errors across participants (the raw values are in Table 6 and the proportions are shown in Figure 13). With this new information in hand, we see that although the accuracy is lower on *D0P^{M}* trials, the distribution of the speed of errors is unchanged. Unfortunately, Figure 13 does not indicate the actual RT of the errors, which would be more relevant to determine if the errors in the *D0P^{M}* trials have a faster RT.

My next step was to approximate the RT of errors in *D0P** and *D0P^{M}* trials if I ignored the other conditions. Doing so allowed me to exclude only 3 participants instead of 6 (i.e., I applied pair-wise exclusion instead of list-wise exclusion). Because of the low number of errors, I used the median of each participant's median RT and found the Confidence Intervals using the Binomial distribution. For a sample size of 16 values, the 92.32% CI is the closest to the typical 95% CI, which is defined by the 5th and 12th values:

$$P(Y_5 < m < Y_{12}) = \sum_{k=5}^{12-1} \binom{16}{k} \times 0.5^{k} \times 0.5^{16-k} = 92.32\%.$$

This CIs of the two conditions are almost identical, but the medians are not. The errors in *D0P** have a median RT of 378.5 ms and a CI of [359; 450] ms, while the *D0P^{M}* trials have a median of 414.75 ms and a CI of [362; 454] ms. Out of curiosity, I also calculated the mean of the medians and use the CI formula based on Student's distribution to obtain a mean median of 403.25 ±43.86 ms for the *D0P** and 444.5±70.80 ms for the *D0P^{M}* trials. The two measures do not reveal a significant difference between the speed of errors on *D0P** and *D0P^{M}* trials. However, both hint that the errors in *D0P^{M}* trials are not faster than those in *D0P** trials, but slower.

This result means that the "strong" change detection effect cannot be involved and instead supports a "weak" change detection effect. Looking at the *D2* trials confirms this

interpretation: the correct answers on *D2P*^{M} trials are indeed faster than expected, as the CDF difference is smaller than 24 ms, and the accuracy of *D2P** and *D2P*^{M} trials are equal.

The very loose definition of a "weak change detection effect" is not formally testable, and the arguments in its favour are based on results that exclude 3 participants. However, it also is the best explanation for the CDF difference between the slower answers in *D0P** and *D0P*^{M} trials. Past the 50th percentile, the CDF difference is greater than 24 ms, which is either attributable to "same" answers having super capacity, or to something slowing down the answers on *D0P*^{M} trials. However, super capacity for "same" trials would predict an improvement of RT as *L* increases, which is not the case (Figure 1). Thus, something must be slowing down the *D0P*^{M} trials. In the proto-model, the only way to slow down a "same" answer is if a mismatch is tested after a few matches. In this scenario, the "different" answer is too inhibited to win the race against the "same" answer, but this extra test results in the "same" answer being given slightly later. In order to explain the CDF difference, however, this extra test would need to occur more often in *D0P*^{M} trials compared to *D0P** trials, and I cannot think of a reason why it would be the case. The weak change detection effect acts similarly as an extra test and is guaranteed to occur on every answer given after the last two letters appear. Thus, it allows for the faster answers in *D0P*^{M} trials (before the 20th percentile, the CDF difference is lower than 24 ms) and explains the slower "same" answers as well.

This effect is a sensible spurious effect of the Delayed Presentation protocol that does not invoke ill-defined concepts or processes, simply relying on the capacity of the brain to detect change and make use of that information. The only effect that seems to contradict it is the lower accuracy of *D0P*^{M} trials, but I will later show how it is in fact congruent and required. The only way to formally test whether the Delayed Presentation protocol induces this effect would be to

run a new experiment with the goal of targeting the presentation in two steps as the independent variable. Until then, the weak change detection effect it is the most likely explanation for the results, slowing down "same" answers and accelerating "different" answers.

2.6.2 Evaluating the hypotheses

I now proceed to evaluate the hypotheses stated in section 2.3. I first test the hypothesis of self-terminating "different" answers, because it will also give me information on the capacity of the decision-making process. Next, I verify whether fixed thresholds are plausible for "same" answers, then show evidence supporting that matches indeed inhibit the "different" responses. Finally, I use the conclusions from these three tests to determine if it is possible that the process(es) responsible for "same" and "different" answers use a single process during the decision-making stage, which tests the letters any number of times, in a serial random fashion.

2.6.2.1 Self-terminating "different" answers.

The idea that "different" answers are generated by a self-terminating process has not been contested in the literature, and this experiment is just another one of the many that make a case in favour of this hypothesis. A simple glance at the *D4* trials rules out the possibility of an exhaustive process, as the correct *D4P** and $D4P^{D}$ answers have the same mean RT.

Still, it is worth looking at the CDF and PDF differences of the *D4* trials. Note that the CDF difference is not only below 24 ms, but that it eventually becomes negative past the 65th percentile. Also, the PDF difference shows that the *D4P** answers have a longer right tail. In other words, this self-terminating process is affected negatively by the number of critical features, a tell-tale of a process with limited capacity.

To confirm this conclusion of limited capacity, I used the CDF and SOA tests. The CDF test compares combinations of *D2P** trials to the *D4P** trials (as explained in subsubsection

2.2.3.3 and detailed in the Appendix) to determine whether there is a redundancy effect of seeing more than one mismatch at a time. The SOA test uses $D4P^*$ and $D4P^D$ trials to find the capacity of the testing process responsible for these answers. Figure 14 plots the results of the CDF test (Equation 3) on the left and of the SOA test (Equation 4) on the right.

As expected, the CDF test shows no sign of the redundancy effect. Combined with the result of the SOA test, we find that a parallel self-terminating process with limited capacity is responsible for the RT of correct "different" answers (as can be seen by comparing Figure 14 and Figure 8).

This new information is bad news for the proto-model, which requires that letters be tested in serial. Fortunately, Sternberg showed that a parallel process with limited capacity can behave like a serial process with unlimited capacity (Sternberg, 1998, p. 411). This phenomenon called "model mimicry" is not simple to deal with and a Double Factorial Paradigm (Little, Altieri, Fific, & Yang, 2017) would be required to differentiate the two types of models with certainty.

I present two arguments in favour of the serial process with unlimited capacity and against the parallel process with limited capacity. The first is that of Occam's razor: even when granted greater flexibility, the best parallel process still has inferior fits compared to a simple self-terminating process with unlimited capacity (Sternberg, 1998, pp. 421–422). The second is that a limited capacity process would see its RT increase with L for "all-different" trials, which is not what we see in Figure 1.

Thus, although I cannot rule-out the parallel alternative, it is likely that the process that generates the "different" answers is serial self-terminating with unlimited capacity.

2.6.2.2 Fixed threshold for "same" answers.

The assumption of a fixed threshold for "same" answers was proposed as an alternative way of implementing exhaustive processing. By defining this fixed threshold, it becomes possible to give a "same" answer on a "some-different" trial without self-termination or varying thresholds. In other words, if this hypothesis is rejected, either "same" answers can be made in a self-terminating fashion or the "same" threshold can vary.

To test this hypothesis, I had planned to compare the speed of errors on *D2P^{M}* trials to those of correct answers on *D0P** and *D0P^{M}* trials. However, as I mentioned in the results section and in subsubsection 2.6.1.1, some participants did not make any mistakes and they cannot be excluded from the analyses without biasing the data. This led me to look for an alternative estimator of central tendency for the RT of errors in the *D2P^{M}* trials. Given that 48.12% of the errors on the *D2P^{M}* trials are faster than the 20th percentile of correct and incorrect RTs combined (Figure 13), I opted to use that 20th percentile as an estimator of the median of errors. This allowed me to extrapolate a median for all 19 participants, including the 6 who were excluded from the calculations in Figure 12. I then applied the method based on the Binomial distribution to calculate the CI of the median. With a sample size of 19, the best approximation of a 95% CI is the 93.64% CI defined by the 6th and 14th values:

$$P(Y_6 < m < Y_{14}) = \sum_{k=6}^{14-1} \binom{19}{k} \times 0.5^{k} \times 0.5^{19-k} = 93.64\%.$$

As before, I complemented these with the 95% CI based on the mean of the (extrapolated) medians. I calculated these four values for the errors on *D2P^{M}* trials and correct answers in *D0P** and *D0P^{M}* trials.

The results, shown in Table 7, were surprising: although the *D2P*^{M} and *D0P*^{M} trials are indistinguishable to the participants during the first 24 ms, the median RT of incorrect *D2P*^{M} trials is more similar to that of correct *D0P** trials (note that this supports unlimited capacity). This led me to expect the median RT of errors on *D2P** trials to also be comparable to that of correct answers on *D0P** trials. This post-hoc assumption was confirmed, as seen in Table 7. Interestingly, the weak change detection effect predicts that the incorrect "same" answers on *D2P*^{M} trials should be slower than those in *D2P** trials, which is arguably the case.

It would be unreasonable to assume that the way these incorrect "same" answers are generated is what happens on every correct "same" answer. If that was the case, then almost all *D2P*^{M} trials would be incorrect. Instead, it suggests that the incorrect "same" answers are not abnormally fast and that a single process is responsible for both correct and incorrect "same" answers. Furthermore, there is evidence that this process is also at work in *D0* trials: Figure 10 shows that the CDF difference between the *D0P** and *D0P*^{M} trials is of less than 24 ms before the 20th percentile, and thus some *D0P*^{M} trials are answered faster than expected.

