The Interplay of Verbal and Visual Information in Cognition and the Role of Intention

Molly A. Delooze

Supervised by Dr. Candice C. Morey and Dr. Dominic Guitard

A thesis submitted for the degree of Doctor of Philosophy

2025

Thesis Summary

This thesis reports multiple experiments from across the field of cognitive psychology which utilise both verbal and either spatial or colour stimuli to investigate how visually presented verbal information interacts with other types of visual information. The first two experimental chapters consist of investigations into the effects of response planning on Stroop-like interference, both when stimuli are presented together and when they are presented apart. These experiments use original verbal-spatial stimuli, and the results have implications both for theories of Stroop interference and for how verbal-spatial Stroop studies can be conducted in the future. The third experimental chapter is a paper published in *Memory & Cognition* with my supervisors and with Professor Nelson Cowan of the University of Missouri, in which we replicated and extended a recent experimental finding of surprisingly rapid forgetting of source information using verbal and colour stimuli. The results importantly corroborate the original finding and support the argument that experiments utilising a 'surprise trial' should be better understood. The fourth experimental chapter is an exploratory analysis of data from a new experimental paradigm used to assess incidental verbal-spatial binding, which was created in response to issues with the previous paradigm of data attrition and failures to replicate key findings. The major implications of this study were that incidental verbal-spatial binding can be detected using a method such as this, and that future analyses on the subject of binding asymmetry should consider task order as a potential factor. Reflection on these findings as a whole suggests that visually presented verbal information is sometimes dominant over other visual stimuli, depending on participants' intentions regarding the responses they will give.

Contents

List of Figures and Tables

Acknowledgments and Thanks

My supervisors, Candice and Dominic for their encouragement and support, and for inspiring me to do this every day. Candice, thank you for inviting me on this journey - it has been one of the best decisions of my life.

Lisa Evans for her brilliant support as my internal assessor (and as a teaching mentor) and Tom Freeman for looking after all the postgraduate students as if they were his own.

All the members of our working memory lab for their support, but especially Teodor Nikolov for all his guidance on anything and everything, and for being a fantastic role model. My postgraduate and early career colleagues for both the academic and emotional support they have provided.

Chris Jarrold and his lab at Bristol University for their feedback and encouragement at joint lab meetings. Professor Nelson Cowan for his wisdom on multiple projects and for building such a wonderful academic ‘family’.

Nia Jones, my undergraduate student intern of 2024, for her help with data collection in association with the project in Chapter 3 and Lotje van der Linden for experiment programming work done in association with the project in Chapter 5.

My family, especially my selfless mother for her unflappable belief in me and in everything I do, and my fiancé for making anything seem possible and helping me to keep my head above water when things get tough. I could not have done this without you.

Disclaimer

The paper listed below has been included in this thesis:

Delooze, M. A., Guitard, D., Cowan, N., & Morey, C. C. (2024). Rapid source forgetting across modalities: A problem for working memory models. *Memory & Cognition*, 1-16. https://doi.org/10.3758/s13421-024-01664-y

Publication status: published work

Publisher's permission: N/A (Open Access)

Presented as Chapter 4.

MAD contribution: 85%

MAD developed the methodology with the assistance of DG (who programmed the study) and CCM. MAD carried out the analyses and led the writing of the paper with input from all co-authors, and advice from reviewers.

