Modelling Naturalistic Visual Working Memory Using Action Thresholds

Author: Manouk Sirag (6716334)
m.sirag@students.uu.nl

Supervisor: Surya Gayet
s.gayet@uu.nl

Second Examiner: Leendert van Maanen
l.vanmaanen@uu.nl

Utrecht University
Master Artificial Intelligence
February 21 2025

Universiteit Utrecht

Abstract

When performing everyday tasks, we make use of our visual working memory (VWM) to temporarily store the information we see. In this study, we have designed the TCC-AT model which can predict naturalistic behaviour where VWM is used with only two parameters: memory strength (how well something is remembered) and action threshold (is the confidence in memory high enough to commit). We also designed and performed an experiment where people used VWM to remember colours and were tasked with reporting on those colours. For every memory task performed participants had to decide if they were confident enough in their memory to commit, either losing or gaining points depending on how well they remembered the colours. By making groups of participants remember a different number of objects at the same time, and making the strictness of the scoring different, the behaviour reflected different memory strengths and action thresholds in the TCC-AT model. The results showed that the model was consistent in predicting memory strength, but not in predicting action threshold. Multiple factors in the model and experiment were identified which could cause the problems in predicting action thresholds.

1. Introduction

When performing everyday tasks, people maintain an internal representation of the objects they need to remember, even when they are not actively looking at them. This is a vital part of our everyday functioning, since without it we would not be able to remember the position and orientation of objects in our direct surroundings. An example of where we use these internal representations, and how challenging our lives would be without them, is simply crossing the road. WIthout the internal representations of vehicles, their locations and their speeds it would be near impossible to cross the road safely, since it would only be possible to know the traffic on one side of the road: the one actively being looked at. The internal representations allow us to capture traffic in the road without having to actively look at it constantly.

The task of maintaining these representations is performed by our visual working memory (VWM). VWM has been studied extensively over the past decades, often focussing on its maximum capacity. A more recent addition to this is the target confusability competition model (TCC) (Schurgin et al., 2020). Simply put, this model can predict the performance of people when they have to use VWM for a range of different tasks using only a single free parameter. An example of such a task, which was also used to test the TCC model, is a continuous colour report task where people had to first remember a certain number of colours, and then after a short delay recall and report one of those colours. An important factor of this model and the task used to test is, is that it is based on forced-choice. This means that participants are forced to make a decision on which colour they saw, regardless of how well they remembered that colour or even if they did not see the asked colour at all.

This model however does not take into account one important factor in our behaviour in these situations: in the vast majority of times we use VWM the outside world will remain available to us. This also means that the required information remains available, which is in contrast with typical VWM experiments. Because the information remains available, use of VWM is not limited to reporting the colour you think you saw, but also making the decision on whether or not to report in the first place. Alternatively, a decision could be made to look at the initial colours again. This discrepancy is vital, since it means the model only addresses a small number of ways VWM is used in a person's daily life. This makes its application limited when used to predict or guide human behaviour in tasks.

We designed a new model as an addition to the original TCC model, the TCC-AT model, to take into account the natural way VWM is used daily. This new model takes into account that there is a point at which people are confident enough in their memory to report their answer. Not reaching this point might mean they will choose to look at the object they are trying to remember again, or if that is not possible either make a best guess or refrain from responding altogether. This model attempts to capture this complex pattern of behaviour using two parameters. The first one is 'memory strength', which was taken from the original TCC model. It represents how strong someone's memory is, how well they remember objects. The second parameter, which is new to this model, is the 'action threshold'. This parameter represents how strong someone's

memory needs to be to use that memory instead of taking the time to recommit it to memory first.

The main question we wanted to answer in this study was: can the TCC-AT model with ‘memory strength’ and ‘action threshold’ be used to model human behaviour in a task without forced choice. Along with this, during the process of analysing the data and the model results, we also questioned whether the model parameters ‘memory strength’ and ‘action threshold’ are independent of each other. If this is not the case, we wanted to know in what way the two variables are dependent on each other, since identifying that can aid in improving the model in the future by incorporating this dependency.

To achieve this, we first had to make the TCC-AT model by altering the existing TCC model. At the same time, to test this model we also needed behavioural data for which an experiment was designed and performed that provided data that would allow us to manipulate the two parameters independently.

2. Theoretical Background

2.1 TCC Model

Visual working memory (VWM) is used to maintain an internal and accurate visual representation of objects, even when they are no longer in our sight (Baddeley & Hitch, 1974)). An example of this is when you cross a road. To ensure a safe crossing, you need to know what is going on both to the left and the right of you, and since it is impossible to look in both directions at the same time, important information of the traffic situation on your left can be stored in VWM while you look right.

This process is not as simple as storing everything that was seen, that would be too costly on cognitive resources and a lot of information we see is not important to the situation we are currently in. When you need to remember the traffic situation on your left, what the trees look like along the side of the road is not important even if you did see them when looking left. We therefore only store a small amount of relevant information in VWM that we need to complete the goal we set when we started looking. In this case that could mean only storing the locations of the cars coming from the left while disregarding the pedestrians walking along the sidewalk.

Many models have attempted to explain and model this working memory, to understand how we store objects in our working memory, and how much we can store. A lot of different reasons for developing these models exist, like predicting how people will behave in certain circumstances, designing certain tasks and user interfaces and also simply to gain a better understanding of how cognitive functions work.

Traditionally, VWM was thought to contain a fixed number of slots, a discrete number of objects could be remembered and nothing more. Objects could thus be categorised in only two ways: either they were remembered or they were not, there was nothing in between. If an object was not remembered, it could only be guessed and if it was remembered, no guessing of any kind was needed.

A more recent view is the resource model of working memory (Bays et al., 2009). These models do not state that VWM has a discrete number of slots for objects, but instead that it has a fixed amount of resources. These resources can be shared across multiple objects, meaning that some objects might have more resources allocated to them and thus be remembered more precisely than other objects with fewer resources dedicated to them. How well an item is remembered (how many resources are allocated to it) is reflected in the precision of its representation, and thus in the precision with which it can be reported upon request. Along with precision, these models also have a guess rate, which is how likely someone is to guess an item. The reason this guess rate needed to be added, is because experiments where participants needed to recall objects from memory showed that there were errors that precision alone could not capture properly. A model that took into account that the resource model still contained objects that had not been remembered at all (where people had to guess), could capture behaviour much better. These models therefore model working memory with two parameters, precision and guess rate.

The TCC model extends on this resource model by showing that behaviour of participants in a forced-recall experiment can, in fact, be well explained using a single parameter called ‘memory strength’ (d’). This was a change to previous models where a ‘guess rate’ was also necessary. The researchers who created this model determined that errors that older models would classify as ‘random guesses’ could actually be predicted by a set of constants in the TCC model.

These constants used to determine the random guesses were determined using experiments where participants had to compare colours at different distances. An example of such a task is a ’fixed-distance triad experiment’, where participants were shown a target colour along with two additional colours. They then had to determine which of the two colours was closest to the target colour. This experiment resulted in a measure that showed how similar any colour on the colour wheel is to any other colour on that wheel based on how far removed they are from each other, regardless of which specific colours are being compared. As long as the distance is the same, the similarity is the same. A key aspect identified here was that this similarity spreads non-linearly. If the two colours shown are both far removed from the target colour, then it is more difficult to determine which of the two is closer on the colour wheel than when the two colours are close to the target colour. For example: if the two colours shown are 120 and 150 degrees removed from the target then participants could not tell which of the two is closer. If the two colours were only 5 and 35 degrees removed from the target then this task became trivial, while the distance between the two colours remained 30 degrees. The way this similarity spreads is constant and was used to replace the ‘guess rate’ from older models since it can be used to predict the ‘guesses’.

With the ‘guess rate’ of older models removed, only ‘memory strength’ is left. ‘Memory strength’ is based on familiarity, a measure of how well an object is remembered. Before any colour is seen, familiarity for all colours is the same. When seeing a colour to remember, the familiarity of that colour increases. However, when for example the colour red is seen and remembered, it is intuitive that colours close to it, like orange, will also have a significant increase in familiarity while colours much further away, like blue, will have (almost) no increase. The way this familiarity spreads over a stimulus space from the target is called the psychophysical similarity.

Since psychophysical similarity is not a physical property of colour but rather a subjective property that is different for everyone, an experiment was used to get a function that approximates this spread over a larger group (Schurgin et al., 2020). When a colour needs to be remembered, the psychophysical similarity is used to determine the familiarity of every colour in the stimulus space. The differences in familiarity (and therefore the subjective familiarity space) is different for each colour, but relatively constant between people. This similarity however is not the only factor that affects familiarity. Colours that are very similar to each other are easily confused, a phenomenon that is caused by perceptual noise, where the colour that was seen becomes slightly corrupted in the process of storing the information in memory. This perceptual noise is what causes incorrect answers and confusion. When memory strength, perceptual noise and the familiarity of each colour with the target are all taken into account, the colour that will be chosen is the colour with the highest familiarity.

One thing that does differ between individuals according to this model is their aforementioned memory strength. What this exemplifies is how the familiarity increases when a colour is sampled. A high memory strength means that the colour actually sampled has a higher increase in familiarity. Since there is a fixed amount of resources available, the area underneath the familiarity distribution has to stay the same, meaning the distribution of familiarity becomes very narrow and steep, decreasing much faster outward from the sampled colour.

What this relationship between memory strength and the shape of the familiarity distribution means is that apart from being different between individuals, memory strength is also related to how much VWM capacity (resources) can be dedicated to the colour to begin with. For example, if three colours need to be remembered, then the familiarity for each target colour will increase less than if only a single colour needs to be remembered since the resources are spread out more. Because of this, as more targets need to be remembered, the memory strength for the targets will be lower.

2.2 Naturalistic Memory Usage

A lot of studies and models (among which also the TCC model) have focused on the maximum capacity of VWM, which is done by asking participants to maximally load the VWM and then maximally testing their memory to see how much of the information is actually contained in VWM (Ma et al., 2014). This however is not the most common way people use VWM, maximum capacity is not used but instead a person will minimise the amount of resources used to what is needed to complete a task.