So far, I have provided arguments for a single process generating all "same" answers. I must now show that a process that is self-terminating or that has a variable threshold could not explain the data. To do so, we only need to look at the CDF difference between *D0P** and *D0P*^{M}.

At a first glance, it seems reasonable to rule out "same" self-terminating processes because they would require that the RT of correct answers be unaffected by the probe (i.e. the RT of *D0P** and *D0P*^{M} would be identical). This argument can be countered by invoking two processes that answer "same": one self-terminating with a low probability of being used, and another exhaustive but with a higher probability of being used (note that this self-terminating process is equivalent to an exhaustive process with a lower threshold). The self-terminating

process would answer faster at the cost of making more errors on "some-different" trials. However, there is no reason for the self-terminating process to be triggered more often on *D0P** or $D0P^{M}$ trials, which means that it cannot explain why the CDF difference would be smaller or larger than 24 ms.

Explanations that involve a process with a variable threshold do not fare any better: they are also incapable of explaining the CDF difference we observe between the *D0* conditions. Furthermore, the threshold distribution would need to be bimodal to account for the fact that almost 50% of the incorrect "same" answers are very fast. While this is not impossible, I am not aware of a cognitive theory that would support this sort of threshold behaviour.

In short, self-terminating processes and exhaustive processes with variable thresholds are unable to account for all the data (and are also overly complex). This means that "same" answers are probably given by an exhaustive process with a fixed threshold. Combined with the discussion on the "different" answers, we now know that to be compatible with the proto-model, "same" answers must be explained by a serial exhaustive process with unlimited capacity.

2.6.2.3 Matches inhibiting the "different" answer.

I now present evidence that matches can inhibit the "different" answers by comparing the $D2P^{D}$ and $D4P^{D}$ trials, as well as the $D2P^{M}$ and *D4* trials.

Considering that $D2P^{D}$ trials are indistinguishable from $D4P^{D}$ trials during the first 24 ms and that "different" answers are self-terminating, there are very few explanations as for why $D2P^{D}$ trials are slower than the $D4P^{D}$ trials. My explanation is that 24 ms is not always enough to encode and test a mismatch, which makes sense given that *L1D1* trials are approximately 50 ms slower than *L1D0* trials (fast-same intercept, subsubsection 1.2.2.1). The accuracy data also supports this idea, as the $D2P^{D}$ trials are only 2.14% more accurate than *D2P** trials. This means

that on most trials, the only impact of showing mismatches earlier is to reduce the number of matches that will be tested before a mismatch is encoded, resulting in less inhibition. If we look at the CDF difference of *D2P** and $D2P^{D}$ trials, it shows that $D2P^{D}$ answers are much faster than expected before the 65th percentile, which is congruent with this explanation. Thus, the reason why even $D2P^{D}$ trials are so much slower than $D4P^{D}$ trials is that they are subjected to inhibition by matches.

Alternatively, given that "self-termination" does not mean "instantaneous", it is also possible that matches tested after a mismatch could cause inhibition. It would be difficult to test which of the two explanations are at work here, although this alternative seems less likely.

Evidence that matches are causing inhibition is much easier to obtain if we compare the $D2P^{M}$ and *D4* trials. If matches did not inhibit the "different" answer, $D2P^{M}$ trials would be at most 24 ms slower than *D4* trials. Even if both matches were tested before the mismatches appeared, the RT difference between $D2P^{M}$ and *D4* answers would not be larger than what is observed between $D0P^{M}$ and *D0P** trials. As this is not the case, I conclude again that matches inhibit mismatches.

2.6.2.4 A serial random testing process.

With every other hypothesis confirmed, it is now time to assess the plausibility of a serial testing process that selects letters in a random order such that they can be tested any number of times (subsection 2.3.2). To support this hypothesis, I need to show that the "same" answers cannot be given by a parallel exhaustive process, since I have already made a case for "different" answers being generated by a serial process with unlimited capacity.

In a parallel exhaustive process, the RT only depends on the slowest feature; no matter how quickly the first 3 letters are encoded and tested, no answer will be given before the last

letter is tested. Thus, for the "same" answers to be explained by a parallel process, the RT difference between the *D0P** and *D0P^{M}* trials should be exactly 24 ms, because two of the letters are seen 24 ms later, shifting the PDFs (and the mean RTs) by 24 ms. Figure 9 seems to support this idea, as the mean RT difference of 26.9 ms is statistically indistinguishable from the expected 24 ms. The PDF difference, however, clearly shows that the two distributions are not simply shifted versions of each other: the *D0P** trials have a denser PDF and a shorter right tail. This could be explained by super capacity, but this is not plausible for "same" answers.

Another result that reduces the plausibility of a parallel exhaustive process is the fast incorrect "same" answers observed on *D2* trials. These answers are given with only two matches and thus the process is not exhaustive. To work around this problem, one can augment this parallel process so that it tests letters many times, making it "exhaustive" in the same sense as the proto-model. However, this would not explain why incorrect "same" answers in *D2P** (based on only 2 matches) are faster than correct "same" answers on *D0P^{M}* trials (see Table 7). Furthermore, this augmented process would not explain the CDF difference between *D0P** and *D0P^{M}* trials.

In conclusion, parallel testing does not seem to be involved in the "same" answers. Because neither "same" nor "different" answers are likely to be generated by a parallel process, the serial random testing process hypothesis holds. Similarly, the hypothesis of the single testing process being used by both answers also holds.

2.6.3 Explaining the same-different data

With every assumption of the proto-model being plausible, it is now time to verify if it can account for the data that was not yet addressed.

2.6.3.1 Accuracy of D0P^{M} trials.

The proto-model has no problem explaining why the accuracy of *D0P^{M}* trials is lower than that of *D0P** trials. Because the only way a "different" answer can be made on a "same" trial is if noise creates a spurious mismatch (or if noise occurs in the response-production stage), the probability of making a mistake increases with the time taken to give the "same" answer. We know that correct answers on *D0P^{M}* trials take longer than *D0P** trials (thanks to the delay in match presentation and to the weak change detection effect), and thus the lower accuracy is expected. Note that the proto-model predicts that these errors will be slower, which could be the case (subsubsection 2.6.1.2). Hence, the lower accuracy of *D0P^{M}* is not in contradiction with the proto-model.

2.6.3.2 RT of D2P^{M} trials.

The RT of *D2P^{M}* trials are better than what would be naïvely expected. Yet, remember the proto-model's first prediction: incorrect "different" answers (given upon seeing matches) should be almost as fast as correct "same" answers (see subsection 1.2.4), and I have already shown that incorrect "different" answers are faster than correct ones (Figure 12). In a *D2P^{M}* trial, however, answering "different" upon testing matches would not be incorrect and would result in a fast "different" answer. Thus, noise is working in favour of the "different" trials, resulting in a CDF difference of less than 24 ms between the *D2P** and *D2P^{M}*.

Because noise is more likely to turn a match into a mismatch than the other way around, the *D2P^{M}* trials are prone to spurious mismatches as much as *D0P^{M}* trials. Whereas these mismatches result in a lower accuracy for *D0P^{M}* trials (compared to *D0P** trials), they benefit the *D2P^{M}* trials in two ways: they remove a source of inhibition and initiate the "different" answer. Of course, this does not happen on every trial, but it generates enough fast "different" answers to

impact the CDF difference. Combined with the change detection effect, this gives two advantages to the $D2P^{M}$ trials.

Note that when noise does not intervene, the inhibition from matches will act the same way as it does in the $D2P^{D}$ trials. Because the only difference between $D2P^{*}$ and $D2P^{M}$ trials is the number of matches that will be tested before a mismatch (subsubsection 2.6.2.3), it is likely that trials where only one match is missing to make a "same" answer will be more frequent in $D2P^{M}$ trials, explaining the trailing right tail of the PDF of these trials.

2.6.3.3 The $D2P^{\pm}$ trials.

I have not touched on the $D2P^{\pm}$ trials, which were added as a way to delve into questions that this project is not concerned with. Although I think their data could be used in a future research, this current project had one purpose: to determine whether it is worth creating and testing a quantitative version of the proto-model. Thus, exploring the results from these trials will be left to the reader and future researchers.

2.7 Conclusion

With all the hypotheses and assumptions of the proto-model being respected and given that it can explain most of the data points without being overly complex, I believe that developing a quantitative and formal model based on its assumptions is warranted.

Although there are no reasons to reject the proto-model outright, it has room for improvements that require more investigation. First, it would be important to know whether "different" answers are given by a parallel model with limited capacity or a serial model with unlimited capacity. Although the later seems more likely, it still needs to be confirmed. Second, the claim that processes that are self-terminating or that have a variable threshold cannot account for the difference in RT between the $D0P^{*}$ and $D0P^{M}$ trials should be verified quantitatively.

This would require adequate implementations of these two models, which fell beyond the scope of this project. Third, a more convincing argument should be made against a parallel exhaustive model for "same" answers. I concluded that it was not likely but did not by any means prove it wrong. Fourth, the serial testing process has the liberty of not testing a letter, but it remains to be known if this can happen during the probe, if it happens at all. Finally, the proto-model's explanation of the RT of $D0P^M$ trials relies on a "weak change detection effect", which I think is unconvincing. Unfortunately, because matches are not critical features, it is much harder to assess the impact of delaying their presentation on RT (whereas the CDF and SOA tests work well for mismatches). Maybe a better explanation can be found by looking at the data, or maybe there is some literature on the matter that I missed that could provide a clearer answer.