1. Introduction

There is evidence from across a wide range of cognitive phenomena that different types of information are not treated equivalently by our minds, specifically that visually presented verbal information seems to be disproportionately persistent compared to other visual stimulus types. Firstly, there is some evidence to suggest that specifically in verbal-spatial memory, associated spatial information may be retained 'for free' over a short period when we make an effort to commit verbal information to memory, but that the reverse does not occur (Campo et al., 2010; Elsley & Parmentier, 2015; Chapter 5 of this thesis; though see Delooze et al., 2022 for a collection of studies wherein this was not the case). It is also commonly demonstrated (e.g., in Stroop, 1935, see Chapter 3 for a fuller discussion) that to-be-ignored verbal information can persist much more strongly than can to-be-ignored colour information. In tasks wherein participants see colour words printed in coloured ink and must respond either to the meaning of the word or the colour of the ink, the response slowing caused by misleading verbal information is generally far greater than that caused by misleading ink colour information. In contrast, this persistence is considerably more equal in versions of the task where verbal stimuli are pitted against spatial stimuli instead. Along similar lines, cognitive researchers have developed a relatively consistent method of inhibiting the action of the verbal rehearsal system (articulatory suppression), but spatial memory is unreliably affected by various tasks designed to inhibit it, such as spatial tapping (e.g., Zimmer et al., 2003), again demonstrating an inequality. Lastly, the phenomenon of visuo-spatial bootstrapping (e.g., Darling & Havelka, 2010) suggests that when a sequence of digits is presented in the form of a familiar spatial layout (e.g., a numbered response pad stored in long-term memory), recall is more accurate compared to when it is presented in a novel spatial layout. However, to the best of our knowledge, there is no evidence for the reverse: familiar verbal structures in long-term memory aiding in the short-term recall of spatial information. From a variety of methods, a common theme emerges: verbal information is frequently treated differently from other information types across cognition. Perhaps with sufficient exploration, we might be able to tie these imbalances together in a meaningful way.

To explain the inconsistency in the verbal-spatial binding they observed, Elsley and Parmentier (2015) favoured what they termed a ‘strong asymmetry’, attributing the binding of task-unnecessary verbal information to task-necessary spatial information (and not the reverse) to the positions of verbal and spatial information within a stimulus hierarchy. They suggest that stimulus types which are higher within the hierarchy cannot bind to stimulus types which are beneath them, so the lack of binding of letters to locations would indicate that spatial information sits below verbal information within the hierarchy. We find this to be an interesting suggestion and want to explore whether there is evidence for such a hierarchy in a wider sphere of cognitive phenomena.

Major themes

Feature binding is the process of integrating the cognitive representations of two or more features or attributes of a stimulus (for instance, its shape and its colour) into one unit. Historically, cognitive psychologists have disagreed about whether the capacity of memory is measured by the number of individual features or the number of integrated units, which is important to establish for memory models. Additionally, the mechanism behind feature binding is also of key interest for study because of its implicated role in various psychological disorders, such as Dementia (e.g., Parra et al., 2009) and Schizophrenia (e.g., Burglen et al., 2004). A better understanding of how binding works may in turn improve our understanding of these conditions and lead to potential treatments.

Working memory is a branch of memory concerned with maintaining information in the short term in service of a particular goal or to guide an action. There have been numerous models and theories created to try to explain its workings since the term was first coined in the 1970s (Baddeley et al., 1974), many of which take vastly different approaches from one another on the subjects of methods of storage, mechanisms of forgetting and mechanisms of remembering. For instance, some models suggest that different types of information are stored in working memory differently, which affects the way that items interact across and within domains (e.g., Baddeley et al., 1974), whereas other models (e.g., Cowan, 1988) take a firm domain-general stance to explaining working memory, suggesting

that there is one ‘store’ for information of all types. Similarly, models can vary considerably in how they explain forgetting. Some propose that forgetting is time-sensitive, with forgetting occurring naturally as a result of time elapsed since encoding (e.g., Barrouillet, Bernardin & Camos, 2004), and others prioritise a limit to capacity, beyond which new items are either not encoded or old items are lost (e.g., Popov & Reder, 2020).