The TCC model, and the underlying theory, do not account for this naturalistic aspect of memory use that people are rarely forced to remember something or to use the information that they remembered. In the vast majority of cases where VWM is used, the external world with the object is still available to look at. For example when an individual is shopping, they will read the grocery list and pick up some of the groceries, but there is nothing stopping them from looking at the list again instead of going purely off memory. When wishing to achieve a goal, a choice can thus be made to either use the information stored in VWM, or to resample the external world for the required information instead. The tipping point between using VWM and resampling the external world can be regarded as an action threshold. The existence of such a tipping point makes the traditional model of VWM that assumes that VWM is always used maximally less valuable when applied to natural human behaviour (Ballard et al., 1995, Van Der Stigchel, 2020). Studies have shown that, when given the choice, people instead rely very little on the imperfect contents of VWM and prefer to resample the external world to strengthen the information (Ballard et al., 1995, Draschkow et al., 2020, Somai et al., 2019*)*. A trade-off happens where in many tasks, the cost of storing something in VWM is higher than the cost of sampling from the external world more often. (Ballard et al., 1995). This trade-off was also shown in studies where the cost of sampling was experimentally manipulated. When sampling the information was made more expensive (i.e. more effort was required to reach the information), the external world was sampled less frequently and thus more VWM was used, even if the information in VWM was considered imperfect (Ballard et al., 1995, Draschkow et al., 2020, Somai et al., 2019*)*. To illustrate, if the shopping list is on a phone with a password lock instead of on a piece of paper, someone may rely more on memory (i.e. remember more grocery items at a time) since it takes more effort to type in the password every time to re-check the list. A model that predicts naturalistic behaviour would therefore need to be able to predict both how well a colour is memorised and how likely it is that it will be used.

To also capture the above mentioned naturalistic behaviour and to determine why an individual resamples, an experiment was conducted where participants performed a memory task where they could choose to resample the target when they wished (Sahakian et al., 2023). At random points in the experiment however, participants were forced to use VWM and recall the target. This showed that even when participants had decided to resample, they still had VWM content they could use. What this means is that people do not use all VWM contents before choosing to resample.

This experiment also changed the cost of sampling the target between participants. Some participants could resample immediately while others had to resample on a delay, making it more expensive to them. This showed that when sampling becomes more expensive, participants will resample less often but look at the external world for a longer period of time. After resampling, these participants could recall more from VWM and critically they also made more attempts to recall items even if they had started making mistakes. This means that when sampling is cheaper, people will remember fewer things as opposed to people who have to deal with a higher sampling cost. This indicates that people can alter to what extent they will use VWM depending on the circumstances surrounding resampling the external world.

The study by Sahakian et al. (2023) therefore proposes that aside from memory strength, everyone also has a context-dependent action threshold, which represents the point at which the decision is made to use VWM as opposed to resampling the external world. This threshold represents how familiar an individual needs to be about an object in VWM before they actually use it. Where the original TCC model uses the memory strength to alter the familiarity space where a choice is forced, action threshold could represent a limit in this same familiarity space that determines if a choice is made or not. This makes the model containing 'action threshold' as well more valuable in research or applications where naturalistic behaviour is needed and/or expected. Any object in VWM whose familiarity reaches the action threshold gets recalled, anything below it does not get used. If nothing in the familiarity space reaches the action threshold then the external world will be resampled instead to correct uncertainty and increase familiarity. This concept of an action threshold is proposed to be the key to reconciling the traditional view of VWM and the naturalistic use of VWM. It can be used to achieve the goal of predicting complex naturalistic VWM-driven behaviour through a simple model with two parameters: 'memory strength' and 'action threshold'.

3. Methods

3.1 Extended model

3.1.1 TCC Model

To implement the TCC-AT model, the open-source code of the TCC model was used as a starting point (*Schurgin et al., 2020* *original model code*). The TCC model already has memory strength implemented, so what was needed was to implement an action threshold in this model as well in a similar way as memory strength. To achieve this, the way the TCC model is implemented had to be understood before the action threshold could be added. Firstly, an important aspect of the TCC model is that the model is pre-trained for each possible value of memory strength. What this means is that for each possible memory strength, a probability density function is calculated that indicates the probability of each possible error for that memory strength.

As mentioned above, the training for the TCC model was done for each possible memory strength (d'), which was set at a predetermined range. The colours used in the memory task were drawn from a circular space (a colour wheel), so that each colour could be represented as a point on a 360 degrees circle. For each of these 360 colours, the psychophysical similarity was multiplied by the current d'. This psychophysical similarity was represented as 360 numbers. These represent possible distances from a target colour, from -179 to 180, and how similar colours at that distance were perceived as. The multiplication of psychophysical similarity by memory strength gives the mean familiarity of each location in the -179 to 180 range, which is highest in the centre (0 degrees error, the target) and radiates out.

The goal at this point is training the model, and since the familiarity spaces we're using contain randomness (due to noise), we could not simulate a single familiarity space and use it to calculate probabilities. Instead many familiarity spaces had to be simulated for each memory strength which could be combined to get a probability density function for each memory strength that properly represented the distribution of noise.

Familiarity spaces were generated using the mean familiarities and noise according to a perceptual correlation matrix. The perceptual correlation matrix is 360 by 360 and indicates for each combination of colours how likely they are to be mistaken for each other. The result of each stimulation was a vector of 360 numbers, each representing familiarity of a colour that is a specific distance removed from the target colour in the centre. A total of 100.000 familiarity spaces were simulated using this method. Due to the addition of noise, familiarity was probabilistic. Locations on the familiarity space closer to the target had a greater chance of containing the highest familiarity since the mean familiarity at those locations was higher to begin with. Through this method, memory strength, psychophysical similarity and the confusion of similar colours were represented in the simulated data.

For each memory strength value, the process of generating 100.000 familiarity spaces was repeated. A lower memory strength means that the distribution of the mean familiarities (before noise was added) was much flatter and therefore that the values overall are lower. What this means is that when the noise was added in the simulations, it was more likely that locations further away from the target contained the highest familiarity. What this intuitively means is that since the amount of noise remains the same, a less precise internal representation means there is a higher chance of mistakes being made. The opposite would apply to higher memory strengths, since they would have much higher peaks relative to the amount of noise and thus it would be less likely that the noise would affect the familiarity values enough to change the location where the maximum familiarity was located.

For each memory strength value, we now had 100.000 familiarity spaces. The location of the maximum familiarity in the familiarity space was determined for each of these simulated familiarity spaces per memory strength, which gave 100.000 numbers indicating the distance from the target that would be picked for each simulation. Most of these values would be close to the target, but again due to noise in the simulation process, some simulations would have results further from the target. The 100.000 errors were then used to generate a probability density function (PDF) representing the probability of each error for that memory strength. This PDF is also a vector of 360 numbers, representing distances from the centre target (the aforementioned -179 to 180).

The memory strength of an individual could then be determined by calculating which of the PDFs for the different possible memory strengths was most similar to the errors seen in the participants' data. This was done by calculating the log-likelihood of the PDFs for each memory strength: getting the probability of each error in the participants' data, taking the log of each of those probabilities and then calculating the sum. The highest log-likelihood represents the dataset where all the errors together have the highest probability. The corresponding PDF is

therefore closest to the errors of that participant and the memory strength for that PDF is the memory strength of the participant.

3.1.2 TCC-AT Model

To extend the TCC model to include the action threshold, the new model had to train on, and thus simulate, not only a range of memory strengths, but also a range of action thresholds. To achieve this some small alterations had to be made to the way the probability density functions were calculated. After the 100.000 maximum familiarities were calculated, every familiarity that is underneath the current action threshold was removed from this list of 100.00 maximum familiarities. These excluded familiarities represent the choices that are not familiar enough. When translating this back to how it corresponds to human behaviour, these would be instances where a choice to resample would be made as opposed to reporting on VWM content. This intuitively means that when the action threshold is 0 (and thus a response is always provided), the extended model has the same PDFs as the original TCC model without the action threshold.

The remaining errors that did reach the action threshold would then be used to generate a PDF that is normalised on the original 100.000 errors retrieved from the simulations. This was done to ensure that the number of times VWM was used as opposed to not was represented in the PDF. Ideally, this would create unique unnormalised PDFs for each combination of memory strength and action threshold. The PDFs would be unique since when the memory strength remains the same but the action threshold changes, different amounts of errors would be removed.

When determining the memory strength and action threshold of an individual, two different methods were tested. The first method is the same as in the original TCC model, by calculating the log likelihood of each possible PDF and using the action threshold and memory strength of the highest scoring PDF. This method was quite slow, since it had to search through a 2D matrix of PDFs instead of a single vector. An alternate method was made where first the best fitting PDF was calculated, assuming an action threshold of 0. This provided a memory strength value. Then, the best fitting PDF was calculated among the PDFs with the already determined memory strength, providing a corresponding action threshold. Instead of searching through the entire matrix, this method only required two searches in two vectors of that matrix (a row and a column).

The range of numbers at which memory strength was trained was kept the same as in the original TCC model, a linear scale from 0 to 4.5 with steps of 0.03. Since the action threshold is in the same space as memory strength, this also had to be in a similar range. To make sure the analysis included all theoretically possible action threshold values, a range from 0 to 8 was chosen. Determining the maximum familiarity value of 8 was done by using the simulated data from the training process of the model. By taking the maximum familiarity from all simulated familiarities, we have an approximation of how high familiarity could theoretically go. Since this was just below 8, setting the maximum action threshold at 8 means that an action threshold at both extremes of the spectrum was possible (always using VWM and never using VWM). After

preliminary analyses of the PDFs, it was determined that lower AT values have barely any effect on the PDFs (*extended figure TBA*). Because of this, a logarithmic scale was also tested. When discussing model results and analysis, the model trained with a linear AT scale will be called V_1 and the model that has been trained with a logarithmic AT scale will be called V_2.