On the other hand, the proto-model was particularly good at predicting the data patterns. Also, the fact that a single serial process can feed both answers make it simpler than most alternatives. Its most important achievement, however, is that it makes predictions on the accuracy and the RT of incorrect answers. Only Ratcliff's model predicts those, with limited success and using a large number of free parameters. If a quantitative version of the proto-model is ever developed, I would be thrilled to compare its performance to the Diffusion model.

3 Closing Remarks

This project is not done. In the first chapter, I detailed the previous research on the same-different task and showed how every theory and model that was proposed failed to explain the data. Yet, most of these ideas were full of brilliant insights. Standing on the shoulders of giants with my larger dataset, I came up with a proto-model that seemed to respect every restriction imposed by the data and previous findings. In the second chapter, I tested the qualitative predictions of this proto-model using the Delayed Presentation protocol and the differential effects of probe contents on the RT distributions and accuracies. By analyzing and combining the RT of correct and incorrect answers to the accuracy, I argued that the proto-model made an acceptable job at fitting the data qualitatively. In the third (and missing) chapter, I would detail the parameters of the model and how they relate to each other. It would also contain an appendix with the code for this model, on which I already spent countless hours, and a comparison of my model's performance versus new contenders in the field, such as Bradley Harding's excellent take on the subject.

With this third chapter absent, it will be hard to convince the community that this proto-model is the answer that we have been searching for the last 50 years. History has shown repeatedly that while theories are perfect *in theory*, their implementations are always imperfect, and often disastrous.

This does not mean that I think my work is subpar. I do think that my proposition is the best made so far. It could account not only for the RT of correct answers, which most models failed at, but also for the accuracy and the speed of incorrect answers. The conclusions and deductions that I drew from the results are based on hundreds of hours of insight and readings. I

made an honest and thorough effort to be the best scientist I could be. Still, I will only be satisfied with this proto-model when it becomes a formal, quantitative model.

This project is not done. Maybe I will get back to it one day. Until then, “so long, and thanks for all the fish”.

4 Appendix

4.1 Creating the Groups of Trials for the CDF and SOA Tests

There are six ways the two mismatches can be positioned in a *L4D2* trial. The CDF test requires that each of these six configurations be treated as unique to prevent assuming that letter positions are equally salient (which is unlikely, given what is known about the way we read words). Getting enough trials to create a CDF for each participant in each probe condition is not realistic in a single testing session, and thus an alternative had to be found.

The alternative I propose is to compare two groups of three configurations, which would effectively triple the number of trials and thus give a reliable CDF. The six configurations can be paired such that each pair has a mismatch in every position. If x is a match and A and B are mismatches, those pairs are AAxx and xxBB, AxAx and xBxB, and AxxA and xBBx.

Because our goal is to compare the group containing A with the group containing B, we are forced to create the two groups by separating the members of each pair, such that the A group contains AAxx, AxAx and AxxA, and the B group contains xxBB, xBxB and xBBx. Unfortunately, this is not adequate because A is always in the first letter position and B never is.

To solve this problem, we can use the fact that A and B are totally interchangeable in the same-different task. Whereas they represent two separable attributes in a visual search, A and B both represent mismatches in our context. Hence, we can use the mirrored pairs (which only exist conceptually): BBxx and xxAA, BxBx and xAxA, and BxxB and xAAx.

Next, because any group of three configurations contains six mismatches, these mismatches cannot be evenly distributed between the four letter positions. If we make two pairs of A and B groups, however, we can balance both the type of trials and the number of mismatches in each position, as shown in Table 8. However, this leaves a possibility that the

arbitrarily chosen pairs happen to give the same conclusion. To be more thorough, we can test every single combination of pairs. There are 6 possible configurations and we wish to take all groups of 3 configurations, for a total of 20 groups (and thus pairs).

The CDF test compares the $D2P^*$ trials to the $D4P^*$ trials, while the SOA test compares the $D4P^D$ trials to the $D4P^*$ trials. Because the position of the mismatches in $D2P^*$ and of the letters in the probe in $D4P^D$ have both have the same 6 configurations, this method of combining the RTs works for the two tests.

Once we have created the 20 pairs, we create the 20 CDFs $F_A(t)$ and $F_B(t)$ and apply the formulas required for the test. We then compare the 20 test results to see if they agree on the same conclusion, which they did for every participant. Hence, I used the mean test results for each participant, then computed the mean across participants to generate Figure 14.

5 Bibliography

Bamber, D. (1969). Reaction times and error rates for "same" - "different" judgments of multidimensional stimuli. *Perception & Psychophysics*, *6*(3), 169–174. https://doi.org/10.3758/BF03210087

Bamber, D. (1972). Reaction times and error rates for judging nominal identity of letter strings. *Perception & Psychophysics*, *12*(4), 321–326. https://doi.org/10.3758/BF03207214

Bindra, D., Donderi, D. C., & Nishisato, S. (1968). Decision latencies of "same" and "different" judgments. *Perception & Psychophysics*, *3*(2), 121–136. https://doi.org/10.3758/BF03212780

Cousineau, D., Thivierge, J., Harding, B., & Lacouture, Y. (2016). Constructing a Group Distribution From Individual Distributions. *Canadian Journal of Experimental Psychology*, *70*(3), 253–277. https://doi.org/10.1037/cep0000069

Decker, L. R. (1974). The effect of method of presentation, set, and stimulus dimensions on "same"-"different" reaction times. *Perception & Psychophysics*, *16*(2), 271–275. https://doi.org/10.3758/BF03203941

Egeth, H. E. (1966). Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, *1*(4), 245–252. https://doi.org/10.3758/BF03207389

Eriksen, C. W., & O'Hara, W. P. (1982). Are nominal same-different matches slower due to differences in level of processing or to response competition? *Perception & Psychophysics*, *32*(4), 335–344. https://doi.org/10.3758/BF03206239

Farell, B. (1977). *Encoding and Comparisons in "Same"-"Different" Judgments*. McGill University.

Farell, B. (1985). “Same”-"Different" Judgments: A Review of Current Controversies in Perceptual Comparisons. *Psychological Bulletin*, *98*(3), 419–456. https://doi.org/10.1037/0033-2909.98.3.419

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*(3), 389–413. https://doi.org/10.1037/0096-3445.136.3.389

Hawkins, H. L. (1969). Parallel processing in complex visual discrimination. *Perception & Psychophysics*, *5*(I), 56–64.

Hock, H. S. (1973). The effects of stimulus structure and familiarity on same-different comparison. *Perception & Psychophysics*, *14*(3), 413–420. https://doi.org/10.3758/BF03211176

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt and Company.

Krueger, L. E. (1978). A theory of perceptual matching. *Psychological Review*, *85*(4), 278–304. https://doi.org/10.1037/0033-295X.85.4.278

Little, D. R., Altieri, N., Fific, M., & Yang, C.-T. (2017). *Systems Factorial Technology: A Theory Driven Methodology for the Identification of Perceptual and Cognitive Mechanisms* (1st ed.). Elsevier. https://doi.org/10.1016/C2015-0-00849-8

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279. https://doi.org/10.1016/0010-0285(82)90010-X

Mordkoff, J. T., Miller, J., & Roch, A.-C. (1996). Absence of Coactivation in the Motor Component: Evidence From Psychophysiological Measures of Target Detection. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 25–41. https://doi.org/0096-1523/96/$3.00

Nickerson, R. S. (1965). Response times for “same”-"different" judgments. *Perceptual and Motor Skills*, *20*(1), 15–18. https://doi.org/10.2466/pms.1965.20.1.15

Nickerson, R. S. (1967). “Same” “Different” Response Times with Multi-Attribute Stimulus Differences. *Perceptual and Motor Skills*, *24*(2), 543–554. https://doi.org/10.2466/pms.1967.24.2.543

Nickerson, R. S. (1978). On the Time it Takes to Tell Things Apart. In J. Requin (Ed.), *Attention and Performance VII* (pp. 77–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nishisato, S., & Wise, J. S. (1967). Relative probability, interstimulus interval, and speed of the same-different judgment. *Psychonomic Science*, *7*(2), 59–60. https://doi.org/10.3758/BF03331077

Proctor, R. W. (1981). A unified theory for matching-task phenomena. *Psychological Review*, *88*(4), 291–326. https://doi.org/10.1037/0033-295X.88.4.291

Proctor, R. W. (1986). Response bias, criteria settings, and the fast-same phenomenon: A reply to Ratcliff. *Psychological Review*, *93*(4), 473–477. https://doi.org/10.1037/0033-295X.93.4.473

Proctor, R. W., & Rao, K. V. (1982). On the “misguided” use of reaction-time differences: A discussion of Ratcliff and Hacker (1981). *Perception & Psychophysics*, *31*(6), 601–602. https://doi.org/10.3758/BF03204200

Proctor, R. W., Rao, K. V., & Hurst, P. W. (1984). An examination of response bias in multiletter matching. *Perception & Psychophysics*, *35*(5), 464–476. https://doi.org/10.3758/BF03203923

Raab, D. H. (1962). Statistical Facilitation of Simple Reaction Times. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574–590. https://doi.org/10.1111/j.2164-0947.1962.tb01433.x

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212–225. https://doi.org/10.1037/0033-295X.92.2.212

Ratcliff, R., & Hacker, M. J. (1981). Speed and accuracy of same and different responses in perceptual matching. *Perception & Psychophysics*, *30*(3), 303–307. https://doi.org/10.3758/BF03214286

Sternberg, S. (1998). Inferring Mental Operations from Reaction-Time : How We Compare Objects. In D. Scarborough & S. Sternberg (Eds.), *An Invitation to Cognitive Science* (2nd ed., Vol. 4, pp. 365–420). Cambridge, Massachusetts: MIT Press.