Interference in the realm of cognitive psychology is the detrimental effect exerted by to-be-ignored information on a response. Usually, a person’s response to a target stimulus is slowed (or sometimes becomes more error-prone) as a result of an interfering stimulus or feature which implies a different meaning, compared to responses made when the accompanying feature implies the same or a neutral meaning. The Stroop effect (Stroop, 1935) is a well-established psychological phenomenon which is an example of interference in cognition. A classic example of a stimulus used within a Stroop task is the word “red” written in blue font: the meaning of the word is *red*, but the meaning of the font colour is different - *blue*. This example of a Stroop stimulus where the meanings of its features (the word and the font colour) are different from one another is known as an “incongruent” stimulus. If instead, the word “red” were presented in red font, the stimulus would be considered “congruent”, because the meanings of the two features are the same. The incongruent stimulus would exert an interference effect on the response given.

Experimental chapters: motivations and goals

Chapter 2 – The Compass Task: A New Direction for the Verbal-Spatial Stroop task

Spatial Stroop tasks are those where the stimuli have both a verbal feature whose meaning pertains to space (e.g., the words “up” and “down”), and a location or directional feature (e.g., appearing at the top or bottom of the screen). These tasks have shown promise in eliciting bi-directional interference, meaning that the tasks of judging both verbal meaning and spatial meaning are impacted by an incongruent to-be-ignored feature, meanwhile colour Stroop tasks produce somewhat inconsistent results in this regard. The way in which bi-directional

interference is achieved in spatial Stroop tasks is by utilising different response methods: one that complements the verbal and one that complements the spatial element of the stimuli. Virzi and Egeth's (1985) translational model of Stroop interference states that different types of stimuli (e.g., colour or word meaning or location) are processed and output by unique processing systems which sit in parallel with one another. Interference occurs when the stimulus and response contradict because the input signal from the stimulus must be translated from one parallel processing system to another to be output as the response. This takes time compared to when a signal can pass straight from stimulus input to response output without leaving its processing system.

The first goal of the experiments detailed in the first experimental chapter was to use the spatial Stroop task which has previously been studied using many different stimuli and methods, and to improve upon it by using simplistic stimuli which lend themselves to straightforward responses. Moreover, we wanted to design a version of the task which would enable us and other researchers interested in this paradigm to run the experiment online, so an important part of this chapter was the comparison between the results of the online version of the task (Experiment 1) and the in-lab version of the task (Experiment 2). The second goal of these experiments was to establish whether Virzi and Egeth's (1985) translational model findings could be replicated using keypress responses (with verbal associations) in place of the vocal response which they utilised in their method. To elaborate, evidence for this would appear in the data as an interference effect only when the type of stimulus to be judged (e.g., location) was in contrast to the type of response to be given (e.g., letter keypress), and not when these two factors match. This had implications for their model because if letter key responses could prove successful in this regard, it would demonstrate that more than one response could be directly connected to their parallel processing systems (i.e. be accessed without the need for translation). Further, it would demonstrate that responses which are physically similar (e.g., pressing letter keys and pressing arrow keys) could be distinct from one another with regard to their links to the processing systems. If we found that more than one response can be directly mapped to a processing system without the need for translation, this opens the floor to asking *how* does a new response become mapped to such an extent that no translation need occur? The final goal of these experiments

(addressed only in Experiment 2) was to assess whether the central image of the compass diagram was integral to eliciting (or at the least involved in boosting the extent of) Stroop-like interference.

Chapter 3 – The Compass Task 2: A Working Memory-Stroop Hybrid

Kiyonaga and Egner (2014) demonstrated in their working memory-Stroop hybrid task that by asking their participants to maintain a colour word in working memory while judging the identity of a colour patch, they could elicit slowed responses when the word and patch had incongruent meanings compared to when they had congruent ones. This is an example of Stroop-like interference taking place even when the interfering item was no longer observable. They compared this working memory Stroop effect to another subset of their data demonstrating the simultaneous Stroop effect and found that it was of almost identical size. They also found that this working memory Stroop effect was susceptible to two of the same factors which affect the simultaneous Stroop effect. First, when participants in the working memory Stroop task experienced a smaller proportion of incongruent trials than congruent trials, the interference effect was larger. Secondly, when participants were required to press the same response key to respond to multiple different colour patches, they showed evidence of additive interference as a result of stimulus incongruence and response incongruence. From these findings, Kiyonaga and Egner (2014) concluded that the working memory Stroop effect which they observed in their method was the same as the simultaneous Stroop effect. To examine this claim, we ran the three experiments detailed in Chapter 3.