3.2 Behavioural Experiment

3.2.1 Experiment goal

An experiment has been designed to validate the TCC-AT model. This means validating that the model could distinguish between differences in both memory strength (low versus high) and action threshold (low versus high). To achieve this, two manipulations had to be made in the experiment. To measure differences in memory strength, the set size was made variable, which means participants had to remember either two or three colours at a time. When three items had to be remembered, cognitive resources had to be spread out more resulting in a lower memory strength compared to when two items had to be remembered, as explained before. This is the same manipulation as was done to test the original TCC model and proved to be a good way to show differences in memory strength.

To be able to measure differences in action thresholds, something had to force participants to be more or less strict about the decision to use their memory content or not (given a constant memory strength). To achieve this a new experimental paradigm was conceived in the form of a scoring system, where different levels of punishment were used (high or low punishment). In the low punishment condition, a larger range of errors would give points and a perfect answer would gain more points compared to the high punishment condition. Similarly, in the high punishment condition a larger range of errors would cause point deductions and the maximum points that could be lost in a single round is higher. Participants of the high punishment condition of the experiment had to be more cautious about which answers they committed, changing their action threshold. This created an experiment with two manipulations: set size and punishment. With two possible set sizes and two possible punishment conditions, this means there were four distinct experimental conditions. With this, the experiment could be used to determine whether the TCC-AT model is able to show (and measure) differences in memory strength and willingness to respond (action threshold). As a secondary goal, it could also be used to provide evidence for whether memory strength and action threshold are independent of each other, and if not in what way they might relate to each other.

3.2.2 Participants

The experiment was designed to be between-subjects, which means that every participant performs only a single experimental condition (a single set size combined with a single punishment condition). This was done as opposed to a within-subject design (every participant does all conditions of the experiment) to prevent the effects of learning. When a single participant performs multiple conditions of an experiment, then the results of the first condition will be normal, but results of all conditions after that will be affected by the fact that participants

have already learned and extensively practised the experiment. In this experimental design, four participant groups needed to be filled. The four participant groups are referred to by their conditions. The L2 group refers to the participant group in the low punishment condition with a set size of 2, The L3 group has the low punishment condition with a set size of 3. H2 and H3 refer to the same thing for the high punishment condition.

Participant recruitment was done through the online platform Prolific. Prolific's built-in screening tools were used to ensure we only included participants who are not colour blind or dyslexic, between the ages of 18 and 35 and with an approval rate of 95-100%. The approval rating is a feature of Prolific, it represents the percentage of studies a potential participant has performed that have been approved. Researchers can formally reject a participant if the participant did not do the experiment properly on purpose, for example if they simply clicked through all the answers to get the payment, without actually reading and answering properly. Only using participants with a high approval rate prevents participants who do not take the experiments seriously from taking part in the study.

We set out to collect usable data from 80 participants in total, 20 per condition. Three exclusion rules were set: (1) participants who experienced an error during the experiment (2) participants who had a 100% or 0% commit rate, or only a single committed trial (3) participants whose mean absolute error of the committed trials is not significantly better than the mean absolute error of non-committed trials. The first rule was set because participants who experienced an error would have either no data at all, corrupted data, or data affected by the error which would make it unusable. The second rule was added because participants who always chose the same option (or only committed once) either failed to understand the instructions, or performed so badly that virtually no stimulus was remembered with sufficient precision to warrant committing the response. Aside from that, participants with a 100% commit rate, a single committed trial or no committed trials at all could not be used to calculate the statistical significance required for the third criterion. This final criterion was made to ensure participants displayed the intended behaviour. Participants had to only commit their trials if they had some certainty that they might gain points from it. If these trials were not statistically better than trials where the participant did not commit, it could indicate that they might not have understood the trial properly, or that there might have been some other reason why they could not perform the experiment as intended (for example floor or ceiling performance). This statistical significance was calculated using a permutation test with a significance level of 0.05. These criteria would also serve to exclude participants who did not try properly but clicked through the experiment. All participants who completed the experiment (regardless of whether their data was included or not) were paid 5.25 GBP for approximately 35 minutes of work. Additional participants were added to the study until there were 20 participants per experimental condition who met all inclusion criteria.

3.2.3 Procedure

The experiment procedure consisted of instructions on how a trial works, practice trials to ensure participants understood the experiment and an exit question. First, the mechanics of a

single trial will be explained in detail, afterwards the entire experiment procedure will be elaborated upon.

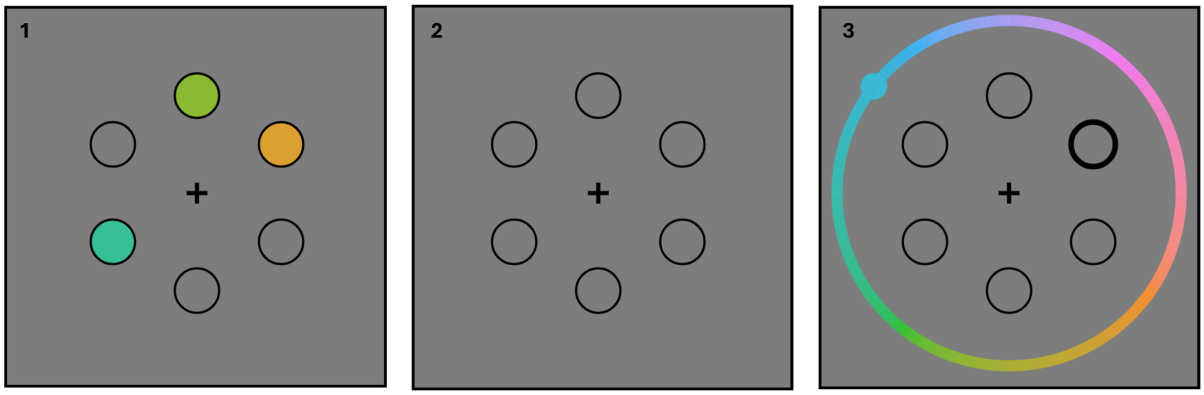

Fig 1: Outline of the first section of a trial (here with set size 3) where participants see the colours they need to remember (1), then have a short delay (2) and finally they get indicated which colour they need to recall and they can report it on the continuous colour wheel (3)

A single trial consists of two sections. In the first section, participants had to remember colours shown on the screen (memory items) and then recall one of those colours. In the second section, participants had to make a decision on how confident they were in their performance during the first section.

The way the memory items were shown in the first section was with placeholders. The centre of the screen contained six equally spaced placeholders (black circles) around a fixation point (a cross) (*fig 1.1)*. The placeholders where the memory items are shown were picked at random to ensure that participants cannot look at specific places on the screen but had to focus on the entire screen to see all the memory items. Placeholders where no memory items were shown remained empty. When a trial started, participants were shown either two or three memory items (depending on set size) for 100ms. The colours then disappeared and a delay time of 800ms was added to ensure participants had to remember the colour and could not rely on an afterimage. During this delay, only the fixation point in the centre of the screen and empty placeholders were still present (*fig 1.2*). After this, a continuous colour wheel appeared on the screen and one of the placeholders that displayed a memory item before was indicated by making its outline bold. Participants had to indicate on the continuous colour wheel which colour they believed was shown in the bold placeholder at the start of the trial. Participants did this by clicking on the colour wheel. A coloured circle that moved along the colour wheel with the cursor indicated to the participant which colour would be selected if they clicked at that moment (*fig 1.3)*. The memory item that was probed was chosen at random from the 2 or 3 memory items shown at the start of the trial.

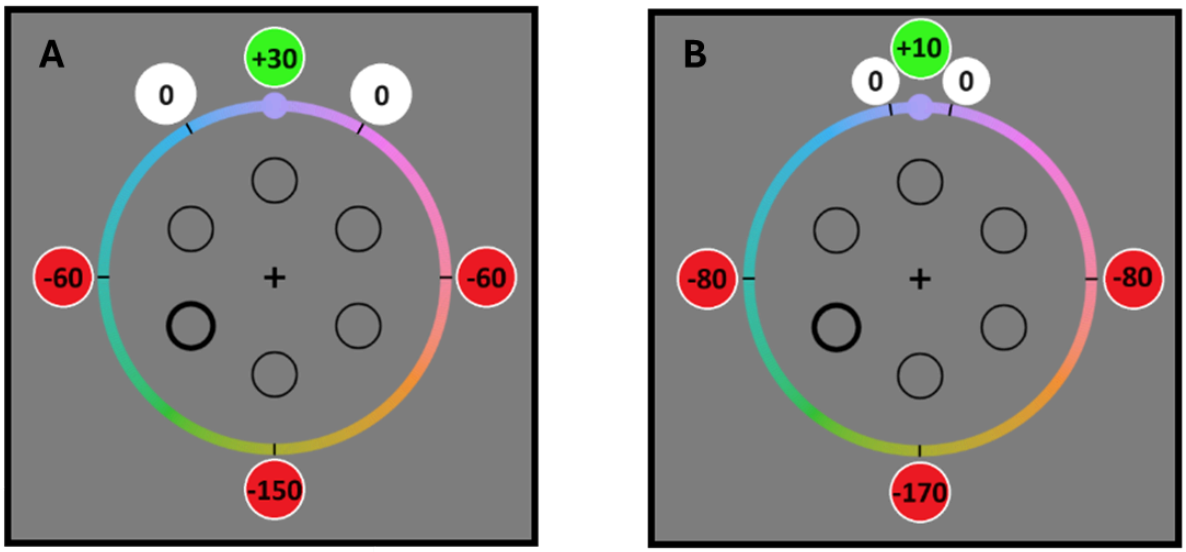

Fig 2: The scoring wheels for both the low punishment condition (A) and the high punishment condition (B). These are the same as used in the instruction screens for the experiment.

The way scoring worked was dependent on the punishment condition, either high or low punishment. In the high punishment condition, the maximum number of points scored for a perfect answer is 10 (*fig 2 B*). For every single degree increase in error, the score goes down 1 point. Since there are 360 degrees, the lowest score a participant could get was -170 for answering on the opposite side of the colour wheel. In the low punishment condition, the highest score possible was 30 (*fig 2 A*). The way points dropped was the same, so for this condition the lowest score possible was -150. If someone has an error of 15 degrees for a trial, this would mean that in the high punishment condition they would lose 5 points (10 - 15 = -5). In the low punishment condition they would gain 15 points (30 - 15 = 15). This indicates the point where different action thresholds would be expected, since in the low punishment condition it would be advantageous to commit the answer while in the high punishment condition it would not be advantageous.