Taylor, D. A. (1976). Effect of identity in the multiletter matching task. *Journal of Experimental Psychology. Human Perception and Performance*, *2*(3), 417–428. https://doi.org/10.1037/0096-1523.2.3.417

Townsend, J. T., & Nozawa, G. (1995). Spatio-Temporal properties of elementation perception An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Ulrich, R., & Miller, J. (1997). Tests of race models for reaction time in experiments with asynchronous redundant signals. *Journal of Mathematical Psychology*, *41*(4), 367–381. https://doi.org/10.1006/jmps.1997.1181

6 Tables

	D0	*D1*	*D2*	*D3*	*D4*
L1	12	12			
L2	12	6	6		
L3	12	4	4	4	
L4	12	3	3	3	3

Table 1: Distribution of trials with *L* letters and *D* mismatches in a typical same-different task in Denis Cousineau's laboratory. By doing this block 8 times, participants go through a total of 768 trials in approximately 40 minutes.

Condition	Mean RT (in ms)	95% mean CI
L1D0	464.80 ±8.47	[456.33; 473.28]
L2D0	481.17 ±7.87	[473.30; 489.03]
L3D0	501.07 ±6.53	[494.54; 507.59]
L4D0	521.51 ±7.32	[514.19; 528.83]
L1D1	518.78 ±6.80	[511.97; 525.58]
L2D1	546.04 ±7.84	[538.20; 553.89]
L3D1	582.00 ±9.07	[572.92; 591.07]
L4D1	608.29 ±12.14	[596.15; 620.43]
L2D2	514.66 ±7.20	[507.46; 521.86]
L3D2	532.95 ±7.85	[525.10; 540.80]
L4D2	554.72 ±9.11	[545.62; 563.83]
L3D3	516.70 ±7.56	[509.15; 524.26]
L4D3	524.95 ±7.69	[517.26; 532.63]
L4D4	522.82 ±9.58	[513.24; 532.39]

Table 2: Mean and 95% confidence intervals of response times for correct answers on trials with *L* letters and *D* mismatches, as shown in Figure 1.

Property	$RT_{L0D0} < RT_{L0D1}$	$RT_{L0D0} = RT_{L0D1}$
Intercept	$a_s < a_d$	$a = a_s = a_d$
Slope	$b_M \leq b_D$	$b_M < b_D$
RT_{L1D0}	$a_s + b_M$	$a + b_M$
RT_{L1D1}	$a_d + b_D$	$a + b_D$

Table 3: Summary of plausible scenarios if we were able to measure the RT of stimuli composed of no letters. If "same" stimuli (*L0D0*) are faster than their "different" counterpart (*L0D1*), the RT advantage of *L1D0* trials is partially (or entirely) explained by the intercept. If *L0D0* are as fast as *L0D1*, then the intercept is 0 and the RT advantage of *L1D0* is exclusively explained by the slopes.
In these formulas, *a* is the non-decision time, *b* is the slope, *s* and *d* refer to "same" and "different" answers and *M* and D refer to matches and mismatches.

	D0	*D2*	*D4*
P^*	432	72	144
P^M	144	72	-
P^D	-	72	144
$P^{\pm}$	-	72	-

Table 4: Distribution of trials in the experiment detailed in section 2.4. Figure 6 details each type of trial.

Difference (in ms)	Interpretation of the CDF difference	Interpretation of the PDF difference
$\text{diff} < 0$	The process has limited capacity and is either self-terminating, or letters can be tested multiple times in less than 24 ms.	The P* trials' PDF is stretched to the right.
$\text{diff} = 0$	If testing a letter takes more than 24 ms, the process is self-terminating with unlimited capacity. If testing a letter takes less than 24 ms, the process is exhaustive and can test letters multiple times.	The PDFs superpose.
$0 < \text{diff} < 24$	Testing a letter takes less than 24 ms.	
$\text{diff} = 24$	The content of the probe had no impact.	The P* trials' PDF is stretched to the left.
$24 < \text{diff}$	Either the process has super capacity (making the P* trial faster) or slowed down by the apparition of the last two letters (making the non-P* trial slower).	

Table 5: Summary of the interpretation of the difference between the CDF and PDF curves shown in Figure 10. The difference is obtained by subtracting the value of the percentile of the non-P* condition to that of the P* condition.

	*D0P**	*D0P^{M}*	*D2P**	*D2P^{M}*	*D2P^{D}*	*D2P$^{\pm}$*	*D4P**	*D4P^{D}*
Fast	173	74	80	81	45	52	50	32
Normal	114	47	19	24	23	20	8	15
Slow	168	74	31	28	34	21	30	41

Table 6: Frequency of fast, normal and slow errors per condition. Errors on "same" trials are distributed similarly, but differ from the "different" trials. The definition of "fast" and "slow" changes per participant; their respective 20th and 80th RT percentiles (including both correct and incorrect answers) were used as threshold values for each condition. Figure 13 shows these values as proportions of fast, normal and slow errors per condition.

	Median (in ms)	93.64% median CI	Mean (in ms)	95% mean CI
Incorrect *D2P**	385	[345; 423]	394.5	[361.1; 428.0]
Incorrect $D2P^{M}$	401	[347; 438]	404.7	[374.2; 435.2]
Correct *D0P**	407	[360; 433]	404.7	[375.6; 433.9]
Correct $D0P^{M}$	438	[389; 456]	431.8	[401.1; 462.5]

Table 7: Median and mean of the median RT, as well as their respective confidence intervals, for trials where the probe only contains matches.

Pair	Trials used to create $F_A(t)$	Trials used to create $F_B(t)$
1	AAxx, AxAx, AxxA	xxBB, xBxB, xBBx
2	xxAA, xAxA, xAAx	BBxx, BxBx, BxxB

Table 8: Example of two pairs of groups of conditions that can be used to create the CDFs used in the CDF and SOA tests. If the result of the two tests are the same for both pairs, then the tests' conclusion likely holds.

7 Figures

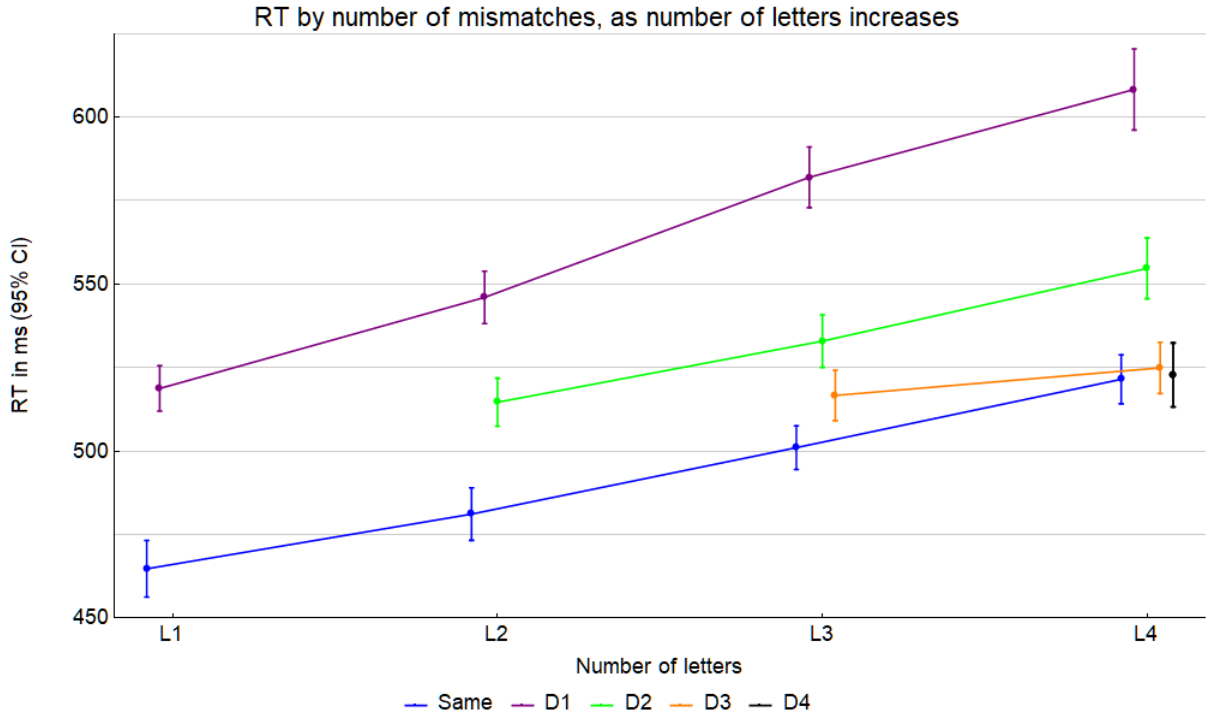

Figure 1: Mean and 95% confidence intervals of response times (in ms) of "same" and "different" trials with L letters and D mismatches between S_1 and S_2. The fact that "same" trials are typically faster than the "different" is called the "fast-same" effect. This figure was made using the data from 8 comparable experiments.