Our first goal was to convert our compass task into a working memory-Stroop hybrid task and establish whether the Stroop-like interference observed in Kiyonaga and Egner's (2014) method could also be seen using verbal-spatial stimuli. Having established in Chapter 2 that the hypotheses outlined in Virzi and Egeth's (1985) translational model are resoundingly supported by data from the simultaneous version of the compass task, we wanted to know whether we would also observe this to be true in the working memory-Stroop version of the task. If Kiyonaga and Egner's (2014) claim that their working memory Stroop effect is the same as the well-known simultaneous Stroop effect is true, we should also be able to observe the data

pattern characteristic of the translational model here. To elaborate, this means that we should see greater response slowing in the judgment task due to incongruent pairs of stimuli when participants must make a response that does not complement the judgment stimulus compared to when they must make a response that does complement it. For example, if participants must make a spatial judgment response (pressing arrow keys) about a letter stimulus, they should be slowed much more when the letter stimulus did not semantically match the location stimulus which they just committed to memory compared to if they were to make a verbal judgment response (pressing letter keys). We first tested this in Experiment 1 using a version of the task which maintained Kiyonaga and Egner's (2014) original recognition-style memory response.

However, Experiment 1 produced inconsistent results across the two domains, and analyses revealed only small effect sizes where the effects of interest to the translational model were significant. From several studies which demonstrated that recognition memory is essentially different from recall memory (e.g., Hall et al., 1976, but see Chapter 3 for a fuller discussion), we hypothesised that we may see interference that is more similar to that documented in simultaneous Stroop tasks if participants intended to recall the interfering information rather than to recognise it. Therefore, the goal of Experiment 2 was to assess whether the replacement of that recognition response with a recall response would produce data which demonstrated the pattern predicted by the translational model more clearly. We hypothesised that if this manipulation were successful, that its mechanism of effect would likely work in one of two ways. This mechanism could either work by increasing the strength of the encoding of the memory item thereby increasing its interference strength, or by reinstating the complementary response planning to the interfering item which we posit occurs very quickly and automatically in the simultaneous Stroop task.

The final goal of the experiments reported in this chapter was to determine whether a vocal response would prove to be a superior 'verbal response' compared to the letter key response type which we had used throughout these spatial Stroop experiments. A vocal response was used in Virzi and Egeth's (1985) study which provided support for their model, and though the experiments reported thus far using a 'stand-in' verbal response had mostly proven successful in eliciting the expected effects, there were a handful of minor unanticipated artefacts within the data which

begged that the comparison be made. Therefore, in Experiment 3, the letter key response type was replaced with a spoken response into a microphone which was recorded and transcribed.

Chapter 4 - Rapid Source Forgetting Across Modalities: A Problem for Working Memory Models

Chen et al. (2018) demonstrated an interesting forgetting phenomenon wherein less than 1s after seeing a stimulus, participants were extremely poor at recognising it from a probe. However, if they had seen that stimulus marginally earlier, they were very accurate at recognising it. Their method involved very briefly presenting pairs of stimuli to participants consisting of a coloured square and a colour word and asking them to indicate whether or not the meanings of the two stimuli were congruent or incongruent. After a set number of trials like this, participants were surprised at the end of one trial by being asked not to make a comparison of the items, but instead to recognise from a set of four options which coloured square they had just seen. When the coloured square had been the first of the two stimuli to be presented on that trial, participants were very accurate in their recognition. However, when the coloured square had been the second of the two stimuli to be presented on that trial (and therefore was seen more recently), participants were very poor in their recognition performance.

