Reliability of Visual Access: Modeling the trade-off between internal storage and external sampling in a Visual Working Memory task

Alex J. Hoogerbrugge
4137477
a.j.hoogerbrugge@uu.nl

November 30 2020

Abstract

We use visual working memory to temporarily store visual information about our environment. However, our environment is mostly visually static and as such, memory can often be 'offloaded' onto the environment. This leads to a trade-off between choosing to internally store information or to externally sample it. In this thesis we explored how this storage/sampling trade-off changes as reliability of access to the environment changes, by submitting participants to a copying task. In this task, participants were instructed to copy a layout of stimuli on the left side of a computer screen to the right side of the screen. The example layout intermittently appeared and disappeared throughout a trial, the timing of which we varied across conditions. We found that, as the example layout disappeared for greater amounts of time, participants sampled it less often (and thus likely memorised more items at once) than in the baseline condition, in which the example layout was always visible. We then designed and ran a computational cognitive model, with which we attempted to simulate the participants' behaviour in such a way that that we could compare the model's results to behavioural observations. The model explored different combinations of possible strategies, namely regarding the number of stimuli it attempted to remember with each gaze toward the example layout, and regarding how many times a stimulus was rehearsed in memory after its first encoding. We then compared human data and model data on three outcome variables: (1) the *number of crossings* from the right side of the screen to the example layout on the left side; (2) the *completion time* of trials; and (3) the *number of fixations per second*. A model was found that fits well to participants' *completion time* and *number of fixations* per trial, but less strongly on the *number of crossings* per trial. Our findings suggest that there may indeed occur a shift in in the usage of visual working memory when reliability of visual access changes, which leads us to believe a storage/sampling trade-off also exists in environments with varying reliability of visual access. We conclude by recommending that future research endeavours take into account a storage/sampling trade-off in conditions with varying reliability of visual access.

Thesis manuscript for obtaining the degree of
Master of Science in Artificial Intelligence

Graduate School of Natural Sciences
Utrecht University

Supervisor: Prof. Dr. Stefan van der Stigchel

Second examiner: Dr. Tanja C. W. Nijboer.

Contents

1 Introduction

We live in a visually rich environment. In order to interpret and process visual information, the objects that we see around us can be temporarily stored in visuospatial working memory; a short-term, quick-access form of memory, considered to be part of the general working memory system (Baddeley & Herring, 1983; Salway & Logie, 1995). In turn, visuospatial working memory consists of two closely related components: spatial working memory (SWM) and visual working memory (VWM). SWM is thought to be responsible for encoding location pointers to visual information in the world, whereas VWM is thought to encode and maintain visual features of stimuli in memory (Baddeley, 2000; Baddeley & Hitch, 1974).

There is ongoing debate about what the constraints and capacity of VWM are and whether the capacity is a rigid number of discrete units that can be stored either accurately or not at all (e.g., Luck & Vogel, 2013), or whether it is limited by available resources in general working memory and the quality of stimuli (e.g., Ma, Husain, Bays, & de Soissons, 2014). Nevertheless, these discussions all describe VWM as a storage medium of limited capacity, such that when a new stimulus needs to be remembered, an old one often needs to make way.

Clearly, it would be very costly to constantly erase and write visual stimuli into VWM, as it not only requires attentional resources (Cowan, 2016), but internal representations may also be encoded incorrectly or be subject to decay (Baddeley & Hitch, 1974). The high cost of internal storage may especially be applicable in repetitive tasks where more objects are present than can be stored and where some of them need to be used frequently. Think of laying a puzzle, where it can be useful to memorize what the edge pieces look like, but still needing to visually sample the inner pieces. By any current theory, even a 100-piece puzzle contains more objects than can be reliably stored in VWM, and thus items need to be sampled, stored, retrieved, and resampled often.

1.1 Storage versus sampling

However, while at the grocery store, taking a stroll through the park, or sitting behind a desk, it is safe to assume that most, if not all, elements will still be there after looking away for a few seconds. The peanut butter on the store shelf, the trees, and the picture frame will still look the same and they will not have moved. With that in mind, O'Regan (1992) proposed that the real world may be used as external memory in order to reduce the cost of handling internal memory.

Nevertheless, using the world as an external memory source requires saccades to be made towards the desired object every time its information needs to be retrieved, which may be considered too costly as compared to storing information internally or vice versa. Therefore, there must be some internal function which models the cost of using the real world as external storage and thus to make more saccades, versus when to store stimuli in VWM ("external sampling" and "internal storage" respectively; Van der Stigchel, 2020).

Previous research into this storage/sampling trade-off has often taken the form of a copying task, in which an example array of stimuli on one side of the screen needed to be copied to a workspace on the other side of the screen (Ballard, Hayhoe, & Pelz, 1995; Gray, Sims, Fu, & Schoelles, 2006; Inamdar & Pomplun, 2003; Melnik, Schüler, Rothkopf, & König, 2018; Somai, Schut, & Van der Stigchel, 2020). It was shown that generally, when instructed to perform quickly and accurately, making more saccades (and thus externally sampling) was participants' preferred strategy. This preference towards external sampling is in accordance with Wilson (2002), who posed that cognition is time-pressured and often offloaded onto the environment. However, when the cost of saccades was increased – e.g., by increasing the distance between example and workspace (Ballard et al., 1995; Inamdar & Pomplun, 2003), or by delaying the appearance of the example (Gray et al., 2006; Melnik et al., 2018; Somai et al., 2020) the dominant strategy shifted towards making fewer saccades, and thus storing more information internally in VWM. These findings conform with earlier research which has demonstrated or argued for the adaptive nature of strategy selection by participants (e.g., Brumby, Howes, & Salvucci, 2007; Cary & Carlson, 2001; Charman & Howes, 2003; Howes, Duggan, Kalidindi, Tseng, & Lewis, 2016).

If we relate these findings to the example of the puzzle; if the puzzle is small enough in size – say it were to fit on a dinner plate – there may be a dominant strategy of looking at a corner piece and all other edge pieces in turn – every time seeing whether it fits the corner piece by making a saccade back towards it. In contrast, if the puzzle is sufficiently large such that all pieces spread out may occupy an entire dining table, the dominant strategy may be to briefly store the features of the corner piece in VWM, then looking sequentially at each edge piece and retrieving the corner piece's features from VWM to see whether the two fit together.

1.2 Reliability of access to visual information

It should be noted that the aforementioned strategies in the copying tasks depend on the stimuli always being present – and in the same place – after a saccade. But what happens when the *reliability of access* to the environment decreases? For instance, in scenarios with moving objects, a stimulus may no longer be present after a saccade has been made elsewhere. In other cases, access to stimuli may be unreliable, such as when other objects regularly move in front of the desired stimulus. For example, when driving, important stimuli such as lane markers, traffic signs or crossing pedestrians may be obscured by other vehicles, by the car's blind spot, or roadside objects such as trees (Senders, Kristofferson, Levison, Dietrich, & Ward, 1967). More generally applicable; rain, snow, or glare from the sun may temporarily obstruct one's view, as could a flickering light or a tall person in front of you at a concert. In these scenarios, visual stimuli are temporarily blocked or occluded, and the frequency of availability of stimuli, although somewhat predictable, is often not known with complete accuracy at the level of milliseconds. We thus describe these scenarios as having varying reliability of visual access.

Given the assumption that one is aware of the degree of reliability of access to the en-

vironment, presumably there also exists a function which models the storage/sampling trade-off in situations where visual access to the environment may only be intermittently available. Therefore, we expect that decreasing the reliability of access to visual stimuli modulates participants' strategy selection in a similar manner to increasing the cost of access as found in previous experiments. If so, we would expect participants to offload memory to the environment as much as possible in a baseline environment, in which stimuli are always visible. Accordingly, low-reliability of access environments are expected to give rise to approaches based more on internal storage.

1.3 Cognitive models

The difficulty of researching mental strategies, however, is that they cannot be directly deduced from observational data. For instance, we can neither directly measure how many stimuli are stored in VWM, nor how accurately they are stored. Luckily, computational cognitive models have been making headway in exploring, explaining and modeling cognitive processes which are not directly observable (McClelland, 2009).