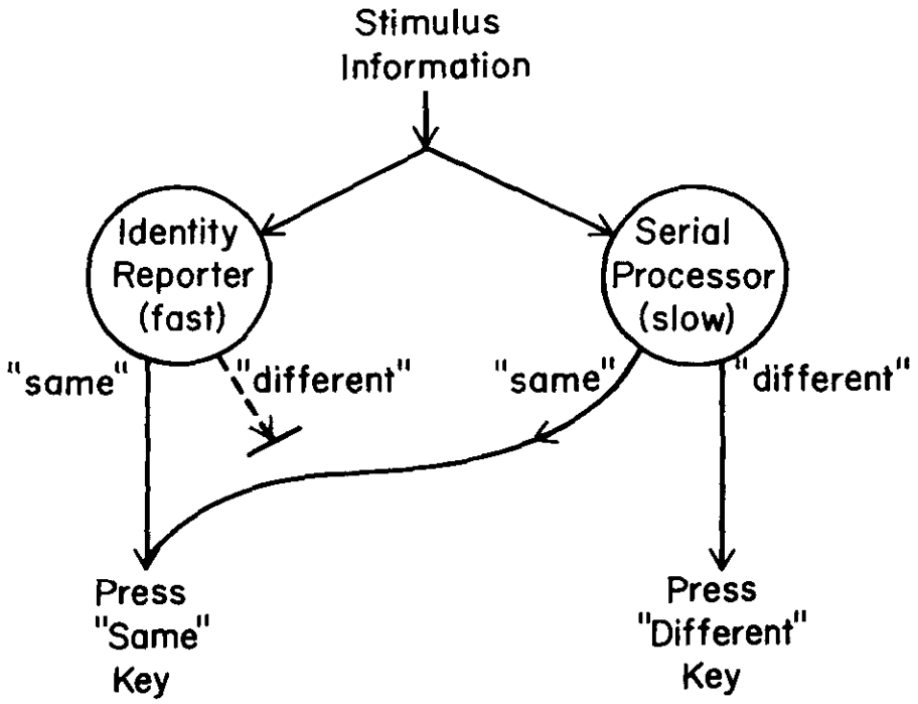

Figure 2: Figure 3 from Bamber's first article on the "same-different" process (Bamber, 1969), in which he details his Identity Reporter dual process model of RT. The Identity Reporter tests letters in parallel and can only answer "same", whereas the serial processor tests letters in serial but can answer both "same" and "different".

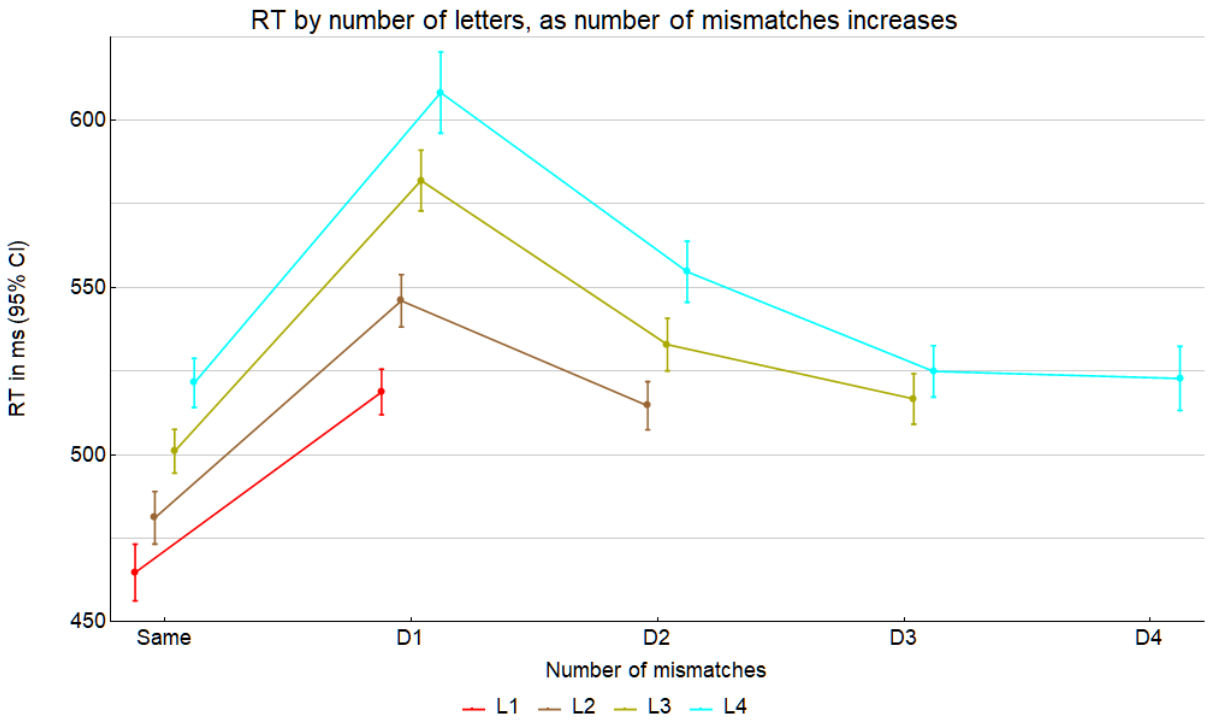

Figure 3: Mean and 95% confidence intervals of response times (in ms) of "same" and "different" trials with *D* mismatches between S_1 and S_2 and *L* letters. The values are identical to those of Figure 1, but were transposed to better illustrate the effect of increasing *D* on the RT.

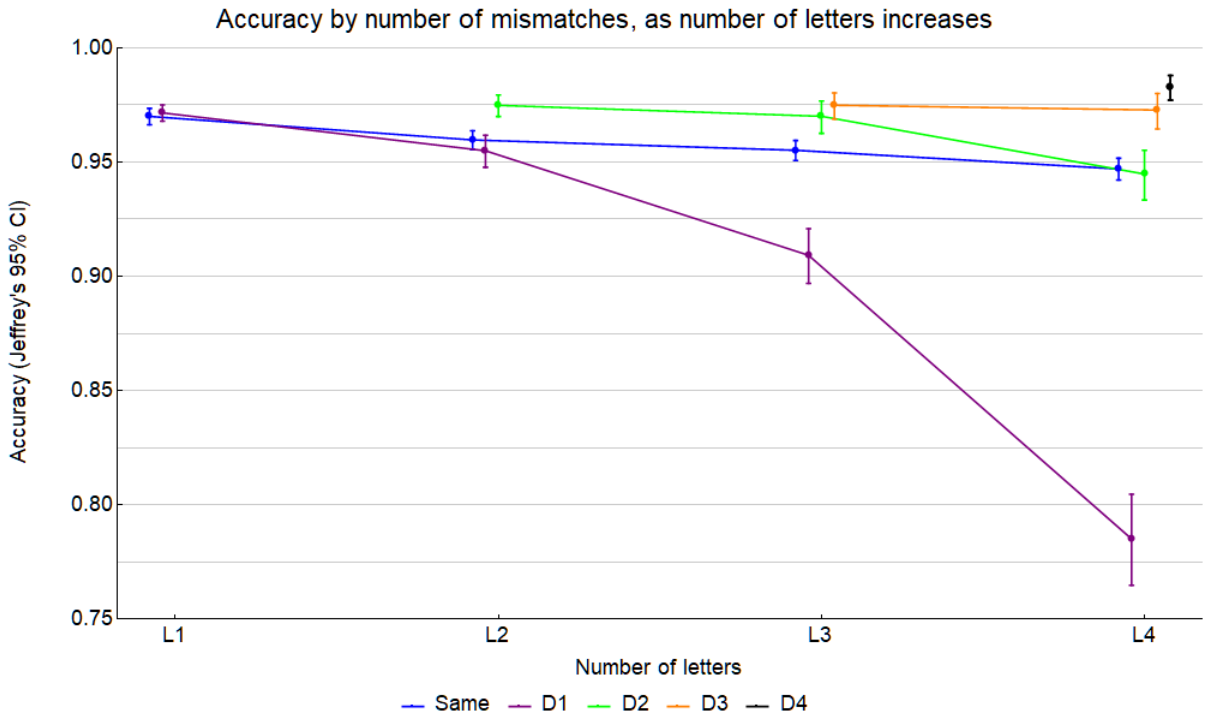

Figure 4: Mean and 95% Jeffrey's confidence interval of the accuracy of "same" and "different" trials with *L* letters and *D* mismatches between S_1 and S_2.