An important non-experimental goal for this chapter was to begin a discussion of this phenomenon with respect to some prominent current models of working memory. Specifically, we tried to consider which elements of the effect can and cannot be explained by the various models. To highlight the importance of discussing this phenomenon, our first experimental goal in this chapter was to replicate the findings reported in Chen et al.'s (2018) second experiment. We wanted to ensure that our experimental program was an adequate reflection of theirs before we made any adaptations to it, and to provide evidence in a new group of participants for the effect, plus, replicating findings is an essential part of modern science.

Next, we ran a version of the experiment which was identical apart from the stimulus which was tested: instead of probing participants' memory for the identity of the coloured square, we were interested in their memory for the colour word. We suspected that participants may remember the identity of the colour word better than

they did the coloured square due to evidence from colour Stroop experiments which use very similar stimuli. Generally, researchers detect stronger Stroop-like interference when participants try to respond to font colour but ignore word meaning than they do when participants try to respond to the word and ignore the font colour (e.g., Gumenik & Glass, 1970; Chmiel, 1984, for a fuller discussion, see Chapter 3). This may speak to an inherent difference in the capacity for interference of these stimulus types, that the interference that words can exhibit on judgments of colours is greater than the interference that colours can exhibit on judgments of words. Therefore, our second experimental goal was to determine whether the very fast forgetting seen in this paradigm would occur to a lesser extent if memory for words was tested.

The goal of the third experiment in this chapter was to test a hypothesis that had been borne out of our discussion of this phenomenon in relation to memory models. Oberauer and Lin's (2017; 2023) Interference model of working memory posits that the way in which item information (in this case, the semantic meaning of the colour which is depicted by the word or coloured square stimulus) is retrieved is through activation of the item's context. Therefore, for an item's meaning to be retrieved, its context *must* be known. We had imagined in this paradigm that 'context' would equate to the item's source (whether it had been presented as a word or as a square), which participants did not seem to be successful in recalling when the item had been presented second. However, we recognised that it was possible that a different type of context might be used instead. Since all stimuli were presented in the same location on-screen and therefore this could not be used to distinguish them, the most likely candidate was the item's serial order information: whether it was presented first or second. Therefore, in Experiment 3, instead of asking participants to recognise which colour or word they saw on the surprise trial, we asked them which item they saw first and which item they saw second. In this way, we tested for evidence of the retention of a different type of context which may have been utilised as a retrieval cue, in-line with the Interference model.

Chapter 5 - A New Approach to Measuring Verbal-Spatial Binding Asymmetry

In the realm of feature binding studies, both Elsley and Parmentier (2015) and Campo et al. (2010) found evidence that unintended binding was more likely to occur under some conditions than others. When their participants were tasked with remembering only the identities of letters within a memory array, they were much more likely to also accidentally remember the locations in which those letters appeared despite being directed to ignore this information. This is indicated by faster (and/or more accurate) responses to the to-be-recognised probe item when both the letter and location information was maintained compared to when only the feature of focus was maintained. Meanwhile the reverse was not true, when participants were tasked with remembering only the locations in which letters appeared, they were not at all likely to also remember the identities of the letters. However, evidence of this asymmetry was not provided by the data reported in Delooze et al. (2022) which came from two experiments using very similar methods to those mentioned above. One of the goals of the experiment reported in this chapter was to once again probe this phenomenon and provide further evidence for or against the binding asymmetry which was demonstrated in two of the previous studies in this area.