In cognitive process modeling, there are generally two approaches: (1) uncovering strategies which lead to some optimal performance metric (such as the shortest completion time) and comparing them to human performance, which then allows statements about the (sub-)optimality of human performance; and (2) uncovering which strategies fit best to human performance, since it is held that humans are boundedly optimal performers (Lewis, Howes, & Singh, 2014; Russell & Subramanian, 1995). The first approach is commonly encountered in artificial intelligence research, where models are often built to perform as optimally as possible on some task, regardless of how the performance is achieved. The second approach is more human-centered; where models are built from a combination of existing theories and newly observed traits of human cognition. Comparing a model's performance to human data then gives an indication of how well a theory explains human behaviour. We define and test a human-centered model on a copying task in which the reliability of access to the example grid is modulated throughout different conditions. The reliability of access is manipulated across conditions by making the example grid disappear and reappear at a different pace for each condition. We then explore how well the model fits to human data under different strategy combinations.

In this thesis, we expect to find that low-reliability of access environments give rise to a dominantly storage-based strategy, and vice versa. The objective of this thesis is then to uncover the parameters of the strategies which may underlie this behaviour by designing a computational cognitive model which approximates participants' behaviour on a task with varying degrees of reliability of access.

The first type of strategy we explore contains how many stimuli are encoded each time the gaze shifts towards the example grid. The second type of strategy we explore is how often an item is rehearsed in VWM after the first time it is encoded. By retrieving an encoded item from memory, its activation strength is increased and it is therefore less likely to be forgotten over time. Finally, we explore different combinations of parameters for

memory encoding. Based on the ACT-R framework, we tune these parameters so that our model can most accurately predict how long encoding items in VWM takes and how long it takes to retrieve those items again (Anderson, 1996; Anderson & Schooler, 1991; Lovett, Reder, & Lebiere, 2012).

Assuming the model's parameters besides strategy – such as duration of saccades and mouse movements – are in accordance with those exhibited by participants, it can be argued that the strategy which best fits to the observed data provides evidence towards uncovering the participants' prevalent strategy (Gershman, Horvitz, & Tenenbaum, 2015; McClelland, 2009). We therefore model the duration of mouse- and eye movements to fit to the data we observed from human participants in our experiment.

2 Methods

2.1 Participants

Participants were recruited via word-of-mouth. There were no prerequisite requirements except that the participants should have normal or corrected to normal visual acuity and could control a mouse and keyboard. We tested 14 participants, of which 7 female (mean age = 31, *SD* = 13.8, range = 22-63). One participant was excluded on the basis of too many missing data points, which we discuss in Section 2.5.

All participants signed an informed consent form and were compensated 7 euros per hour. The experiment was approved by the Faculty Ethics Review Board of the Faculty of Social Sciences, Utrecht University.

2.2 Apparatus and stimuli

The experiment was programmed in Python 3.7 (Python Core Team, 2019) using the PyQt5 library (Riverbank Computing Limited, 2019) for visual presentation and interaction with the mouse and keyboard. PyGaze (Dalmaijer, Mathôt, & Van der Stigchel, 2014) was used to interface with an Eyelink 1000 eye tracker (SR Research Ltd., Canada), which measured at a sampling rate of 1 kHz. Data processing was performed in Python 3.7, using the Pingouin 0.3.8 package (Vallat, 2018) for statistical analyses.

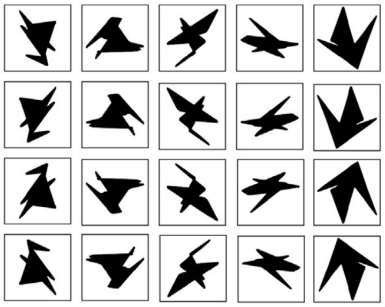

Figure 1: Stimuli as adopted from Arnoult (1956); Somai et al. (2020). There are 4 unique shapes, each rotated at a multiple of 90 degrees, creating 16 stimuli in total.

The experiment was run on a Windows 10 Enterprise computer with an Intel Core i7-4790 CPU and 16GB RAM, and displayed on a 27 inch ASUS PG278Q LCD monitor at a resolution of 2560×1440 pixels @ 60Hz. Participants placed their heads in a fixed chin

rest in a dimly lit room at 70 centimetres from the screen, such that each 100×100 pixel stimulus occupied a visual angle of approximately $1.75°$ to $1.95°$ on both the horizontal plane and the vertical plane, dependent on its position on the screen.

The stimuli used in this experiment were adopted from Arnoult (1956) (Figure 1). As discussed in the original paper and in Somai et al. (2020), these stimuli were used in an attempt to prevent participants from using mnemonic devices and to increase the reliance on visual working memory. Additionally, they should have been novel enough that participants had not yet acquired an internal representation through prior experience, since prior experience could offload working memory (Arnoult, 1956; Wilson, 2002). Although the set of stimuli contained 20 images, they consisted of five unique shapes, each shape additionally mirrored horizontally, vertically, or both.

2.3 Task

The experiment consisted of a copying task in which participants were asked to copy a layout of 4 stimuli in a 3×3 grid on the left side of the screen to an empty 3×3 grid on the right side of the screen. These areas are referred to as the 'example grid' and the 'working grid', respectively. The centres of both grids were located at a visual angle of $12°$ (635 pixels) from the centre of the screen, with each of the grids occupying approximately $7.3° \times 8.8°$ (415×460 pixels) of the visual field. On the bottom right of the screen, the same stimuli were presented as in the example grid, but in randomized order. We refer to this area as the 'resource grid'. The participants' task was to recreate the layout of the example grid in the working grid, by dragging stimuli from the resource grid to their correct location in the working grid.

In the *baseline* condition (0), the example grid was always visible. In order to experimentally manipulate the reliability of access, the example grid was either present or occluded at specified intervals throughout a trial. In the three experimental conditions the example grid was (1) repetitiously visible for 4 seconds and subsequently occluded for 2 seconds, such that the reliability of access was *high*; (2) repetitiously visible for 3 seconds and occluded for 3 seconds, such that the reliability of access was *medium*; (3) repetitiously visible for 2 seconds and occluded for 4 seconds, such that the reliability of access was *low* (see Table 1). Appearance and occlusion of the example grid were repeated until the trial ended (see Figure 2). In order to further decrease reliability of access, the predetermined visibility time was multiplied by a noise factor drawn from a Gaussian distribution ($\mu = 1.0$, $\sigma = .1$) for each trial, with the occlusion time being adjusted accordingly such that the sum of visible time and occlusion time was always 6000 ms. As such, the visible/occluded times in the *medium* condition could, for example, be $[2900, 3100]$ in one trial and $[3001, 2999]$ in the next – and so on.

A trial ended whenever the grid was fully copied or a predetermined timer of 20 seconds ran out, the latter of which forced a sense of urgency on participants and is likely an important factor in optimization of behaviour (Janssen & Gray, 2012; Melnik et al., 2018).