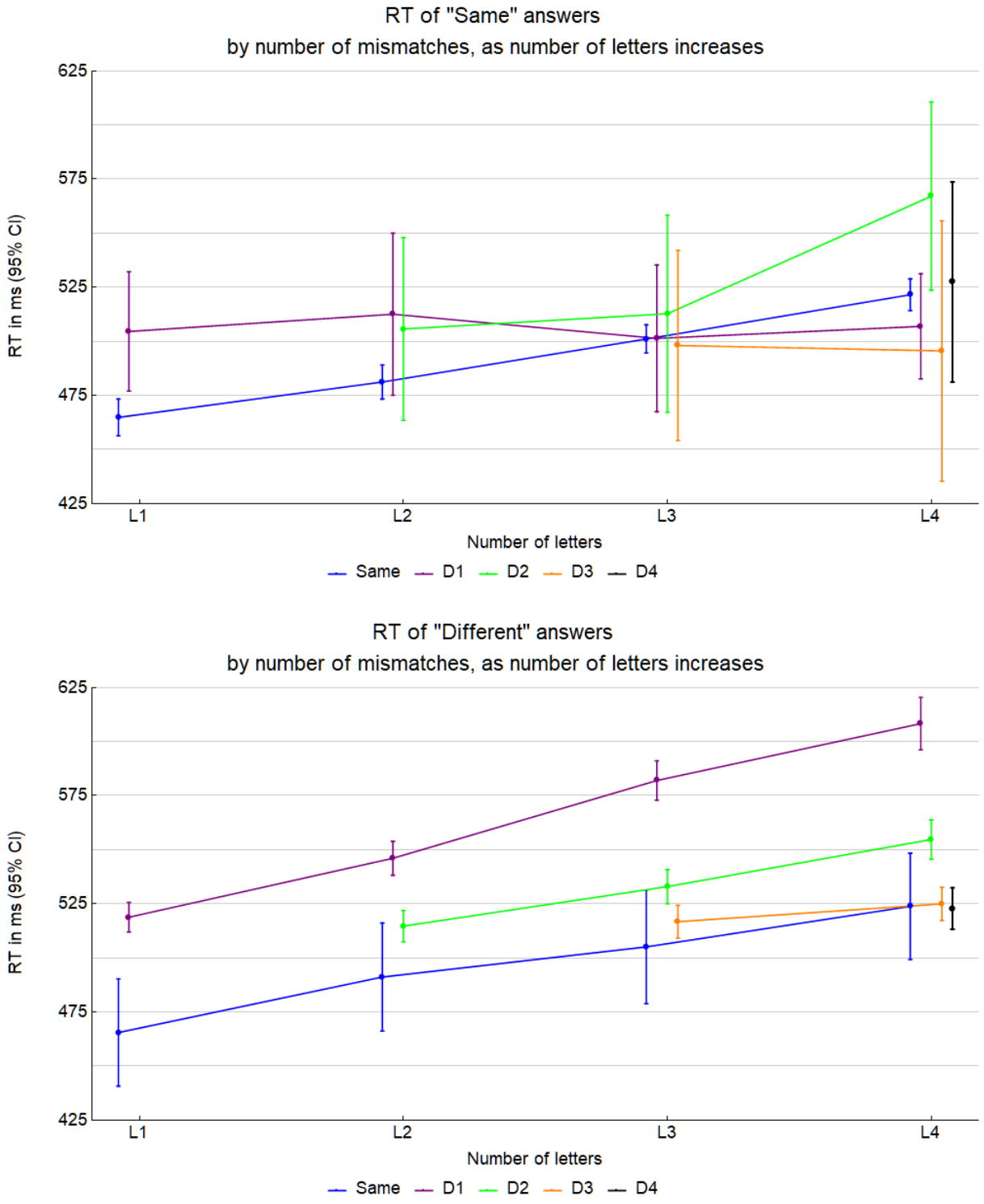

Figure 5: Top: RT of "same" answers given in all 14 conditions. The blue line represents correct answers, the others are errors. Bottom: RT of "different" answers given in all 14 conditions. The blue line represents errors, the others are correct answers.

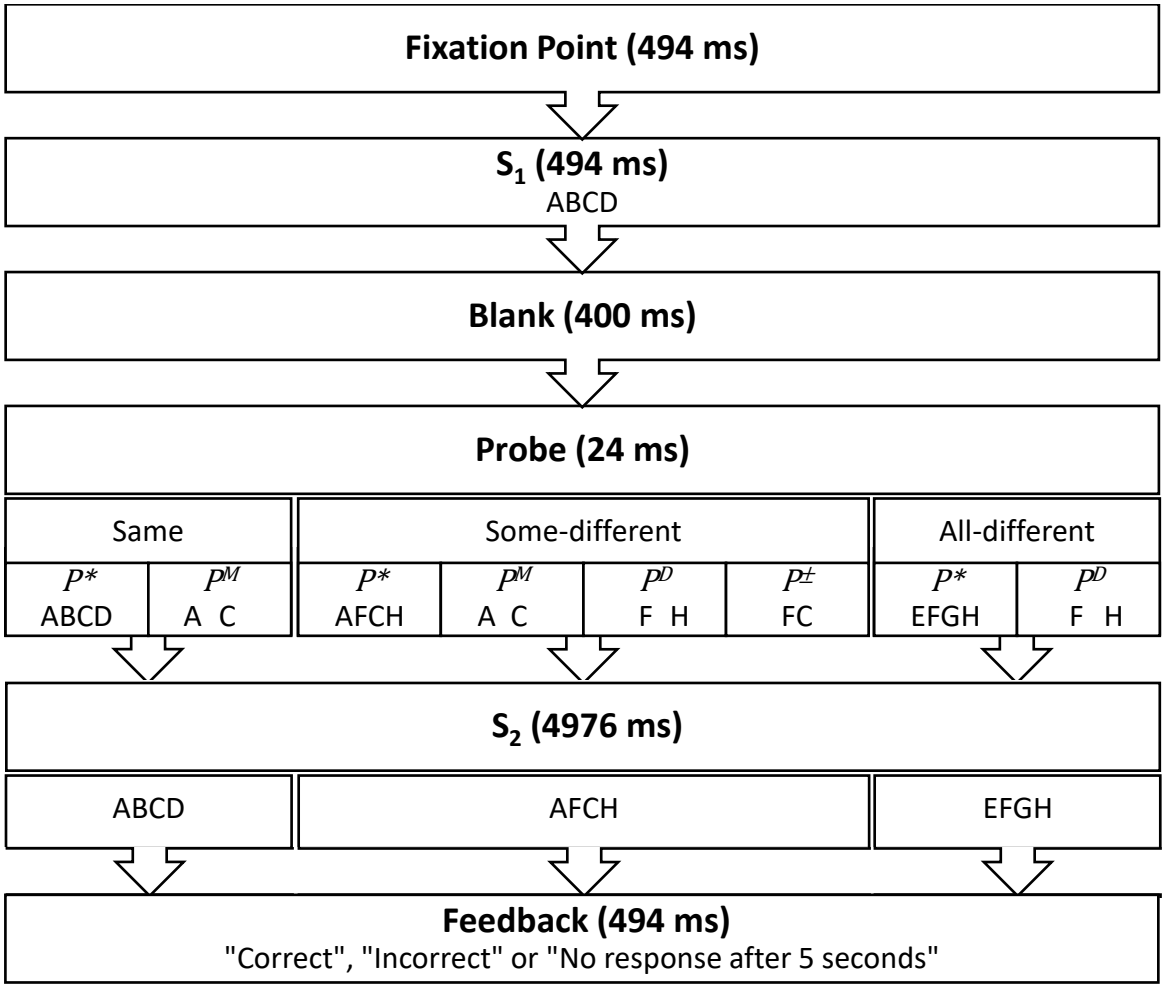

Figure 6: Detailed timing of an individual trial in the Delayed Presentation protocol. When the entire content of S_2 is shown in the probe (the *P** conditions), the trial is identical to a typical same-different trial. When the probe contains only partial information on S_2, this protocol can help identify properties of the process making the answers through the RT of correct and incorrect answers and the accuracy.

TABLE 1

Model's Predictions for the CDF and SOA Tests

Model	Brief Description	CDF Test	SOA Test
Race models	Faster of two racers determines *RT*	Always passed	Always passed
Triggered-moment models	Faster of two racers triggers a moment of duration *Q*. Sensory input arriving during this moment is accumulated. A response is initiated at the end of this moment, and response speed increases with the amount of accumulated sensory input.	May be violated	May be violated
Distinct-signals models	Similar to race models but additionally assumes a redundant racer in redundant-signals trials	May be violated	Always violated
Superposition model	Each stimulus starts a Poisson process. Detection occurs when a criterion of $c > 1$ pulses is reached.	Always violated	Always passed
Limited capacity models	Fixed amount of central capacity is shared between channels (exponential detection times)	Always passed	Always violated
Super capacity models	Double stimulation increases the amount of central capacity (exponential detection times)	Always violated	Always violated

Figure 7: Table 1 of Miller and Ulrich (1997) showing how combining Miller's Inequality with their Stimulus Onset Asynchrony test allows to discriminate between 12 types of models.

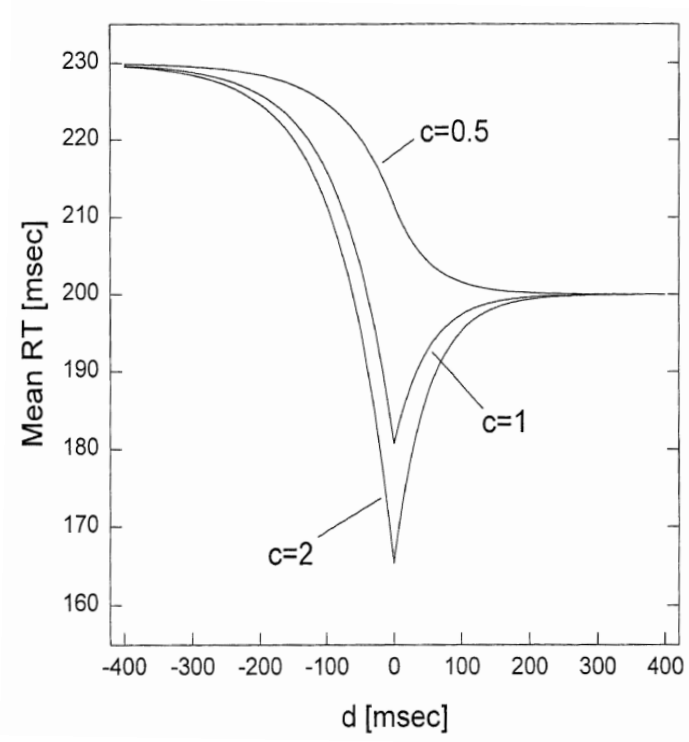

Figure 8: Figure 6 of Miller and Ulrich (1997) showing RT as a function of the delay (*d*) between the apparition of the two critical features of a stimulus, depending on the capacity of the process. The SOA test uses the RT to discriminate processes with super (c = 2), unlimited (c = 1) and limited (c = 0.5).