An issue with the paradigm that all three of these previous studies used is that it discards a lot of data. This high rate of data removal is due to the necessity of the inclusion of trials wherein participants must give ‘no’ responses to recognition probes, despite the fact that the only trials which are useful in measuring binding are those wherein a ‘yes’ response must be given. The consequence for the participant is that many more trials are required than would be expected, which can cause fatigue and potentially impact the quality of the data produced. The experiment detailed in this final experimental chapter was an exploration of a new method for measuring incidental binding which does not rely on such a small proportion of the collected data. To measure participants’ memory performance, our new paradigm asked participants to recall all the items in the memory array by dragging four of the possible letters into four of the possible locations, instead of asking participants to indicate whether they recognise a single probe item. If a participant placed the ‘correct’ letter in its ‘correct’ location, that demonstrated binding in either task. By

testing memory for all four display items at once, this paradigm can feasibly collect much richer data, none of which need be discarded because whether binding occurred or not was inherent to every trial's response. We hoped that this richer data might also be able to provide a candidate for explanation of why our attempts to replicate may have failed in the past.

Foreword

Altogether, these studies shed new light on the interplay of visually presented verbal information and other visual stimuli, taking into consideration how intentions to remember and the nature of our planned responses influence these relationships. This work explores this verbal-visual imbalance across the themes of interference, working memory and feature binding. Building on the experimental chapters, a final chapter will discuss and speculate on what can be drawn from our findings as a whole.

References

Baddeley, A. D., Hitch, G. J., & Bower, G. A. (1974). Working memory. *Recent advances in learning and motivation, 8*, 47-89. https://doi.org/10.1016/S0079-7421(08)60452-1

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Burglen, F., Marczewski, P., Mitchell, K. J., Van der Linden, M., Johnson, M. K., Danion, J. M., & Salame, P. (2004). Impaired performance in a working memory binding task in patients with schizophrenia. *Psychiatry research, 125*(3), 247-255. https://doi.org/10.1016/j.psychres.2003.12.014

Campo, P., Poch, C., Parmentier, F. B. R., Moratti, S., Elsley, J. V., Castellanos, N. P.,…Maestú, F. (2010). Oscillatory activity in prefrontal and posterior regions during implicit letter-location binding. *Neuroimage, 49*, 2807-2815.

Chen, H., Carlson, R. A., & Wyble, B. (2018). Is Source Information Automatically Available in Working Memory? *Psychological Science*, *29*(4), 645–655. http://doi.org/10.1177/0956797617742158

Chmiel, N. (1984). Phonological recoding for reading: The effect of concurrent articulation in a Stroop task. *British Journal of Psychology, 75,* 213-220. https://doi.org/10.1111/j.2044-8295.1984.tb01894.x

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104*, 163-191. https://doi.org/10.1037/0033-2909.104.2.163

Darling, S., & Havelka, J. (2010). Visuospatial bootstrapping: Evidence for binding of verbal and spatial information in working memory. *Quarterly Journal of Experimental Psychology*, *63*(2), 239-245. https://doi.org/10.1080/17470210903348605

Delooze, M. A., Langerock, N., Macy, R., Vergauwe, E., & Morey, C. C. (2022). Encode a letter and get its location for free? Assessing incidental binding of verbal and spatial features. *Brain Sciences*, *12*(6), 685. https://doi.org/10.3390/brainsci12060685

Elsley, J. V. & Parmentier, F. B. R. (2015). Rapid Communication: The asymmetry and temporal dynamics of incidental letter-location bindings in working memory. *Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://dx.doi.org/10.1080/17470218.2014.982137

Gumenik, W E., & Glass, R. (1970). Effects of reducing the readability of the words in the Stroop Color-Word Test. *Psychonomic Science, 20,* 247-248. https://doi.org/10.3758/BF03329047

Hall, J. W., Grossman, L. R, & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4* (5), 507-513. https://doi.org/10.3758/BF03213211

Kiyonaga, A., & Egner, T. (2014). The working memory Stroop effect: When internal representations clash with external stimuli. *Psychological science*, *25*(8), 1619-1629. https://doi.org/10.1177/0956797614536739

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological review*, *124*(1), 21. https://doi.org/10.1037/rev0000044