Each condition was tested in its own block of trials and the block order was random-

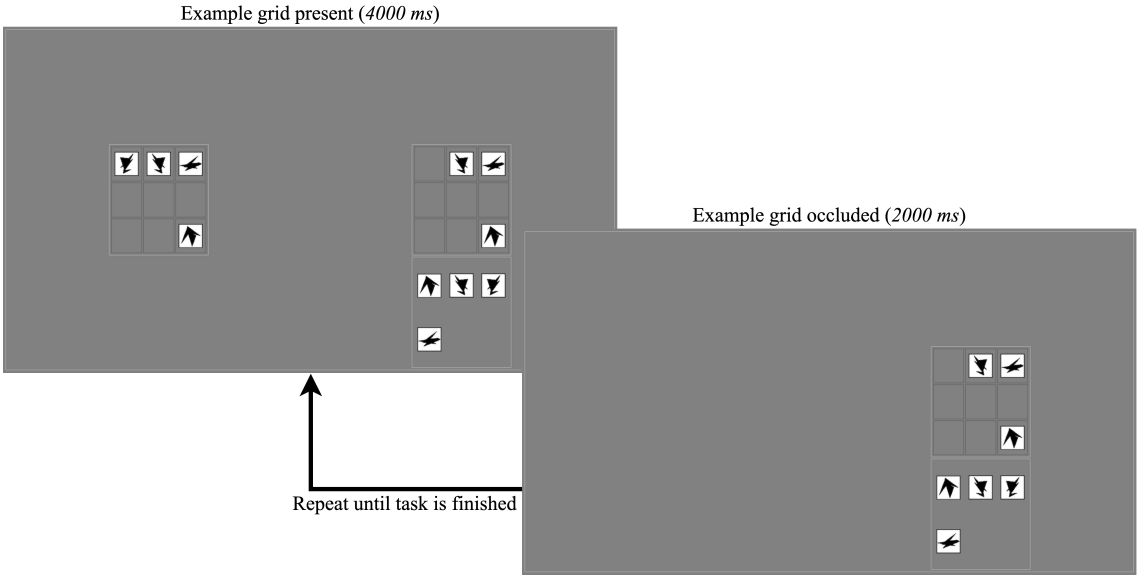

Figure 2: Example overview of a partially completed trial in the *high* reliability of access condition. In this example, three items have already been dragged to their correct positions, with the top-left one remaining.

ized between participants. In every trial, the example grid was randomly generated with 4 randomly chosen stimuli. In order to allow participants time to establish a (rough) estimation of the reliability of access within a condition, data of the first three trials of each block were discarded from analysis.

2.4 Procedure

Participants were instructed about the goal of the task, how to control the task, and about the 20-second limitation. They were also instructed that misplaced stimuli could be removed from the working grid by either right-clicking the misplaced item or dragging the correct item over top of it. Finally, participants were instructed to not move their head in the chin rest after measuring had started. A short break could be taken after each block of trials, with a longer 5-10 minute break halfway through the experiment. Additionally, they could indicate whether they needed a quick break during a block of trials to relax from their static posture.

Each participant was first subjected to five practice trials in the baseline condition (always visible). After confirming that they understood the task and were familiarized with

Table 1: Visibility of the example grid across conditions in terms of *high*, *medium* and *low* reliability of access. As the example grid becomes occluded for longer, the reliability of access decreases.

	Condition			
	0 (*baseline*)	1 (*high*)	2 (*medium*)	3 (*low*)
Visible (ms)	Always	4000	3000	2000
Occluded (ms)	Never	2000	3000	4000

the controls, they started the actual experiment. In the experiment, the four conditions were presented in blocks of 35 trials. After each trial ended, a blank screen with a message would appear. If the trial was correctly finished, the message would instruct the participant to press the space bar to continue to the next trial. If the trial was not completed in time, it would read "You timed out," paired with the standard message. After 35 trials, the message indicated that the block was completed and that participants could take a break.

The eye tracking was calibrated and validated before the start of each block, and was validated (and if necessary, re-calibrated) approximately every 3 minutes, and at the end of each block. The experiment took approximately 45-90 minutes to complete, dependent on task speed, calibration time, and the number and length of breaks.

2.5 Analysis

Due to a software bug in the experiment, the first seven participants experienced some data loss. Participants were excluded if there was data from fewer than 15 trials remaining in at least one of the conditions. One participant was excluded from analysis based on this criterion.

We analyse six key variables for significantly different outcomes between conditions. (1) The *number of crossings*, which is calculated by counting how many times within a trial the participant made a saccade across the centre of the screen from the right side to the left side of the screen. In effect, this variable represents how often participants sampled externally, by looking toward the example grid after focusing on the working- and resource area. (2) The *total dwell time per crossing* is measured in milliseconds, and is calculated as the total fixation time on the left half of the screen, after each time the centre of the screen is crossed from right to left. This variable acts as a proxy for how much information participants took in each time they shifted their attention toward the example grid. (3) *Completion time (seconds)*. Although completion time is expected to increase as the occlusion time increases, the time in which the example grid was occluded is not considered lost. Participants could, for example, still drag stimuli to their correct spot or correct their mistakes. Therefore we include the pure completion time. (4) The *number of fixations per second*. (5) *Median saccade velocity* is calculated over a whole trial as a representation of mental workload, since a decrease in (peak) saccade velocity has been shown to be linked to an increase in mental workload (Di Stasi, Antolí, & Cañas, 2011; Di Stasi et al., 2010). Finally, (6) the number of *errors per trial* is defined, in which an error constitutes placement of an item in an incorrect position.

These six variables first needed to be aggregated per participant, per condition. Since the data for all but one of the variables was not normally distributed, we calculated the median value of a variable over all trials a participant has performed within each condition. The exception is that the mean was calculated for *number of errors*, since using the median drew results close to zero.

We report averages and standard deviations for each of the variables over all participants and perform a Shapiro-Wilk normality test and a Mauchly's sphericity test. To test

Table 2: Descriptive statistics for all variables, reported as 'mean (SD)'. The highest value for each variable is in bold.

	Condition			
	baseline	*high*	*medium*	*low*
Number of crossings	**5.0** (0.68)	4.62 (0.92)	4.54 (0.91)	3.65 (0.82)
Dwell time per crossing (ms)	293 (67.0)	319 (96.0)	314 (76.0)	**430** (159)
Completion time (s)	6.89 (1.53)	7.14 (1.80)	8.10 (2.06)	**9.21** (2.33)
Fixations per second	**3.80** (0.51)	3.76 (0.52)	3.64 (0.56)	3.46 (0.53)
Saccade velocity	**165.1** (27.5)	152.7 (17.6)	148.4 (18.6)	137.2 (14.9)
Errors per trial	0.11 (0.07)	0.20 (0.13)	0.29 (0.18)	**0.38** (0.25)

whether there is a significant effect of condition on each of the variables, we report the results of a Repeated Measures ANOVA per variable. However, if the assumption of either normality or sphericity are violated for at least one of the conditions within a variable, we report the results of a Friedman Chi-squared test, which is considered the nonparametric counterpart to the Repeated Measures ANOVA (Friedman, 1937). η^2 and Kendall's W are calculated as indications of effect sizes for the ANOVA and Friedman's tests respectively. As a post-hoc analysis to observe precisely between which conditions variables differ significantly, we report either a one-tailed paired samples t-test or a one-tailed Wilcoxon signed-rank test for each condition pair, dependent on normality. We perform one-tailed post-hoc tests because we expect directional effects across conditions.

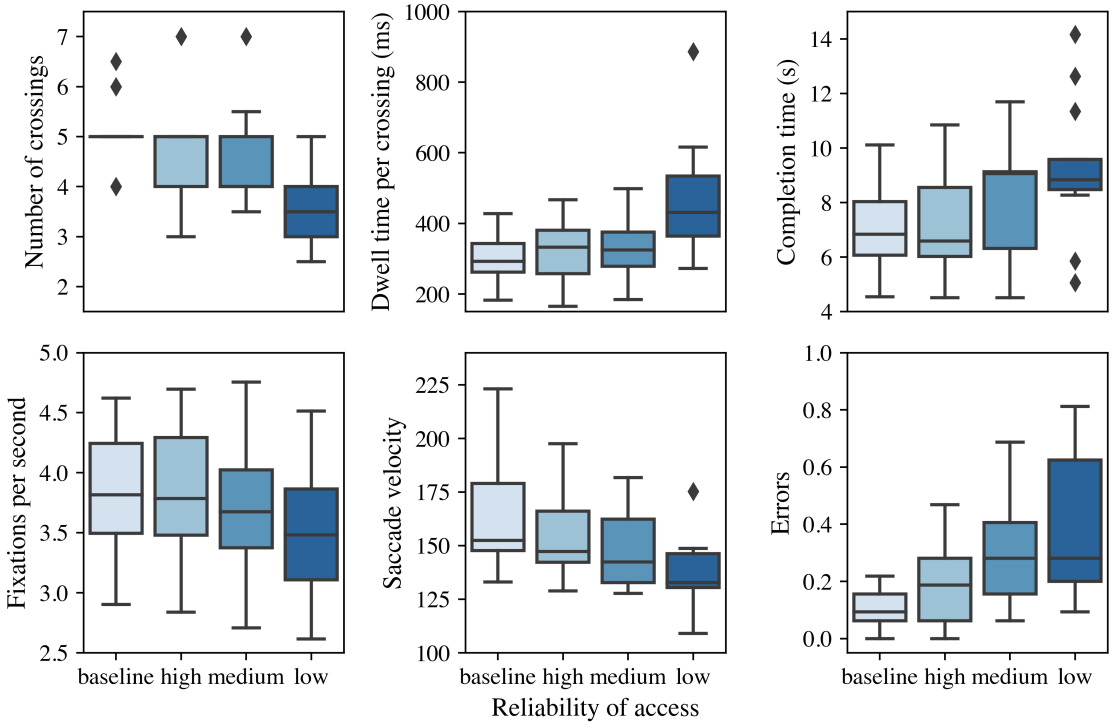

Figure 3: Box-and-whisker plots for each variable, per condition. $N = 13$ for all conditions and all variables. Data was first aggregated per participant by calculating the median value over all trials (except for *Errors*, for which the mean was calculated.