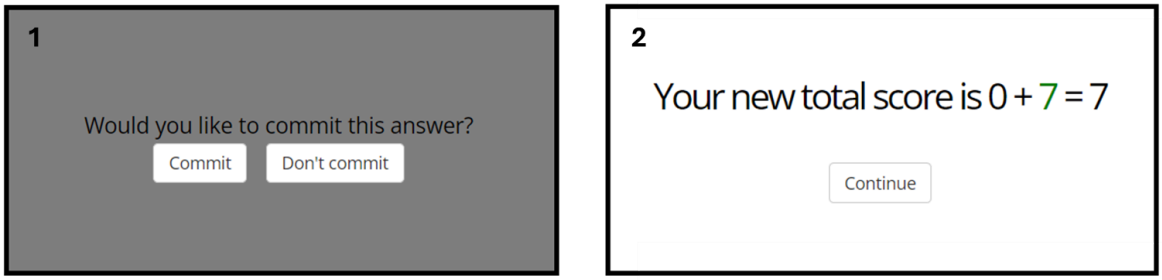

Fig 3: Outline of the second section of a trial, where participants could choose whether they wanted to commit an answer or not (1) and then how their total score changed (2). In this example, a screen for a committed trial is shown where the trial score is +7. When a score is not committed, the total score will not change (the screen shows the total score with 0)

Once participants reported the colour of the probed item by clicking on the colour wheel, the second section of the trial was initiated. Participants were given an option to commit their answer or not (*fig 3.1)*. Committing meant that the score for that trial would count towards the total score, while not committing meant that the trial score was discarded. Crucially, participants were not shown their trial score before committing, they had to decide based on how confident they were in their answer. This decision is where the action threshold can be determined, since the different punishment conditions would affect the confidence participants needed in their answer before they were willing to commit their answer. After deciding whether to commit or not, participants were shown how that trial affected their total score (*fig 3.2*). This would conclude the current trial and the experiment would continue to the next trial (or the experiment end if it was the final trial)

When a participant entered the experiment from Prolific, they were welcomed and asked to provide informed consent for the experiment and the collection of non-personalized data. When a participant agreed with this, they were first given the instructions. This was done in two steps. First the participants were shown a screen with explanations on how they would be presented colours to remember, how they should report the probed colour and how they could choose to commit their answer or not. This was followed by a single practice round where the colours were visible for a longer time to make it easier, but no decision to commit had to be made yet.
The second set of instructions explained how committing an answer affected the score. Participants in the high and low category had different instructions here, reflecting the strictness of their point system. Two additional practice rounds were then added with a special score screen. On this screen, participants were shown two coloured circles, the left one showing the colour that was probed and the right one showing the participants' answer. The way the scoring for this trial is calculated is explained below that, as well as how it would affect the final score. After this there were ten normal practice rounds, set up the exact same way as the actual experiment and with a total score being tracked. After the practice rounds, participants were informed that their score had been reset to zero and that the experiment would begin now that they understood the mechanics.

The experiment itself was 200 trials in which participants collected points. After every trial, they would see their trial score and how their total score was updated. Every 25 trials, a break screen was added where participants could also see the score wheel again. After all 200 trials were done, the participants were shown their final score and asked an exit question. This question showed nine images of score wheels, like the ones shown in the instructions and break screens but simplified.Participants were asked which of those images correlated with the scoring of their experiment. It was vital to ensure participants understood how the scoring worked and were able to remember how the scoring for their trials worked. This is the case because a crucial part of the analysis of the model revolves around participants with different punishment conditions using different strategies, which would only be possible if they knew throughout the entire experiment how strict their scoring was. The experiment was then ended and participants were directed back to Prolific

3.2.4 Apparatus and Stimuli

The experiment was programmed in JavaScript using the libraries jsPsych (version 7.3.0) and HSLuv (version 1.0.1). It was hosted on the online web service Gorilla. Participants were strongly recommended to use a desktop or laptop and all participants indicated that they did. Since many different kinds of devices and displays were used, it was assumed that participants had different screen sizes and that the experiment could look somewhat different for some participants. One way this effect was lessened was by determining the sizes of the elements on the page based on a scaling factor. This made sure the different elements on the page kept the same relative size to each other. This way, the experiment screen would not distort, even on smaller screens. Due to different display and brightness settings it was inevitable that between participants the same colour values would result in different monitor outputs. Since there was no achievable way to avoid this in our experiment, the effect this difference between participants might have had on the results of the experiment had to be considered when analysing the results.

The colours that had to be remembered were picked at random. The colours were picked from the HSLuv colour space. In this space, colours are made up of three components: saturation, lightness and hue. Simply put, saturation represents the intensity of a colour, lightness represents the brightness (the blackness/whiteness components of colour) and hue represents the dominant colour component. A continuous colour wheel was divided into 360 values, corresponding to the 360 possible integers for the hue value in HSL. The saturation was always set at 90% and lightness always at 70%. Due to the way HSLuv colours are generated, this means the colours that are an equal distance removed from each other on the 360 hue scale are perceived as equally different from each other. When randomly picking multiple colours (two or three, depending on set size) for a single trial, the colours had to be at least 15 degrees apart from each other to ensure participants are able to distinguish them. Other than that there are no requirements for the colours, the randomness along with the high number of trials (200 per participant) ensured that trials with both large and small differences in colour were represented. This is important to ensure the entire familiarity space is covered, so there is data to test all ranges of errors.

3.2.5 Analysis

Analysis of the data was done in twofold, first a general analysis of the experiment data itself independently from the TCC-AT model, then analysis of the TCC-AT model without the experiment data and finally also the TCC-AT model using the experiment data. The initial analysis was done using python (version 3.12.6) along with libraries Numpy (version 1.26.4), Pandas (version 2.2.2), Scipy (version 1.13.1) and Matplotlib (version 3.9.2). The testing of the model was done in MATLAB (version 2024b), since that is where both the original TCC and the TCC-AT model were programmed in.

The initial experiment analysis started with the computations of the metrics that were used for participant inclusion. First, that the errors of committed questions were significantly lower than the errors of non-committed questions. This was done with 1-sided Welch's t-tests as. Welch's t-test was used instead of standard Student's t-test since calculations showed that the assumption of equal variances does not hold in this data. This assumption is necessary for an accurate result of the Student's t-test, but not for the Welch's t-test. The hypothesis tested was that the mean absolute error of committed errors was greater than the mean absolute error of non-committed errors. The alternative hypothesis was that the mean distribution of committed errors was lower. A significance level of 0.05 was used. Participants where the errors of committed trials were not significantly greater than the errors of non-committed trials were excluded from further analysis and replaced.

An important factor we wanted to determine before analysing the model is whether there were differences in performance and commit rates between the four conditions and whether these were as expected. For this, the report error was used. This error is the angular degrees difference between the reported colour (the answer given by the participant) and the presented colour (the requested memory item). The differences between the conditions were analysed by comparing the mean absolute errors as well as the standard deviations of the report errors for each condition. This was done both for all errors together and for committed/non-committed errors separately. We expected to see higher overall absolute errors in the trials where the set size was three as opposed to two, which would correspond to lower memory strengths for set size three when compared to memory strengths for set size two. Along with this we also expect that in the high-punishment trials, more trials were not committed overall, which would mean the action threshold of the high punishment condition is higher. This part of the analysis would allow us to determine whether the manipulations of set size and punishment condition had the intended effects on behaviour. This is a prerequisite for the model to be able to distinguish the two parameters before we start the analysis of the model using the data.

Since the model is pre-trained, the experiment data could be used to find a combination of memory strength and action threshold where the model would be best able to predict data from the same source. To verify the quality of the fits for the best fitting models, model statistics were used like the log-likelihood score and AIC. The log-likelihood score represents the probability that the model would predict the data provided to it. The higher the log-likelihood score, the more probable the data is using the parameters of the corresponding PDF. When comparing different models (specifically the TCC model and the TCC-AT model), the Akaike Information

Criterion (AIC) was used to determine which is best. AIC is a method that calculates a score using the maximum likelihood estimations (the log-likelihood scores), along with the number of independent variables. What this means is that it allowed us to compare the original TCC model and the extended TCC-AT model while taking into account the different number of parameters (only memory strength for TCC, memory strength and action threshold for TCC-AT). This is important because models with fewer parameters are preferred if the fits are otherwise the same. If adding the action threshold to the model does not improve the performance of the model, then the addition would not be useful. This provided a comprehensive view of how well the model fit the behavioural data, and whether varying different aspects of the model improved or worsened the performance.

A bootstrapping procedure was used to obtain variance estimates for all metrics of interest (e.g. confidence intervals for AT and d' values). The way this was done was by resampling the data of individual participants, and the total number of resamples used was 20.000. For each of these 20.000 runs, the errors from each participant were resampled with replacement, while keeping the commits the same. This resulted in 20 participants per condition like before bootstrapping, but with 200 report errors taken from each participants' original 200 report errors, randomly with replacement. Since this was done 20.000 times, the result of this process was 20.000 sets of memory strengths, action thresholds and log likelihoods corresponding to the experiment data (each based on the entire participant pool). By doing so, every metric of interest had a singular parameter estimate (the mean across bootstrap samples) as well as an associated measure of variability (the confidence intervals of the parameter mean).

Since the model parameters are both rooted in theory, their values also have meaning, especially when comparing the parameters of different datasets (e.g. conditions or participants). Aside from the fit of the model to the experiment data, it was also important that the parameters were in line with the underlying theory explained previously. If this were not the case it would mean that while the model fits the data in practice, we would not understand what the corresponding parameters actually refer to in human behaviour. The model could then still have uses in predicting performance, but it would not aid in understanding the underlying cognitive functions. What this means is that the action thresholds for the high punishment conditions had to be higher than the action thresholds for the low punishment conditions. If this was not the case, it meant the model did not correspond with the theory. While the same thing would count for the memory strength, we already know the model results correspond to the theory since the model's memory strength results were already extensively analysed and tested in the TCC model. If the differences between conditions matched with our expectations and theory, it meant that the model could aid in explaining cognitive function through memory strength and action thresholds.