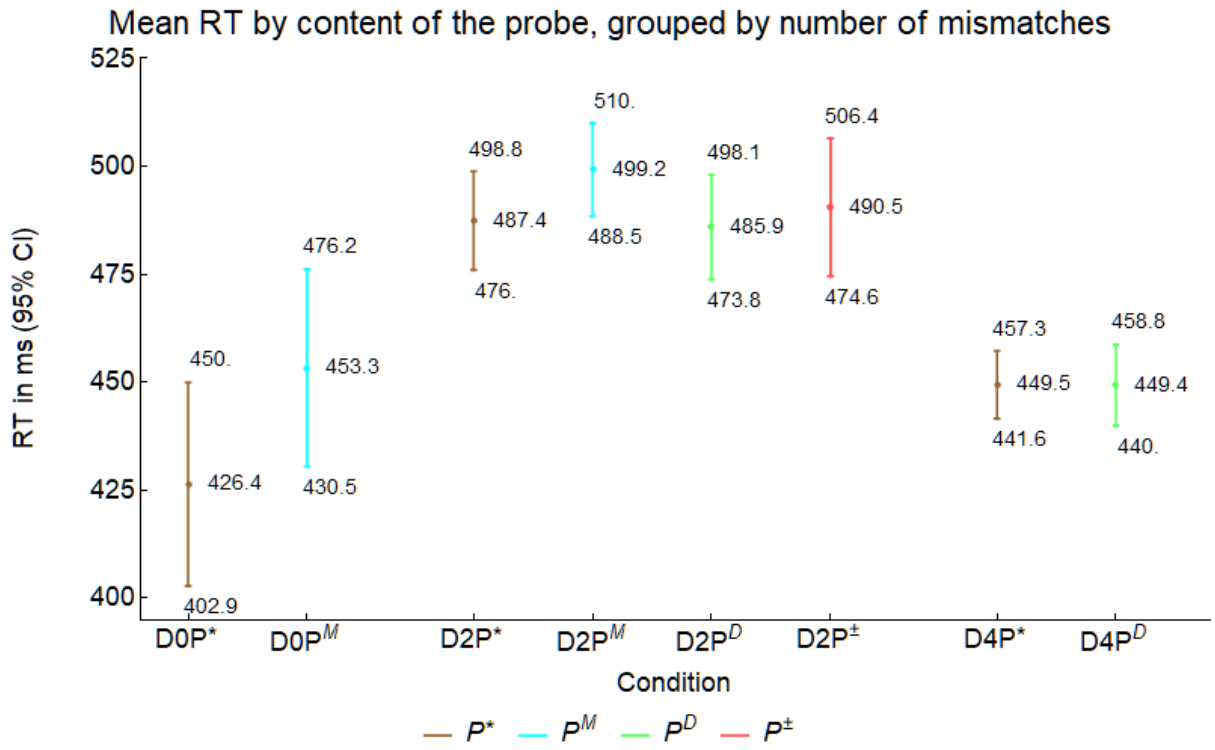

Figure 9: Mean and 95% confidence intervals of RT for trials with 4 letter stimuli only, depending on the content of the probe, grouped by the number of mismatches.

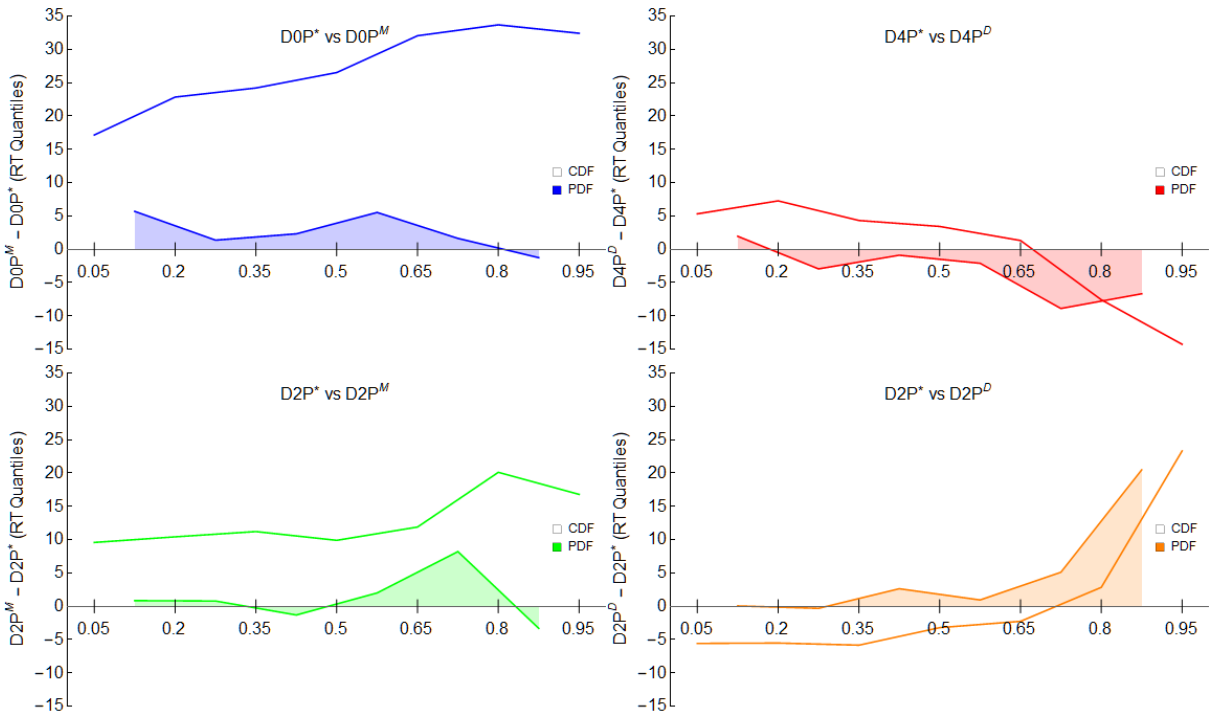

Figure 10: Averaged CDF and PDF differences between pairs of conditions. To obtain these values, I first calculated the 5th, 20th, 35th, 50th, 65th, 80th and 95th percentiles of each participant for each condition. To obtain the CDF difference, I subtracted the percentile values between conditions for each participant, then plotted the means of the differences for each condition. For the PDF difference, I took the difference between each percentile for each participant and condition, which gave me the slope of the CDF (i.e., the PDF). I then calculated the difference in slopes between conditions and once again plotted the means of the differences. I used the middle point between the percentiles as the x-axis value. The differences plotted are the curves of the non-P* minus the curves of the P* trials. The interpretation of this figure is summarized in Table 5.

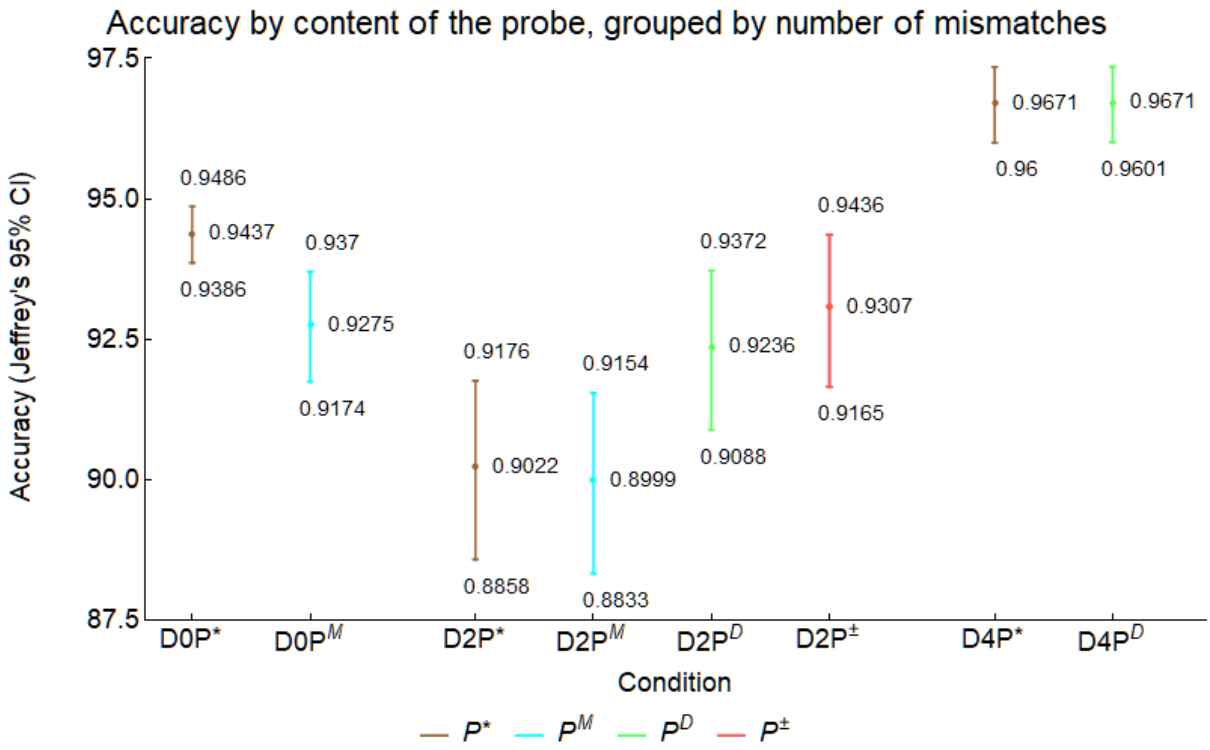

Figure 11: Mean and 95% Jeffrey's confidence intervals of accuracy for trials with 4 letter stimuli only, depending on the content of the probe, grouped by the number of mismatches. If two conditions have their means outside of the other condition's confidence interval, the difference between the means is statistically significant.