3 Results

Of 13 participants performing 32 trials per condition (1,664 trials total), data of 81 trials (4.9%) were lost. On average, data of at least 31 trials remained for each condition except *low* reliability of access, where on average 28 trials remained. We report the means and standard deviations for each of the six outcome variables in Table 2. In Figure 3 we report the box-and-whisker plots for each variable, per condition. We give an overview of our findings in Table 3 and Table 4.

3.1 Nonparametric tests

A significant effect of condition on the number of crossings was found, $\chi^2 = 23.0$, $p < .001$. This effect is considered medium to large, Kendall's $W = .59$. Post-hoc Wilcoxon signed-rank tests showed that, at significance level $\alpha = .05$, there was a difference in distribution between conditions (*baseline*, *medium*). At $\alpha = .01$ there was a difference between conditions (*high*, *low*) and (*medium*, *low*), and at $\alpha = .001$ there was a difference between conditions (*baseline*, *low*). This means the number of crossings was significantly affected by reliability of access, and there was a significant drop in the number of crossings in the *low* reliability of access condition as compared to the *baseline*, *high* and *medium* conditions, as well as between the *high* and *low* reliability conditions.

The effect of condition on median *dwell time per crossing* was significant, $\chi^2 = 22.0$, $p < .001$. This effect is considered medium to large, Kendall's $W = .56$. Although the assumption of sphericity was violated, the data showed no violation of normality. As such, post-hoc paired samples t-tests were used to show that, at significance level $\alpha = .01$ there was a difference between conditions (*baseline*, *low*), (*high*, *low*) and (*medium*, *low*). This means the time participants viewed the example grid after each crossing was significantly affected by reliability of access, and there was a significant increase in dwell time in the *low* reliability of access condition as compared to the *baseline*, *high* and *medium* conditions.

Lastly, a significant effect of condition on errors per trial was found, $\chi^2 = 18.7$, $p < .001$. This effect is considered medium, Kendall's $W = .48$. Since the assumption of normality was not violated, t-tests for paired samples were used as post-hoc tests. These tests show that, at significance level $\alpha = .05$, there was a difference in distribution between conditions (*baseline*, *high*), (*high*, *low*) and (*medium*, *low*). At $\alpha = .01$ there was a difference between conditions (*baseline*, *medium*) and (*baseline*, *low*). This means there was a significant effect of condition on the number of errors per trial, and the number of errors significantly increased between all conditions but the *high* and *medium* reliability of access conditions.

3.2 Parametric tests

The last four variables showed no violation of normality or sphericity, and were therefore tested with Repeated Measures ANOVAs and one-tailed t-tests for paired samples.

Firstly, a significant effect of condition on *trial completion time* was found, $F = 22.3$, $p < .001$. This effect is considered large, $\eta^2 = .65$. Post-hoc paired samples t-tests show

Table 3: Friedman (χ^2) and Repeated Measures ANOVA (F) results, respectively. Effect sizes are reported as Kendall's W for the Friedman test, and as η^2 for the Repeated Measures ANOVA.

	χ^2	F	df	p	W	η^2
Number of crossings	23.0		3	$< .001$	.59	
Dwell time per crossing	22.0		3	$< .001$	.56	
Completion time		22.3	3	$< .001$		.65
Fixations per second		17.7	3	$< .001$		.60
Saccade velocity		17.8	3	$< .001$		.60
Errors per trial	18.7		3	$< .001$	.48	

that, at significance level $\alpha = .01$ there was a difference between conditions (*high*, *medium*) and (*medium*, *low*), and at $\alpha = .001$ there was a difference between conditions (*baseline*, *medium*), (*baseline*, *low*) and (*medium*, *low*). This means the total time participants needed to complete a trial was significantly affected by reliability of access, and there was a significant increase in completion time between all conditions except for between the *baseline* and *high* reliability of access conditions.

Secondly, the effect of condition on the *number of fixations per second* was significant, $F = 17.7$, $p < .001$. This effect is considered large, $\eta^2 = .60$. Post-hoc paired samples t-tests show that, at significance level $\alpha = .05$ there was a difference in distribution between conditions (*high*, *medium*). At $\alpha = .01$ there was a difference between conditions (*baseline*, *medium*) and (*medium*, *low*), and at $\alpha = .001$ there was a difference between conditions (*baseline*, *low*) and (*high*, *low*). As such, the number of fixations per second within trials was significantly affected by reliability of access, and there was a significant decrease in the number of fixations per second between all conditions except for between the *baseline* and *high* reliability of access conditions. This implies that participants either fixated for greater amounts of time or made slower saccades as reliability of access decreased.

Finally, there was a significant effect of condition on *median saccade velocity*, $F = 17.8$, $p < .001$. This effect is considered large, $\eta^2 = .60$. Post-hoc paired samples t-tests show that, at significance level $\alpha = .01$ there was a difference between conditions (*baseline*, *high*), (*baseline*, *medium*) and (*medium*, *low*), and at $\alpha = .001$ there was a difference between conditions (*baseline*, *low*) and (*high*, *low*). This means the average velocity of each saccade within a trial was significantly affected by reliability of access, and participants made slower saccades as reliability of access decreased, except for between the *high* and *medium* reliability of access conditions.

3.3 Discussion of results

We have shown that there was a clear effect of condition on each of the six variables. This finding lends support for our hypothesis that there is a significant change in participants' behaviour depending on the condition of the task. When we delve deeper, we find that only saccade velocity and peak velocity showed a difference in distributions between conditions *baseline* and *high*. There was also little overall effect across variables between *high*

Table 4: Post-hoc Wilcoxon test *W*-values and *T*-test values for each condition pair. Superscripted ^{a}, ^{b} and ^{c} signify $p < .05$, $p < .01$ and $p < .001$ respectively. Empty cells were not found to be significant and are therefore not reported.

		Condition pair					
	Test	*baseline*, *high*	*baseline*, *medium*	*baseline*, *low*	*high*, *medium*	*high*, *low*	*medium*, *low*
N. of crossings	*W*		39.0^a	91.0^c		55.0^b	64.5^b
Dwell time p/crossing	*T*			-3.1^b		-3.9^b	-2.7^b
Completion time	*T*		-4.8^c	-6.3^c	-3.4^b	-5.7^c	-2.8^b
Fixations p/second	*T*		3.8^b	5.8^c	2.3^a	6.7^c	3.7^b
Saccade velocity	*T*	3.0^b	3.9^b	5.4^c		5.3^c	3.1^b
Errors per trial	*T*	-2.6^a	-3.2^b	-3.6^b		-2.6^a	-2.3^a

and *medium* reliability of access conditions. Not unexpectedly, the effects became more pronounced as the difference in reliability of access between conditions became more pronounced. For instance, the results in the *low* condition were found to differ significantly from all other conditions and on all variables. This may imply that the difficulty in the *low* condition passed some threshold such that a larger change of behaviour was required as compared to between other conditions.