4. Results

4.1 General analysis experiment

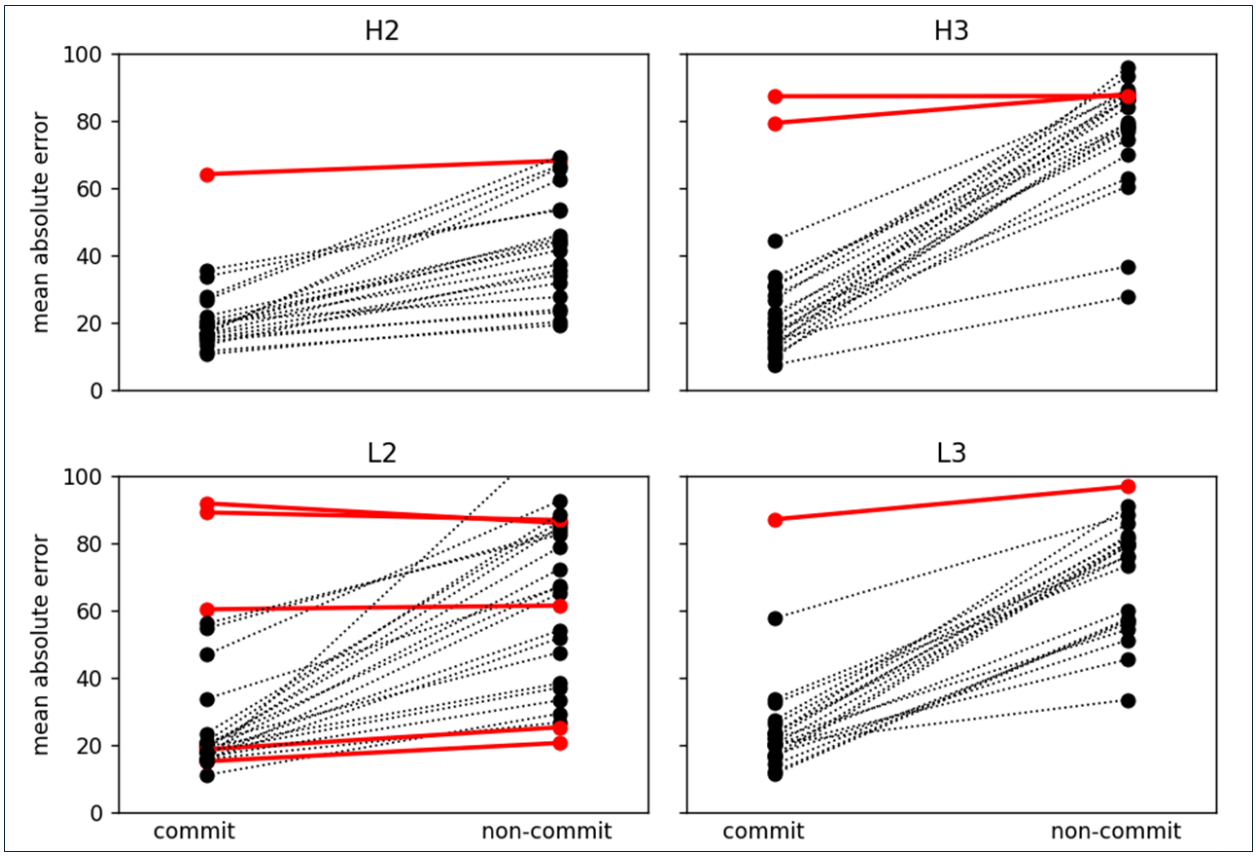

fig 4: For each participant, the mean absolute error for the committed trials is connected to the mean error of the non-committed trials. Lines in black represent participants who passed the third exclusion criterion, participants in red did not pass the criterion. Statistical significance is determined using a permutation test with a significance level of 0.05.

Before the analysis of the included participants can start, we had to decide which participants to exclude from this analysis. As mentioned before, this was done through three exclusion rules. Participants were added to the experiment until each condition contained 20 participants. The numbers below indicate how many participants overall were excluded per rule during the entire experiment before the required amounts were reached. Firstly, all participants who experienced some experiment error were excluded. This resulted in three participants being removed. Secondly, all participants who had a 100% commit rate, a single committed trial or a 0% commit rate, were removed as well. This step excluded another 10 participants. Thirdly, *(fig 4)* shows the analysis of the final exclusion rule, where the mean absolute error of the committed trials had to be significantly higher than the mean absolute error of the non-committed trials (using a

permutation test with a significance level of 0.05). In total, another nine participants were excluded through this method.

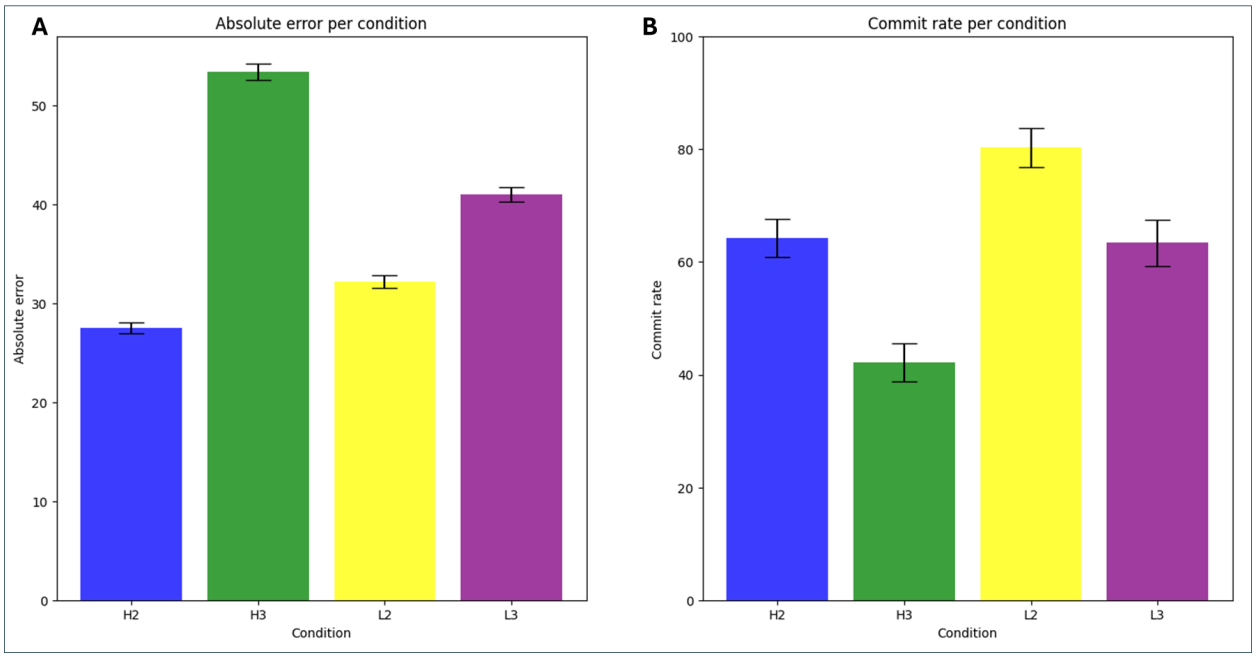

Fig 5: A) Mean of the absolute error across participants, separated by experimental condition. B) Mean commit rate (percentage of trials in which a response was committed) across participants, separated by experimental conditions. Error bars in both panels depict the standard error of the mean (SEM)

To check whether the manipulations in the experiment have the intended effect on the behaviour of participants, a two-way ANOVA was performed. This allows us to assess the statistical evidence for the hypothesised effects of set size and punishment on behaviour. We analysed the effect on both the absolute mean error per participant and the commit rate per participant. What we expected to see in participant behaviour was that the set size mainly affects the absolute error while the punishment level mainly affects the commit rate. In an ideal situation, this would have been shown through statistical significance, where only set size has a statistically significant effect on the absolute error and only the punishment has a statistically significant effect on the commit rate. The ANOVA analyses revealed that the punishment level did not have a statistically significant effect on the mean absolute error ($p = 0.23$), but the set size did ($p = 8.79e\text{-}7$). The second analysis determined that both the punishment level ($p = 2.02e\text{-}6$) and set size ($p = 7.52e\text{-}7$) have a statistically significant effect on the commit rate. This means that punishment only affected the commit rate, while set size affected both the absolute errors and the commit rate. Comparing this with our previously stated expectations, we do see the effect of punishment on commit rate and of set size on absolute error, but an unexpected outcome was the additional effect of set size on commit rate. This can be explained by the fact that a set size of 3 is more difficult than a set size of 2, meaning participants make greater errors and thus also commit less often.

An additional statistic the ANOVA analysis can provide is the interaction between punishment and set size. This measure interprets whether the effect of one independent variable (punishment) on the absolute error or commit rate changes depending on the other independent variable (set size), or vice versa. The interaction of the two previously mentioned ANOVA analyses showed that when it comes to absolute errors, there was a statistically significant interaction ($p = 0.01$), meaning that the effect of set size on the absolute error differs between punishment levels. The analysis however shows no statistically significant interaction ($p = 0.4$), meaning the effect of punishment on the commit rate does not depend on the set size. An important

4.1.1 Interim discussion

What the ANOVA analysis on absolute error along with (*fig 5 A)* shows is that for the absolute error, a set size of two provided lower absolute errors when compared to conditions with a set size of three. Punishment level however had no significant effect on the absolute error directly, but it did affect it indirectly by changing the effect of the set size. The ANOVA analysis on commit rate and *(fig 5 B)* show that for the commit rate, both set size and punishment have a direct effect, where a lower set size and a lower punishment level both increase the commit rate independently of each other.

We theorised that participants had lower errors in the low punishment condition compared to the high punishment conditions under the same set sizes (L2 vs H2 and L3 vs H3). The data does not support this, as it can be seen that the mean absolute error of L2 is actually higher than H2, while the mean absolute error of H3 is higher than L3. This is advantageous to the application of the model, since it provides evidence for the theory that while the differences in mean absolute error can indicate memory strength, differences in action thresholds will need to be related to whether or not errors have been committed.