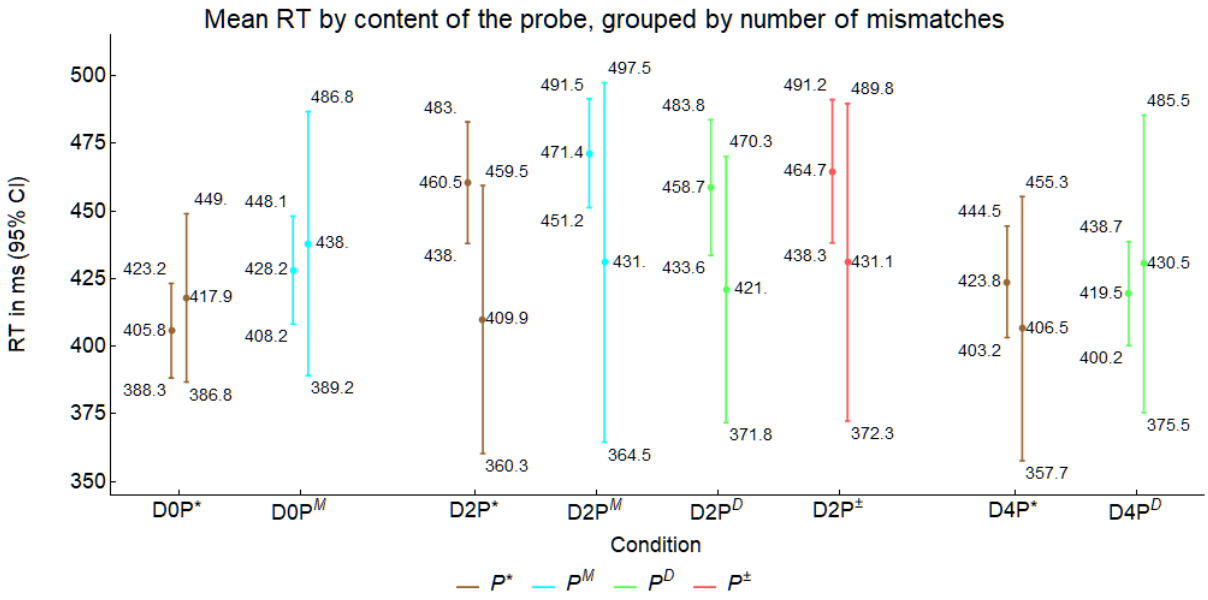

Figure 12: RT of correct and incorrect trials for each of the 8 conditions, including only participants that made at least 1 error in each condition. The mean RTs in this figure are faster than those in Figure 10, which means that participants with a higher accuracy also took more time, a typical speed-accuracy trade-off. The only significant results are that the following answers are faster than all correct *D2* answers: correct and incorrect *D0P**, correct $D0P^{M}$, correct and incorrect *D4P** and correct $D4P^{D}$ answers. The correct *D2P** answers are also faster than all correct *D2* answers excepted for correct $D2P^{D}$ answers. The difference between this pair is barely non-significant and is very likely to be significant from a theoretical point of view.

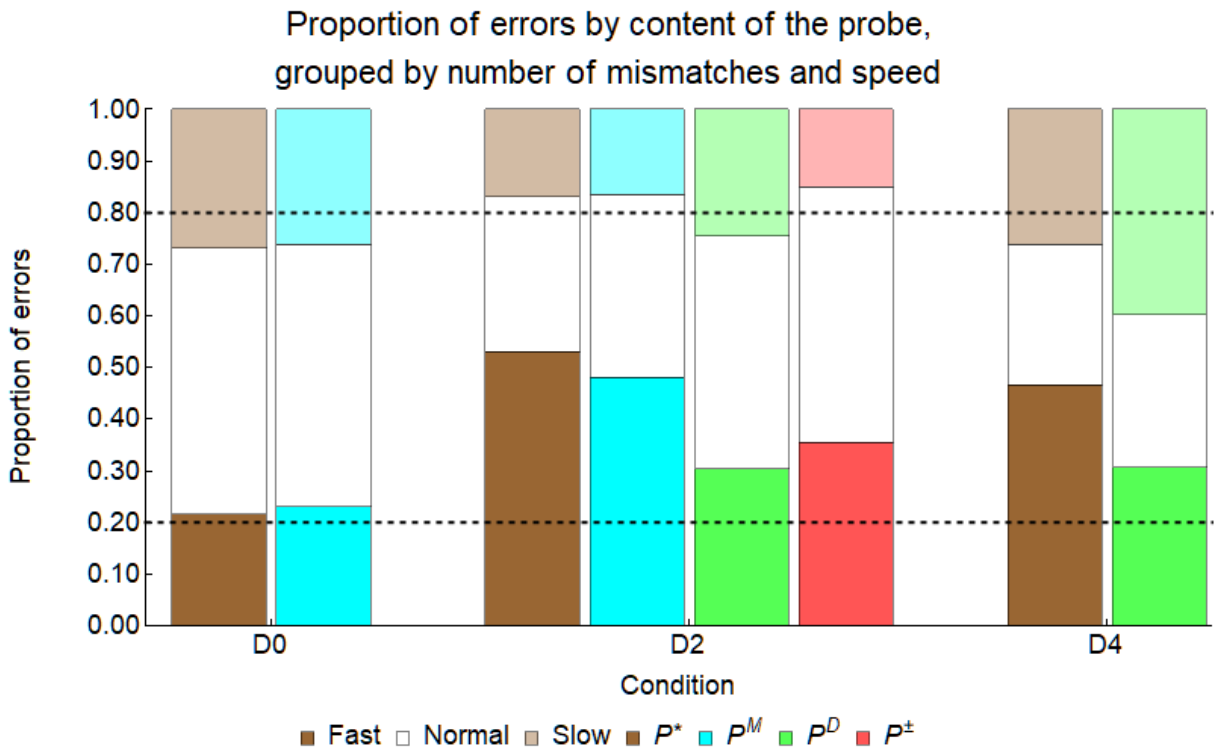

Figure 13: Distribution of the errors for trials with 4 letter stimuli only, depending on the content of the probe and grouped by number of mismatches and the speed of the error. There are no error bars for these values as this graph is simply a visual representation of count data found in Table 6.

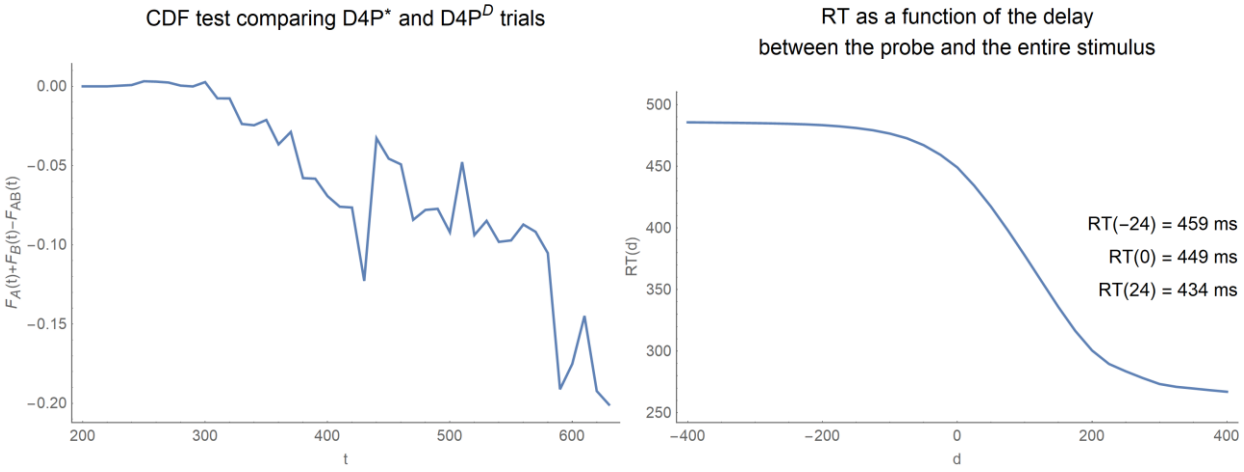

Figure 14: Left: Result of the CDF test using Equation 3. The value being almost exclusively below 0, this means that there is no redundancy effect in the “different” trials. Right: Result of the SOA test (Equation 4). Comparing this curve to Figure 8 reveals that the process responsible for these answers probably has limited capacity.