As reliability of access decreased, the data showed a slight increase in the mean number of uncompleted placed items per trial. However, we found the number of unfinished trials to be negligibly low (22 unfinished trials overall), and as such this variable could not be analysed with valid statistical results.

4 Modeling trade-off strategies

Our observational data provides support for the hypothesis that participants adapt their memory strategies as reliability of access changes. However, we cannot directly observe what each of those strategies exactly entail. For example, we cannot directly infer from eye tracking data how many items are encoded in visual working memory at any given moment. To that end, we designed and ran a computational cognitive model, which attempts to simulate participants' behavioural processes which underlie performance on the task. It approaches this problem by breaking down human behaviour into several small processes – such as making an eye movement or encoding an item in VWM – which are executed serially. By varying the model's strategies in terms of a storage/sampling trade-off, we can uncover which strategies lead to outcomes that are most similar to those outcomes observed by measuring human participants.

Our model was designed as a rational performer with systematically controlled variations, based around both theoretical frameworks of memory and observed motor functions. In the following paragraphs, we first describe how the model's parameters were established and subsequently describe the model in general and how the parameters fit in.

Table 5: A subset of all 35 encoding schemes. Encoding schemes range from $k = 1$ to $k = 4$ and are constrained by the rule that, as reliability of access decreases, k may only be greater than or equal to k in the preceding condition.

	Condition			
	baseline	*high*	*medium*	*low*
Scheme 1	[1	1	1	1]
Scheme 2	[1	1	1	2]
Scheme 3	[1	1	1	3]
...	...	...	...	...
Scheme 33	[3	3	4	4]
Scheme 34	[3	4	4	4]
Scheme 35	[4	4	4	4]

4.1 Encoding schemes

Our hypothesis is that in high reliability of access conditions, participants' dominant strategy would be to internally store few items at once and to rely on external sampling. However, in the condition where the reliability of access is low, behaviour would likely shift towards a strategy of relying more on internal storage and memorizing multiple items at once.

An *encoding scheme* represents, per condition, the amount of items k which a participant attempts to memorize while looking at the example grid. In Table 5 a subset of the finite amount of possible encoding schemes is represented, with the limitation that the m-th value of a scheme must always be either greater than, or equal to, the $m-1$-th value, as we expect that more items would be internally stored in low-reliability of access conditions than in high-reliability conditions. In the case of four conditions and k ranging from 1 to 4, there exist 35 unique strategies.

4.2 Memory parameters

The storage capacity and speed of storage in, and retrieval from, working memory is highly dependent on context, such as task difficulty, item complexity and storage recency. As such, we used a simplified form of the memory theory as described in the ACT-R cognitive architecture and which has shown good results in comparison to different memory theories (Anderson & Schooler, 1991; Gray, Schoelles, & Myers, 2003; Gray et al., 2006; Lovett et al., 2012). In this theory, encoded items are subject to decay, activation noise and a linear latency factor, such that the time it takes to retrieve an item RT_i follows

$$RT_i = F \times e^{-a_i} \tag{1}$$

in which F is a factor by which the reaction time is scaled dependent on activation, such that items with low activation tend to require longer retrieval times. The value a_i is then

the activation value of item i, which is calculated as

$$a_i = log_n(\sum_{j=1}^{n} t_j^{-d}) + \varepsilon^2 \tag{2}$$

in which, for each time j the item was activated in memory, t is the time since its activation, which is influenced by decay factor d, which causes items to fade from memory over time. Finally, ε is added as a noise component such that the activation of an item can fluctuate over time. This noise component is drawn over a logistic distribution, centered around 0, and where the variance of the distribution is the parameter to be explored. For simplicity, ε will be used to denote the variance of this factor.

Finally, the memory theory returns whether a retrieval was successful or not, based on whether activation a_i is above a specified threshold T. In case of a failed retrieval, RT_i is still calculated in order to reflect the time that failed retrieval took.

We explore F in the range of (.1, .4) with steps of .1, d in the range of (.5, .9) with steps of .1, threshold T in the range of (.175, .275) with steps of .025, and we explore the variance of ε in the range of (.26, .30) with steps of .02. The first encoding of an item is set to 50 ms, as defined by Lovett et al. (2012).

4.3 Memory rehearsals

As an item is encoded into memory, it may be retrieved one or more times in order to increase its activation in working memory. We expect that, as reliability of access decreases, the importance of correctly memorizing an item increases. Therefore, we explore the option that the number of rehearsals after encoding an item varies across conditions. With the same limitations as in generating the encoding schemes, and the maximum number of rehearsals $r = 3$, we create a set of 15 unique rehearsal schemes.

4.4 Eye- and mouse movements

4.4.1 Eye movements

In order to model the duration of eye movements, we fitted a linear regression to saccade duration as a function of Euclidean distance (in pixels) between the start of a saccade and the end of a saccade, based on observational data. Since we found in Section 3 that participants' mean saccade velocity and peak velocity varied across conditions, we fitted a linear model on saccade data from each condition separately. Accordingly, when the model makes a saccade over a certain distance, the duration of that saccade is calculated as

$$Duration = (a + b \times distance) \times \varepsilon \tag{3}$$

where a and b are taken from the linear model and ε is a noise parameter drawn from a gaussian distribution with $\mu = 1.0$ and $\sigma = .25$.

4.4.2 Mouse movements

Fitts' law is a widely used measure for modeling human mouse movement. In the Shannon and Weaver (1949) formulation this law is defined as

$$ID = log_2(\frac{D}{W} + 1) \tag{4}$$

where D is distance to target and W is the width of the target (Fitts, 1954). ID then represents the difficulty of making a movement towards the target, which increases with distance to the target, but decreases with size of the target. Additionally, Shannon's law in full contains an extension which allows calculation of the expected duration of a movement. We calculate movement time MT as

$$MT = (a + b \times ID) \times \varepsilon \tag{5}$$

where a, b and ε are the intercept, coefficient and noise parameters, respectively. Shannon's variant of Fitts' law is the preferred formula as it provides a better fit to human data than the original (e.g., MacKenzie, 1989, 1992; Shannon & Weaver, 1949; Soukoreff & MacKenzie, 2004). The values for a and b were obtained by fitting a linear model over mouse movement data, for each condition independently. Additionally, ε is drawn from a gaussian distribution with $\mu = 1.0$ and $\sigma = .25$. Clicking and releasing a click are both set to 150 ms, as defined in Gray et al. (2006).

4.5 Computational cognitive model

The proposed computational cognitive model describes a step-by-step theoretical model of how a human performer may execute a trial within the copying task. The model starts by selecting how many items to encode at once and how many rehearsals to perform after encoding, based on the condition and the current encoding- and rehearsal schemes. It will then start the simulation of a trial. A trial consists of two main sub-tasks which repeat until all items are placed correctly or time runs out. See Algorithm 1 for a pseudo-code overview of the proposed model.

Sub-task (1): *Encode k items*, starts with shifting the gaze to the centre of the example grid, the duration of which is calculated with Eq. (3). Then, k items are chosen randomly – as long as they have not been placed yet – in which k is based on the encoding scheme. For each of the chosen items, the gaze is shifted from its current position to the centre of the new item, and the item is encoded in memory (50 ms). The item is then rehearsed in memory r times, based on the rehearsal scheme. The duration of rehearsals is calculated with Eq. (1).

Once k items have been stored or if the example grid disappears, the model moves on to sub-task (2): *Place encoded items*. This sub-task starts with shifting the gaze to the resource grid. Then it retrieves each of the items i which were stored in sub-task (1), and tries to match it to one of the items in the resource grid. It does this by moving its gaze to

an item on-screen, retrieving item i from VWM, and checking whether there is a match. If there is no match, or if the retrieval is unsuccessful, the gaze is shifted to the next on-screen item, and so on, until an item is found which matches i.

Once a match is found, the model moves the computer mouse to the matching item (Eq. 5), picks the item up by clicking on it (150 ms), and moves both the mouse and the gaze to the appropriate location in the workspace grid, where it is dropped (150 ms). Realistically, the process of moving both the mouse and the gaze to the same position would likely not be sequential, but somewhat parallel. However, modeling that process is outside the scope of this research. Once an item is successfully placed, the model retrieves the next item i from VWM and tries to match it to an item again.