For commit rates we theorised that a higher punishment would correlate to a lower commit rate, since it was supposed to force participants to adopt a stricter strategy when deciding to commit a trial, and thus commit less likely. This is seen in the behavioural data as well, as seen in *(fig 5 B)*, the higher punishment conditions correspond to a lower mean commit rate. What was not theorised however was the effect of set size on the commit rate, namely that the conditions with set size three also had significantly lower commit rates (at significance of 0.05). Because of these two effects on commit rates, conditions H2 and L3 have almost the same commit rate, which is unfortunate since it makes it more difficult to clearly distinguish between the conditions once we start analysing them with the model

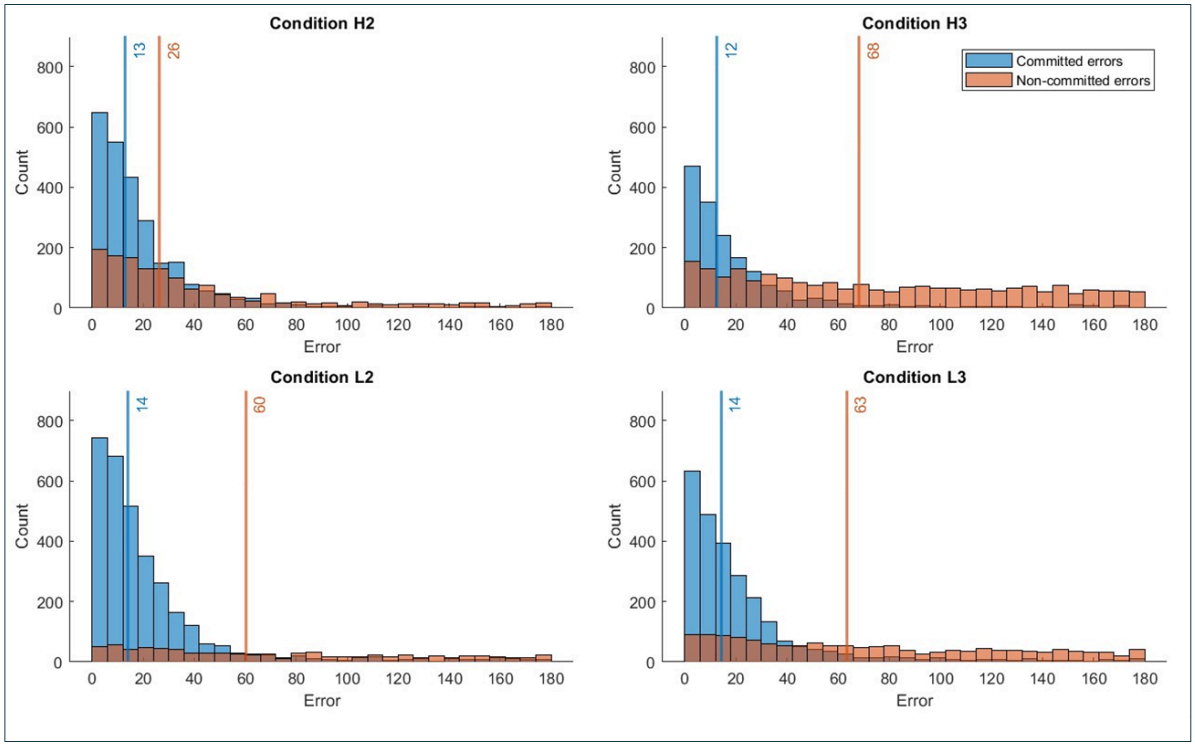

fig 6: For each condition, the distribution of errors (in degrees from the correct answer) for both the committed and non-committed errors are overlaid for comparisons. Counts calculated by taking the absolute value before binning. Vertical lines represent the median for both the committed errors (blue) and the uncommitted errors (orange) separately. The added number is the rounded error value of the median.

We also analysed which trials participants in different conditions chose to commit or not (*fig 6*). We theorised that participants in low punishment conditions would commit more overall, and also commit trials with higher errors compared to the high punishment conditions. This was because participants in low punishment conditions could afford to be less strict in their strategy for when to commit without losing more points compared to the high punishment condition. This should result in overall higher errors in the committed trials. To determine this, we pooled the data for the two punishment conditions together and performed a one-sided Wilcoxon rank-sum test. This test determines whether there is a statistically significant chance of the two datasets (high and low punishment errors) belonging to a distribution with the same median, as opposed to the alternative hypothesis that the low punishment errors were drawn from a distribution with a higher median. The significance level used for this test was 0.05. The result of this proved that the committed errors from the low punishment conditions were statistically greater than those from the high punishment conditions ($p = 7.27e\text{-}5$). This points towards the theory that participants with different punishment conditions apply differing strategies for committing trials

4.2 General Model Analysis

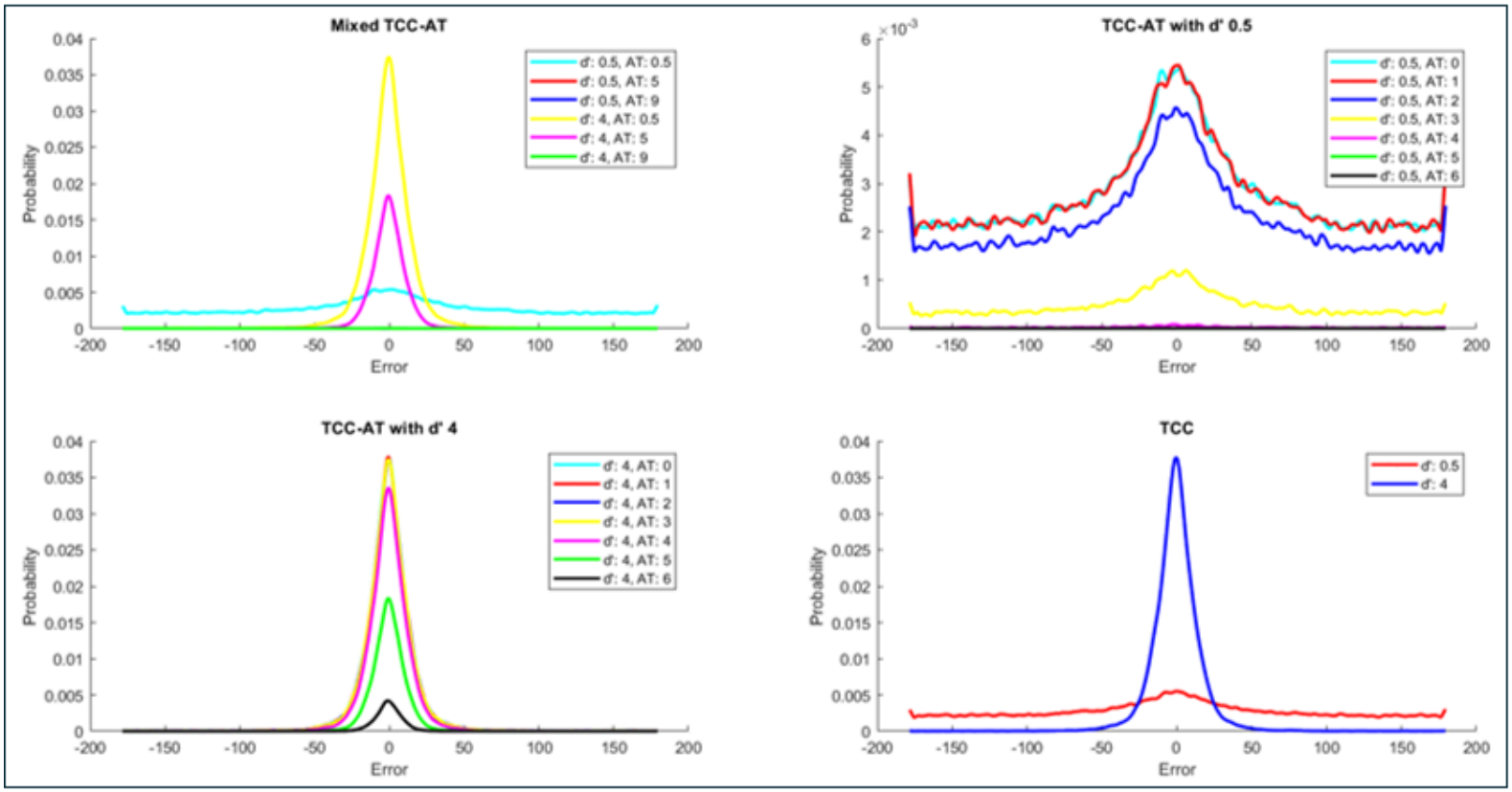

fig 7: Probability density functions (PDF) obtained with the TCC and TCC-AT model. Panels A, B and C depict hypothetical PDFs for the TCC-AT, reflecting different combinations of d' and AT values. These are based only on the trained model, not the experimental data. Panel D depicts the same hypothetical PDFs for the original TCC model. A) TCC-AT model with different combinations of d' (0.5 and 4) and AT (0.5, 5 and 9). B) TCC-AT model showing the effect of different ATs between 0 and 6 on a low d' (0.5). C) TCC-AT model showing the effect of different ATs between 0 and 6 on a high d' (4). D) original TCC model with different d' values (0.5 and 4)

Understanding the trained TCC-AT model allows us to also better understand the results of the experiment. This can be done by comparing the effect of different combinations of memory strengths and action thresholds on the shape of the PDFs, showing how memory strength and action threshold affect the distribution of errors. To achieve this, different possible combinations of d' and AT have been visualised.

Theoretically, both parameters affect a different aspect of the probability density function. The memory strength (d') has the same effect in both the TCC and TCC-AT models since it has

been designed so that the TCC-AT model with an AT of zero is the same as the original TCC model. As can be seen from the TCC model analysis (*fig. 7 D*), different memory strengths have probability density functions that differ in how steep they are. A higher d' means a sharper graph that reaches higher. The action threshold on the other hand changes just the height of the graphs, as can be seen in (*fig. 7 B and C*). These differences are much more pronounced when d' is higher, since low d' functions are already flat and thus closer to zero across the entire x-axis.