Finally, if the example grid is not visible, the model will attempt to place items which are already stored in VWM. If there are no items to place and the example grid is not visible, the model waits until the example grid becomes visible again. In reality, participants may sometimes be going over the placed items once more while waiting, but since the array of options during this waiting period is so extensive, we did not model this for the sake of simplicity.

4.6 Error modeling

Since we found that the number of errors per trial increased as the reliability of access decreased, we implement a probability of error into the model. We consider the mean number of errors per trial for each condition (as reported in Table 2) as the probability that an error is made within a trial. The probability of an error for a single item is then the probability of error for the entire trial, divided by the number of items to be placed.

If an item is placed incorrectly, the model takes steps to fix it. However, modeling the method participants used to correct errors is difficult, as there is a large variation in possible strategies. Participant could realize their mistake immediately, they could notice at the end of the trial, or they could notice halfway through a trial. We simplify approach and assume that a participant notices their mistake immediately.

After a mistake, the incorrectly placed item is right-clicked (150 ms) in order to remove it from the working grid. the gaze is shifted toward the example grid, where the relevant item is again encoded in memory. The gaze and mouse then shift back to the resource grid, from where the item is dragged to its correct position. The model then resumes its process from the point where the mistake was made.

4.7 Modeling methods

The computational cognitive model was designed, run and analysed in Python 3.7.

Our goal was to find which encoding schemes and rehearsal schemes provided results most similar to those observed from human participants. However, the parameters required in the ACT-R theory of memory are not constant throughout different tasks; e.g., memorizing colored blocks may be quicker than memorizing abstract stimuli. Therefore, all parameter combinations needed to be explored by the model. With 35 unique encoding

schemes, 15 rehearsal schemes, and $4 \times 5 \times 5 \times 3 = 300$ memory parameter options, the entire search space consisted of $35 \times 15 \times 300 = 157,500$ parameter combinations. Simple grid search was implemented to find the optimal combination of parameters.

Each parameter combination was run 32 times to account for implemented noise and to produce statistically relevant results. We then compared three outcome variables between participants and the model: (1) *number of crossings*, (2) *completion time in seconds*, and (3) *number of fixations per second*.

Observed data was scaled between 0 and 1 for each outcome variable and each condition separately, to which the simulated data was then matched by mapping it to the same scale. This scaling step was performed in order to standardize the outcome metric for each of the three variables and allow their metrics to be directly compared to each other.

The squared error was calculated per condition for each variable. As such, we could calculate the scaled Root Mean Squared Error (sRMSE) over all four conditions, for each of the three variables. The analysis thus provided three sRMSE values (one per outcome variable), and the model fit for each parameter combination is reported as the mean of these three sRMSE values.

We report the result of the three best models and the worst performing model. Moreover, our goal was to uncover which encoding- and rehearsal schemes provided the best model fit, and finding the ACT-R memory parameters were not directly part of this goal. To ensure the best model's performance was not merely a result of accurate approximation of the memory parameters, we report how the best performing model compares to the mean performance of all other models with the same memory parameters but indifferent of encoding- and rehearsal schemes.

5 Model results

The best model achieved a mean scaled RMSE score of $.1695$, with sRMSE = $.187$ for number of crossings, sRMSE = $.150$ for completion time, and sRMSE = $.171$ for fixations per second. This best performing model used encoding scheme $[1, 1, 2, 3]$, rehearsal scheme $[2, 2, 2, 3]$ and memory parameters $F = .1$, $d = .9$, $T = .175$, $\varepsilon = .30$. Thus, the best model encoded one item after each crossing and rehearsed the item twice in the *baseline* and *high* reliability of access conditions. Subsequently, it attempted those place that single item in the workspace grid before shifting its gaze toward the example grid again. In the *medium* condition it switched to an approach where it encoded two items per crossing, rehearsed them twice, and attempted to place both of those items in the workspace grid. In the *low* reliability of access condition, the model encoded three items after each crossing and rehearsed each item three times in memory. It then attempted to place those three encoded items in the workspace grid before shifting its gaze toward the example grid again. See Figure 4 for a comparison of observational data and the best model's results. We report the results and parameter combinations of the best three models and the worst model in Table 6 and Table 7.

Table 6: Mean of scaled RMSE over all three outcome variables for the best three models and the worst model.

	Scaled RMSE			
Rank	**Mean**	N. of crossings	Completion time	Fixations p/s
1	**.1695**	.187	.150	.171
2	**.1751**	.219	.127	.179
3	**.1769**	.156	.178	.197
...	...	...	...	...
157,500	**.6101**	.254	.903	.673

Table 7: Parameter combinations for the best three models and the worst model.

Rank	Encoding scheme	Rehearsal scheme	F	d	T	ε
1	$[1, 1, 2, 3]$	$[2, 2, 2, 3]$	.1	.9	.175	.30
2	$[1, 1, 2, 3]$	$[1, 2, 2, 2]$	.1	.8	.250	.30
3	$[1, 1, 1, 4]$	$[1, 1, 1, 2]$	.1	.8	.20	.26
...	...	...	...	...	...	...
157,500	$[1, 1, 1, 1]$	$[3, 3, 3, 3]$	.4	.5	.275	.26

Taking the results of all models with the same memory parameters as in the best model but disregarding encoding- and rehearsal schemes ($N = 525$), we found their average performance was worse, sRMSE=.226 (SD=.016). Comparing the best model to all models with both the same memory parameters and the same rehearsal scheme, but disregarding encoding scheme ($N = 35$), their average performance was worse than that of the best model, sRMSE=.224 (SD=.016). Finally, comparing the best model to all models with the same memory parameters and the same encoding scheme, but disregarding rehearsal schemes ($N = 15$), we found their average performance was also slightly worse, sRMSE=.212 (SD=.016). This tells us that disregarding encoding- and rehearsal schemes would deteriorate model performance. Additionally, disregarding the encoding schemes, but with a fixed rehearsal scheme and fixed memory parameters, would decrease the models' average performance more strongly than if the encoding scheme was fixed and the rehearsal schemes were varied.

6 Discussion

We investigated whether reliability of access to visual information influences a trade-off between internal storage and external sampling. Relying on internal storage entails that stimuli are encoded in visual working memory (VWM), whereas external sampling relies on offloading VWM and choosing to sample from the environment instead. Previous studies have provided support for the theory that people generally prefer to externally sample from the environment when the cost of making saccades is relatively low, but that reliance on internal storage increases as the cost of making saccades increases (Ballard et al., 1995; Gray et al., 2006; Inamdar & Pomplun, 2003; Melnik et al., 2018; Somai et al., 2020).

In order to investigate a storage/sampling trade-off in conditions with varying relia-