4.3 Analysis model using experiment

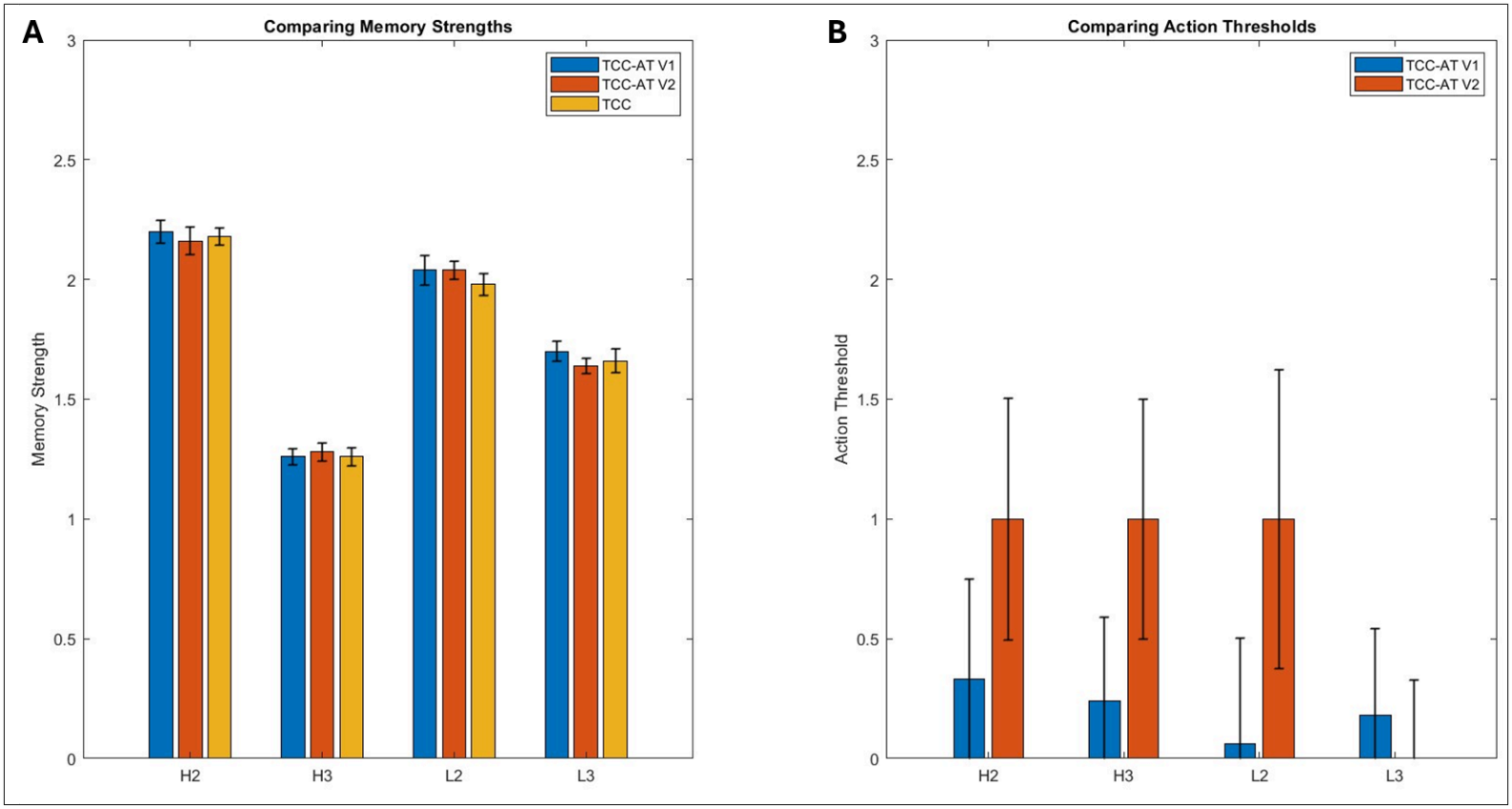

fig 8: Comparing memory strengths (A) and action thresholds (B) between models. TCC: the original TCC model when passed all the errors. TCC-AT V_1: Model searches through the entire 2D search space with linearly spaced d' and AT using all errors. TCC-AT V_2: Model first searches for the best fitting d' using all errors assuming an action threshold of 0, then it finds the best matching AT with that d' using the committed errors. Error bars are created by bootstrapping the errors for each participant 20.000 times and calculating the standard deviation per condition. Error bars cut off at 0, since a negative action threshold is not possible.

The memory strength and action threshold values cannot be verified on their own since they cannot be calculated from the experiment data, they can only be verified by comparing the parameters to those of other models and to each other. Two main ways of determining these values have been established, one for TCC-AT V_1 and one for TCC-AT V_2. V_1 uses a linear

search space for both d' and AT, and searches through the entire space to find the best fit using all errors. V_2 has been adapted to use a logarithmic scale and has a different search function. In V_2, the model first searches for the d' that best fits all the data, assuming an action threshold of 0 (based on all data, including both committed and non-committed trials). Then, the committed data is used to find the best AT given the previously found d'.

The values of memory strength can be compared to the values of the old TCC model (*fig. 8 A*). This showed that the new model finds memory strength values very close to the values found by the TCC model, regardless of the search function used. This is not surprising. Since both the training method and search method are based on the same principles. This method of analysis through comparisons was used since the results of the TCC model were already verified in the original paper for the TCC model, meaning that it provided a simple and reliable sanity check for the TCC-AT model.

The estimated action thresholds parameters do differ greatly between versions, which is to be expected as well. Since the V_2 AT scale is logarithmic, the lower range of the action threshold (between 0 and 2) has less options meaning they all fall back to the closest fit, being either 1 or 0. V_1 on the other hand has a linear scale with steps of 0.03, meaning there are many more options in between 0 and 1.

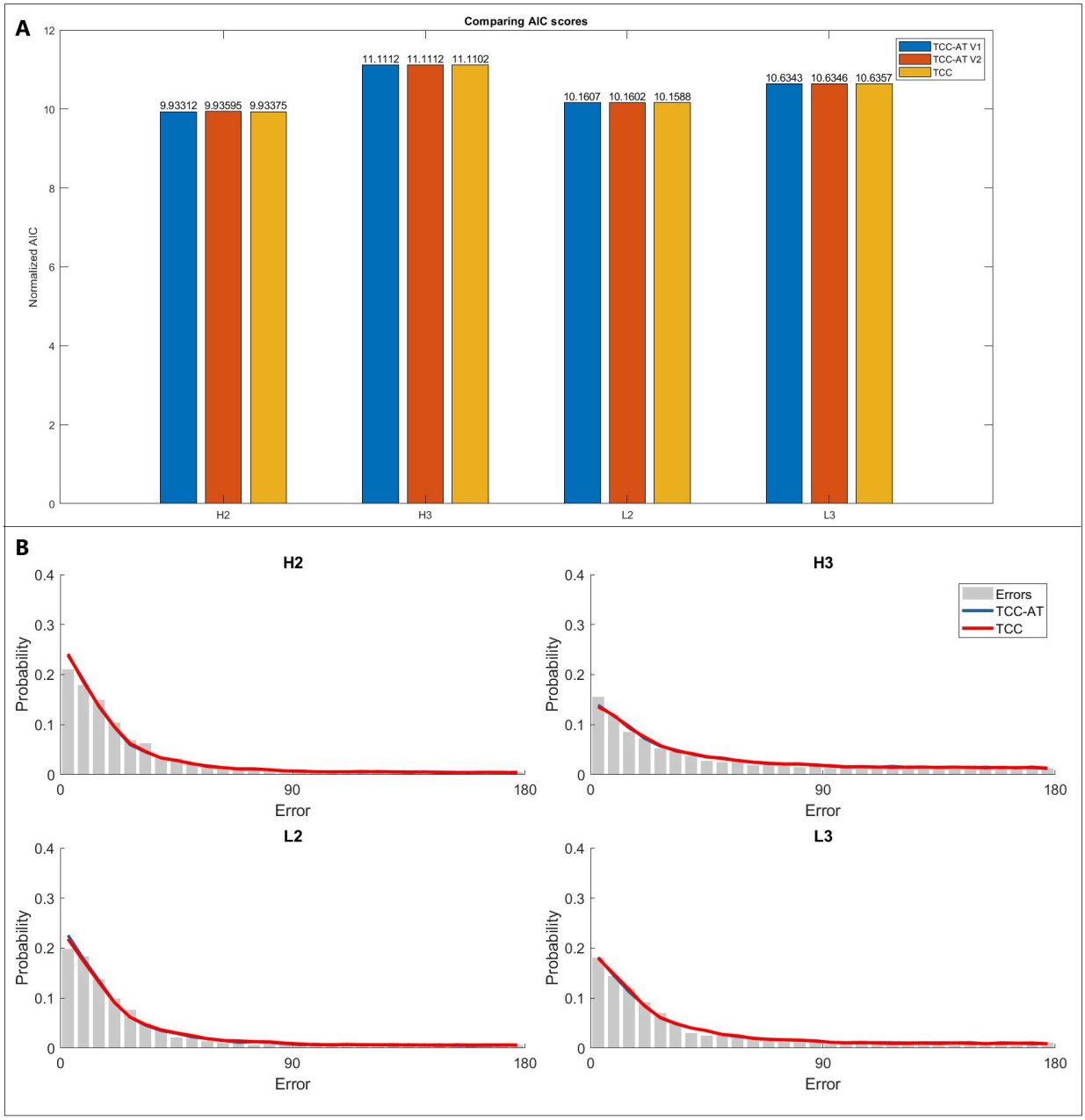

fig 9: A) Comparing AIC between different models. AIC score normalised on number of observations per condition. Note that standard deviations following the bootstrapping procedure were so small that they are not visible in the figure (e.g. STD for H2 V_1 was 0.02). B) Plots visualising the fits of the model to the actual error data per condition, comparing the original TCC model with the TCC-AT V_2 model.