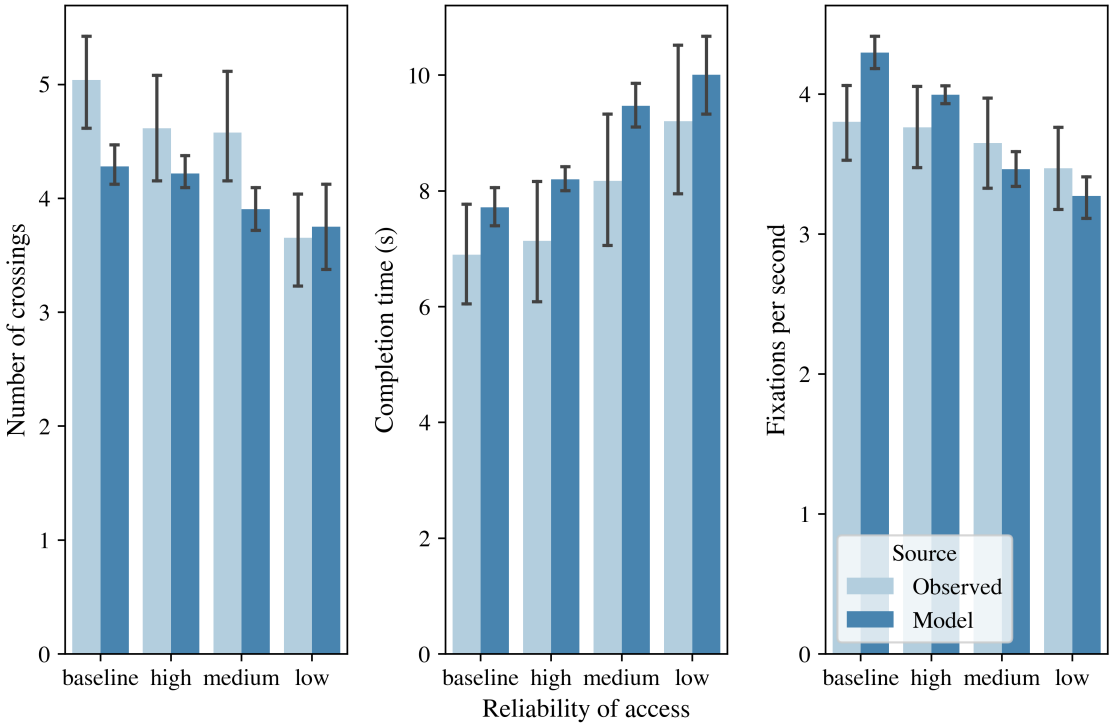

Figure 4: Comparison of observed results versus the best model's results for each of the three outcome variables, per condition.

bility of access to visual information, we designed a study in which participants were subjected to a copying task. In this task, participants copied an example grid to a workspace grid. During a trial, the example grid would disappear and reappear at a set interval. We manipulated the reliability of visual access across conditions by varying the interval with which the example grid disappeared and reappeared.

In this thesis we provide support for our theory that there is an effect of reliability of visual access on the storage/sampling trade-off, which we find to be similar to how varying the cost of saccades affects the storage/sampling trade-off. In our baseline condition, where stimuli were always visible, participants shifted their gaze towards the example grid often and fixated on it for relatively short amounts of time. In the condition with the lowest reliability of visual access, participants shifted their gaze towards the example grid less often than in the baseline condition, but fixated on it for greater amounts of time. This suggests that participants relied primarily on external sampling in the baseline condition, but as reliability of visual access decreased across conditions, participants' reliance on external sampling decreased and they shifted towards relying on internal storage more strongly.

Our second goal in this thesis was to investigate whether a computational cognitive model could be designed to uncover how this storage/sampling trade-off evolved across conditions. We hypothesized that the best model we could find would vary the number of items it encoded – and the number of times it rehearsed those items after encoding –

across conditions.

We designed our cognitive model to as closely mimic human behaviour on the copying task as possible. We then tested each model by comparing the outcomes of three variables between human results and model results. We found that, in the baseline condition, our best-fitting model encoded one item from the example grid and then directly placed it in the workspace grid before shifting its gaze back to the example grid and encoding the next item. In contrast, the same model encoded three items at a time in the *low* reliability of access condition. It would attempt to place those three items in the workspace grid and then shift its gaze back to encode and place the remaining item. We pose that this supports the theory that external sampling is the preferred strategy in conditions with high reliability of access, but that the strategy shifts towards reliance on internal storage as reliability of access decreases. Furthermore, in the baseline condition, this best model rehearsed each item twice in memory after encountering it for the first time. In the low reliability of access condition, it rehearsed each item three times, which implies that the accuracy of encoding items also plays a role in the storage/sampling trade-off.

We found that models with different encoding- and rehearsal schemes than that of the highest-ranked model performed worse overall. This finding provides evidence that the inclusion of storage/sampling strategies was an important factor in our model. Furthermore, we noticed that models with no internal variation within their encoding scheme tended to perform worse than those that did have internal variations in in their encoding scheme. These findings lead us to believe that encoding- and rehearsal strategies indeed change dependent on reliability of visual access within the environment.

We found that the best model slightly overestimated the completion time of trials, but that it did accurately reflect the increase in completion time as the reliability of access decreased. The model approximated the number of fixations per second participants made in the task fairly well, and did somewhat accurately capture the decrease of fixations per second as reliability of access decreased. For a better fit on *completion time* and *fixations per second*, the latency scaling factor F would need to be smaller than .1, the decay rate d would need to be greater than .9, or both. However, we could find no experimental evidence for values in that range.

Furthermore, the best model did not accurately estimate the number of crossings made from the right side of the screen towards the example grid, although it did capture the decrease in number of crossings as reliability of access decreased. Given the deviations between model results and human performance, and that the memory parameters pushed the limits of experimental evidence, we recommend that future research investigate refinements of the model in order to reduce the degree to which it simplifies the cognitive process.

Analysis of the computational cognitive model was limited by the fact that we incorporated three different variables over which the mean error was calculated. This approach did indeed find a model which performed the best overall, but we also found that models with different parameter combinations may have been optimised on different outcome

variables. This is noticeable in Tables 6 and 7, where models with rank 1 and 2 scored best on *completion time*, but the model with rank 3 scored best on *number of crossings*. Furthermore, as the mean errors of these models lay very close together, and the model has built-in noise components, it is not possible to say with confidence which of these models represents the cognitive processes underlying human performance most accurately. It could be mere luck of the draw that one model performed slightly better than its successor. Nonetheless, our findings provide an indication of what the true trade-off may be.

The variance exhibited in human behaviour was higher than the variance exhibited by the model. This can be partially explained by the fact that there was a clear difference in performance between human participants; some were quicker than others and thus showed lower completion times and a greater number of fixations per second. It is evident that some participants were quicker at memorizing items, were more comfortable with handling the computer mouse, or may have used different encoding- and rehearsal strategies altogether. We did not model for the existence of between-participant differences; the model results we discussed in this thesis were calculated over 32 trials per parameter combination with the only variance intentionally being introduced by way of our noise components.

Additionally, when fitting linear regressions to our observed saccade data, there were small groups of saccades which spanned relatively short distances, but which had long durations (e.g., 300 ms over a 10° visual angle). These observations could not be explained by grouping them as smooth pursuit eye movements (i.e., dragging an item along the screen and visually following it), nor could they be ascribed to one or two participants making especially slow saccades. Finally, we determined some of these saccades were 'turn-around saccades'; an eye movement that travels in multiple directions without a fixation being detected in between. Since the prevalence of this type of saccade was relatively low, we did not incorporate them in our model, but for more accurate findings it may bear future incorporation.

We make two recommendations for future to the experiment discussed in this thesis. In the current research, some participants, and in some trials, could memorize all four stimuli with just one crossing from the right- to the left side of the screen. This seemed to nudge those participants somewhat towards trying to complete trials with a single crossing, which came at the cost of accuracy. The first of our recommendations is therefore to explore the optimal number of stimuli within a trial where the condition with the lowest reliability of access exceeds some threshold such that not all stimuli can be memorized at once. Moreover, as mentioned in Section 3.3, the observed difference in behaviour between the *baseline* and *high* reliability of access conditions and between the *high* and *medium* conditions was less pronounced than between the *medium* and *low* conditions. This may imply that the *low* reliability of access condition exceeded some threshold for a bigger change in behaviour. Ideally, each condition passes such a threshold so that behavioural change can be more distinctly observed. The second recommendation we therefore make is to explore bigger variations in reliability of access between conditions.

In conclusion, we found a clear effect of reliability of visual access on behaviour within our copying task. Accordingly, we found that the performance of our best-fitting model as compared to other models could be accounted for by the incorporation of this storage/sampling trade-off. This implies that reliance on visual working memory changes as the reliability of visual access changes. Although many computational cognitive models already incorporate or explore some form of strategy selection, it bears taking into account that humans may select memory strategies based on the reliability of visual access and that this theory is incorporated in cognitive modeling efforts going forward.