Comparing the model fits using AIC scores allows us to determine which model has a better fit, taking into account the number of parameters in the model (*fig 9 A*). The two TCC-AT models have two parameters while the TCC model has only one. While ideally, the TCC-AT model would work with only the committed trials, the current model design only works by using all data for the d' calculations. It is clear from this that the two models have incredibly similar fits, with

very little variation. This is also clearly seen in (*fig 9 B*), the fits found by the two models almost completely overlap.

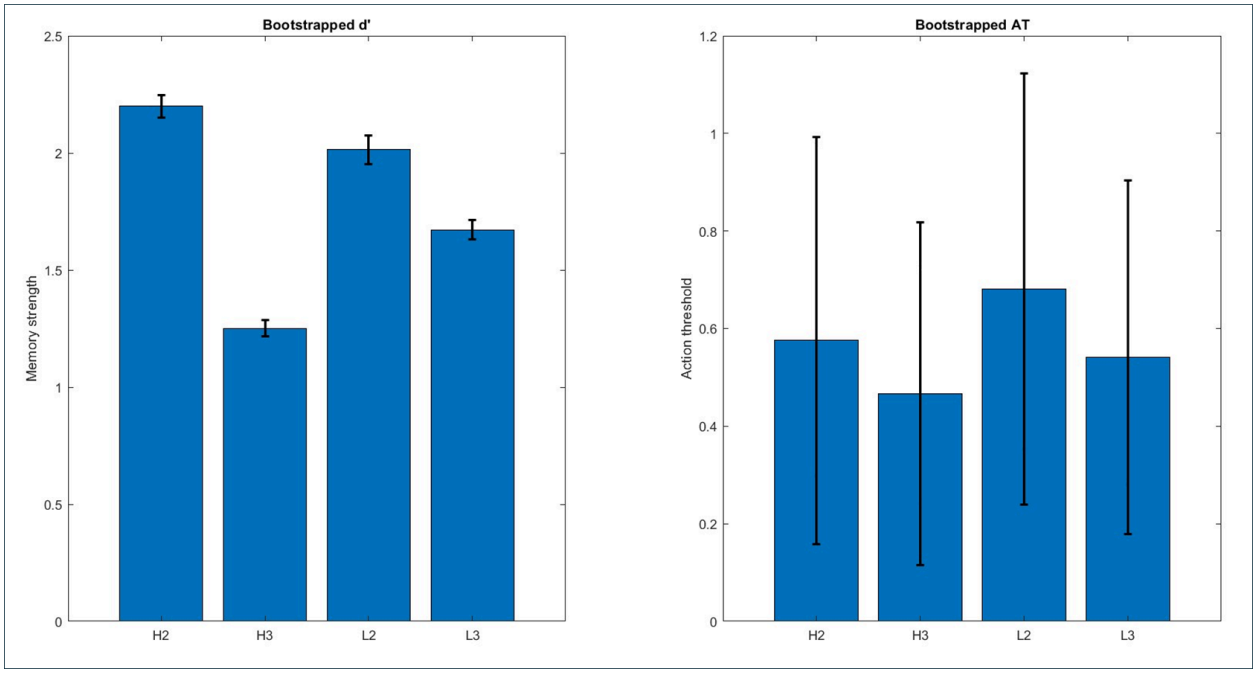

fig 10: Results of bootstrapping with 20.000 runs, randomly sampling errors per condition but keeping the distribution of commits/non-commits the same. The mean and standard deviations of the memory strengths and action thresholds are shown. Figure shows results of TCC-AT V_1.

To analyse the stability of the different parameters, we bootstrapped 20.000 runs, for each run randomly sampling the errors for each participant, but keeping commits the same. This kept the commit rate per participant the same, but changed the errors. From this, a d' and action threshold were calculated and the means and standard deviations of all 20.000 runs were calculated (*fig 10*). What this showed is that while memory strengths were quite stable, having low standard deviations over all the runs, the action threshold does not. Among all the 20.000 runs the action thresholds would vary quite widely. The mean action thresholds from bootstrapping does show a trend however. It shows that action thresholds would increase when set size became higher, and the action threshold would also increase when punishment level decreased. This is counter to what we theorised, since we believed the action threshold would be higher for higher punishment conditions.

5. Discussion

We wanted to answer whether two parameters (memory strength and action threshold) would suffice to model naturalistic working memory guided behaviour. Here, naturalistic refers to a task where participants are not forced to use working memory, but can instead choose to resample the external world. To this end we developed a computational model called the TCC-AT model and conducted a proof-of-principle experiment to test this model. This experiment was designed with four participant groups, each manipulated through set size and punishment intended to show predictable differences in memory strength and action thresholds. We found that while the TCC-AT model performed as well on memory strength as the original TCC model, it showed significant variance in action thresholds. Along with this it showed an unexpected dependency between set size and commit rate.

The experiment we had designed had to show differences in confidence between participant groups. The data provided through this experiment provided us with the different ways set size and punishment affect commit rate (confidence) and participant errors. We found that set size affected both report errors and commit rate. Punishment on the other hand only affected the commit rate directly. While punishment did not affect report error directly, it did affect report error indirectly by changing the effect set size has on report error. Not all of these findings were in line with our initial expectations. Ideally, punishment and set size would affect commit rate and report error respectively, allowing for clear distinctions between conditions to base our analysis of the two free parameters.

For the experiment data to be usable in the analysis of the model, we had to know if the model itself could represent different memory strengths and action thresholds. For memory strength, we could rely on the original TCC model since they had already verified the usage of memory strength and we did not change the component of the model that determines memory strength. For action thresholds, we showed that their effect on the probability density functions was the same regardless of memory strength.

When testing the model with experiment data, we knew the TCC-AT model would fit at least as well as the TCC model, since the TCC-AT model with action threshold set to zero is the exact same as the TCC model. The main question we wanted to answer is whether the changes we made to make the TCC-AT model could also give us an action threshold fitting with the theory, and providing a better fit than the TCC model on the experiment that includes whether participants were confident enough in their answer. The results we found showed that the model was still able to accurately predict memory strength, coming close to the original TCC model predictions. The action threshold results were less conclusive, varying greatly and being inconsistent with the theory. The action thresholds gained from simply using the participant data directly would not be reliable due to this, so it is better to draw conclusions from the bootstrapped data. Memory strength in these bootstrapped results behaved as expected: higher set sizes corresponded to lower memory strength since more items had to be remembered. With the action thresholds the results were not as theorised. It showed that action thresholds would decrease for the higher set size.

Theoretically, these findings where the high punishment condition provided lower action thresholds would indicate that participants who were punished more harshly would be less strict in which answer they committed. This is counter to the theory, and it could be explained in a multitude of other ways. It could be possible that this is due to chance, given the high (and overlapping) variances in action thresholds. The discrepancy could also be caused by unintended behaviour in the experiment. Participants in the high punishment condition might have become frustrated with the scoring of the game and therefore started taking it less seriously and not trying as hard. This is a possibility since the total scores participants achieved in the experiment varied greatly between punishment levels, with participants in the high punishment condition losing more points than they gained (around -3000 in total), while participants in the low punishment condition gained more points than they lost (around 3000 in total). This could have resulted in participants in high punishment conditions becoming much more frustrated with the task and thus approaching the task with different motivation levels.

The effect of set size on the commit rate, where a higher set size resulted in a lower commit rate regardless of the punishment level can be explained by the fact that when resources are spread out over more items, the chance is higher that a participant did not see the colour being asked about properly. If they could only see two of the three colours properly (due to distractions or losing focus during the experiment, for example) then they might choose not to commit. Since this is more likely to happen when there are three items to remember than when there are only two. this could result in the lower commit rate for the higher set size

All of these results allowed us to draw a couple of conclusions. The first one being that when looking at participant data alone, we see a distinct difference between the two parameters controlled on (commit rate and report error). The high variance in action thresholds and the action threshold results being so close together however prevent us from definitively answering the question of whether memory strength and action threshold can be used to model naturalistic VWM use. We have however been able to more clearly establish the effect of set sizes and punishment level on commit behaviour.

Additional research is needed to definitively answer the research question. Firstly, the way the action threshold alters the probability density functions in the model could be reviewed, changing the way the model is trained and how it searches for the best memory strength and action threshold. This study provided a more comprehensive view on how people with different strategies behaved when it came to committing trials, and more specifically which trials would be committed and which would not be. Knowing the distribution of committed errors versus non-committed errors in different conditions would allow the probability density function to be altered by different action thresholds in a way that fits with the behavioural data better than the current method does. Secondly, a different experiment could be designed that makes the differences between commitment strategies larger. This experiment should then also ensure that the different conditions do not introduce additional factors that could unduly affect the commitment strategies or report errors. An example of such a factor in our experiment was that

participants in the low punishment condition would consistently gain points while the participants in the high punishment condition would mostly lose points.

Further research and such changes would allow the specifics of behaviour in naturalistic memory tasks to be learnt and also provide cleaner behavioural data to analyse the model and its underlying theory. This would bring us closer to both understanding and being able to further utilize the cognitive processes surrounding our memory.

6. References

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *The Psychology of learning and motivation/The psychology of learning and motivation* (pp. 47–89). https://doi.org/10.1016/s0079-7421(08)60452-1

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7. https://doi.org/10.1167/9.10.7

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2020). When natural behavior engages working memory. *Current Biology*, *31*(4), 869-874.e5. https://doi.org/10.1016/j.cub.2020.11.013

Ma W.J., Husain M., & Bays P.M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Sahakian, A., Gayet, S., Paffen, C. L., & Van Der Stigchel, S. (2023). Mountains of memory in a sea of uncertainty: Sampling the external world despite useful information in visual working memory. *Cognition*, *234*, 105381. https://doi.org/10.1016/j.cognition.2023.105381

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Somai, R. S., Schut, M. J., & Van Der Stigchel, S. (2019). Evidence for the world as an external memory: A trade-off between internal and external visual memory storage. *Cortex*, *122*, 108–114. https://doi.org/10.1016/j.cortex.2018.12.017

Van Der Stigchel, S. (2020). An embodied account of visual working memory. *Visual Cognition*, *28*(5–8), 414–419. https://doi.org/10.1080/13506285.2020.1742827