The author would like to thank Stefan van der Stigchel and Tanja Nijboer for their advice and supervision. The author thanks Sanne Böing, Roderic Hillege, Chris Janssen, Timo Kootstra, Andre Sahakian, and the Attentionlab team for their feedback and advice, and Son Luong for the lab support. The code for this thesis is publicly available at https://github.com/higher-bridge/copying-task-uu

References

Anderson, J. R. (1996). A Simple Theory of Complex Cognition. *American Psychologist*, *51*(4), 355–365. doi: 10.1037/0003-066X.51.4.355

Anderson, J. R., & Schooler, L. J. (1991, nov). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408. doi: 10.1111/j.1467-9280.1991.tb00174.x

Arnoult, M. D. (1956, apr). Familiarity and recognition of nonsense shapes. *Journal of Experimental Psychology*, *51*(4), 269–276. doi: 10.1037/h0047772

Baddeley, A. D. (2000, nov). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. doi: 10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Herring, S. R. (1983). Working memory. *Philosophical Transactions of the Royal Society of London. Biological Sciences*, *B 302*(1110), 311–324.

Baddeley, A. D., & Hitch, G. (1974). Working memory. The psychology of learning and motivation. *New York, NY: Academicp*.

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

Brumby, D. P., Howes, A., & Salvucci, D. D. (2007). A cognitive constraint model of dual-task trade-offs in a highly dynamic driving task. In *Conference on human factors in computing systems - proceedings* (pp. 233–242). New York, New York, USA: Association for Computing Machinery. doi: 10.1145/1240624.1240664

Cary, M., & Carlson, R. A. (2001). Distributing working memory resources during problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 836.

Charman, S. C., & Howes, A. (2003). The adaptive user: an investigation into the cognitive and task constraints on the generation of new methods. *Journal of experimental psychology: Applied*, *9*(4), 236.

Cowan, N. (2016). *Working memory capacity: Classic edition*. Psychology press.

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014, dec). PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods*, *46*(4), 913–921. doi: 10.3758/s13428-013-0422-2

Di Stasi, L. L., Antolí, A., & Cañas, J. J. (2011, nov). Main sequence: An index for detecting mental workload variation in complex tasks. *Applied Ergonomics*, *42*(6), 807–813. doi: 10.1016/j.apergo.2011.01.003

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., … Pannasch, S. (2010, apr). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation Space and Environmental Medicine*, *81*(4), 413–417. doi: 10.3357/ASEM.2579.2010

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, *47*(6), 381.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. doi: 10.1080/01621459.1937.10503522

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi: 10.1126/science.aac6076

Gray, W. D., Schoelles, M. J., & Myers, C. W. (2003, oct). Meeting Newell's other challenge: Cognitive architectures as the basis for cognitive engineering. *Behavioral and Brain Sciences*, *26*(5), 609–610. doi: 10.1017/S0140525X03280134

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological review*, *113*(3), 461.

Howes, A., Duggan, G. B., Kalidindi, K., Tseng, Y. C., & Lewis, R. L. (2016). Predicting Short-Term Remembering as Boundedly Optimal Strategy Choice. *Cognitive Science*, *40*(5), 1192–1223. doi: 10.1111/cogs.12271

Inamdar, S., & Pomplun, M. (2003). Comparative search reveals the tradeoff between eye movements and working memory use in visual tasks. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25, pp. 599–604).

Janssen, C. P., & Gray, W. D. (2012). When, what, and how much to reward in reinforcement learning-based models of cognition. *Cognitive science*, *36*(2), 333–358.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311. doi: 10.1111/tops.12086

Lovett, M. C., Reder, L. M., & Lebiere, C. (2012, jun). Modeling Working Memory in a Unified Architecture: An ACT-R Perspective. In *Models of working memory* (pp. 135–182). Cambridge University Press. doi: 10.1017/cbo9781139174909.008

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391–400.

Ma, W. J., Husain, M., Bays, P. M., & de Soissons, P. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347. doi: 10.1016/S0031-9406(10)63634-6

MacKenzie, I. S. (1989). A note on the information-theoretic basis for Fitts' law. *Journal of motor behavior*, *21*(3), 323–330.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction*, *7*(1), 91–139.

McClelland, J. L. (2009, jan). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, *1*(1), 11–38. doi: 10.1111/j.1756-8765.2008.01003.x

Melnik, A., Schüler, F., Rothkopf, C. A., & König, P. (2018). The world as an external memory: the price of saccades in a sensorimotor task. *Frontiers in behavioral neuroscience*, *12*, 253.

O'Regan, J. K. (1992). Solving the" real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *46*(3), 461. doi: 10.1037/h0084327

Python Core Team. (2019). Python: A dynamic, open source programming language

[Computer software manual]. Retrieved from https://www.python.org/

Riverbank Computing Limited. (2019). PyQt5 [Computer software manual]. Retrieved from https://www.riverbankcomputing.com/software/pyqt/

Russell, S. J., & Subramanian, D. (1995, may). Provably Bounded-Optimal Agents. *Journal of Artificial Intelligence Research*, *2*, 575–609. doi: 10.1613/jair.133

Salway, A. F. S., & Logie, R. H. (1995, may). Visuospatial working memory, movement control and executive demands. *British Journal of Psychology*, *86*(2), 253–269. doi: 10.1111/j.2044-8295.1995.tb02560.x

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., & Ward, J. L. (1967). The attentional demand of automobile driving.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information. *Urbana: University of Illinois Press*, *97*.

Somai, R. S., Schut, M. J., & Van der Stigchel, S. (2020). Evidence for the world as an external memory: A trade-off between internal and external visual memory storage. *Cortex*, *122*, 108–114. doi: 10.1016/j.cortex.2018.12.017

Soukoreff, R. W., & MacKenzie, I. S. (2004, dec). Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human Computer Studies*, *61*(6), 751–789. doi: 10.1016/j.ijhcs.2004.09.001

Vallat, R. (2018, November). Pingouin: statistics in python. *The Journal of Open Source Software*, *3*(31), 1026.

Van der Stigchel, S. (2020). An embodied account of visual working memory. *Visual Cognition*, 1–6.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636. doi: 10.3758/BF03196322

A Appendix: Algorithm

Algorithm 1 Computational cognitive model

for EACH CONDITION **do**
  SELECT k FROM $encoding\_scheme$ BASED ON CONDITION
  SELECT r FROM $rehearsal\_scheme$ BASED ON CONDITION

  procedure TRIAL
    SET LOCATION OF GAZE TO CENTRE OF SCREEN
    repeat
      if $k >$ NUMBER OF REMAINING (UNPLACED) ITEMS **then**
        $k \leftarrow$ NUMBER OF REMAINING (UNPLACED) ITEMS

      procedure SUB-TASK 1: ENCODE k ITEMS
        SHIFT GAZE TO EXAMPLE GRID ▷ Eq. (3)
        repeat
          PICK A RANDOM UNPLACED ITEM
          MOVE GAZE TO NEW ITEM ▷ Eq. (3)
          ENCODE ITEM IN VWM ▷ 50 ms
          REHEARSE ITEM IN VWM (r TIMES) ▷ Eq. (1)
        until k ITEMS STORED OR EXAMPLE GRID DISAPPEARS

      procedure SUB-TASK 2: PLACE ENCODED ITEMS
        SHIFT GAZE TO RESOURCE GRID ▷ Eq. (3)
        for EACH STORED AND UNPLACED ITEM i IN VWM **do**
          repeat
            MOVE GAZE TO NEW ITEM ON SCREEN ▷ Eq. (3)
            TRY RETRIEVING ITEM i FROM VWM ▷ Eq. (1)
          until VIEWED ITEM AND ITEM i IN VWM MATCH
          MOVE MOUSE TO TARGET STIMULUS ▷ Eq. (5)
          PICK UP ITEM [CLICK] ▷ 150 ms
          SHIFT GAZE TO WORKSPACE GRID ▷ Eq. (3)
          DRAG MOUSE TO WORKSPACE GRID ▷ Eq. (5)
          DROP ITEM [RELEASE CLICK] ▷ 150 ms

          if PLACEMENT IS INCORRECT **then** FIX MISTAKE ▷ Section 4.6

      if EXAMPLE GRID IS OCCLUDED AND NO UNPLACED ITEMS IN VWM **then**
        WAIT UNTIL EXAMPLE GRID REAPPEARS

    until ALL ITEMS PLACED OR TIME RUNS OUT