Oberauer, K., & Lin, H.-Y. (2023). An interference model for visual and verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001303

Parra, M. A., Abrahams, S., Fabi, K., Logie, R., Luzzi, S., & Sala, S. D. (2009). Short-term memory binding deficits in Alzheimer's disease. *Brain*, *132*(4), 1057-1066. https://doi.org/10.1093/brain/awp036

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127(1),* 1–46. https://doi.org/10.1037/rev0000161

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18(6),* 643-662. https://doi.org/10.1037/h0054651

Virzi, R. A. & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13(4),* 304-319. https://doi.org/10.3758/BF03202499

Zimmer, H. D., Speiser, H. R., & Seidler, B. (2003). Spatio-temporal working-memory and short-term object-location tasks use different memory mechanisms. *Acta Psychologica*, *114*(1), 41-65. https://doi.org/10.1016/S0001-6918(03)00049-0

2. The Compass Task: A New Direction for the Verbal-Spatial Stroop Task

Introduction

Understanding how information to which we are not attending influences our behavior and cognition is an important goal. There are many real-life scenarios in which it is essential that we understand and account for how irrelevant information affects our behavior and decisions; think of a high-powered executive making important decisions under the pressure of many incoming streams of data, or even just our everyday lives trying to avoid suspicious links while answering e-mails or browsing the internet. From a theoretical standpoint, understanding how information which we are trying to ignore still permeates our cognition may influence the structure of models of attention and working memory.

The classic Stroop paradigm (Stroop, 1935) provides a vivid example of irrelevant information influencing behavior. The task requires participants to declare aloud the color in which color words are written. To accomplish this, participants must inhibit the pre-potent response to read the word aloud, which results in compromised reaction times when the meaning of the word and the text color are incongruent. Meanwhile, responses are quick and anecdotally easy when the word and color information are congruent. A Stroop task measures “interference” to determine whether inhibition is occurring: this is done by comparing the length of time taken for participants to respond to a congruent stimulus to the length of time taken to respond to an incongruent stimulus. If the response time to incongruent stimuli is greater on average than the response time to congruent stimuli, interference is said to be occurring, which reflects the need for inhibition to be enacted.

Stroop experiments have been conducted in many guises, using stimuli which go beyond colors and color words. In a spatial variant of the Stroop task, the stimuli consist of two parts: first is the words which have a spatial *meaning* (commonly used examples are “up”, “down”, “left”, “right” and the cardinal directions: “North”, “South”,

“East” and “West”) but are ultimately verbal stimuli; second is the words’ locations, which correspond to the words selected (i.e., an experiment using only “up” and “down” would have words placed in the top and bottom of the screen). Participants can be asked to respond based on the location or the word meaning, and in this way, every Stroop stimulus has both relevant and irrelevant information, depending on the task goal. An interesting feature of spatial Stroop tasks which is not commonly seen in the more classic color-word version is bi-directional interference. Stroop’s (1935) original paper on the color-word Stroop phenomenon concluded that extensive training is required in the skills of color naming and reading inhibition to elicit interference from color information on word reading, known as a “reverse Stroop effect”. This is not the case in spatial Stroop tasks, where interference from spatial information on verbal judgments is commonly present alongside interference caused by verbal information on spatial judgments. This puts spatial Stroop tasks in a unique position to probe some important interference-based questions.

An early spatial Stroop study by Shor (1970) investigated the effect of varying task focus in a spatial Stroop task. They found that their participants, who were asked to verbally state their response, experienced a considerably greater interference effect when asked to focus on which direction an arrow was pointing and ignore the meaning of the accompanying word, than when they focused on the word meanings and ignored the arrow directions. Palef and Olson (1975) also varied whether participants focused on the verbal or spatial content, but their results indicated the opposite pattern: instead of irrelevant verbal information interfering selectively with responses to relevant spatial information (as had been found in Shor’s study), they found that irrelevant spatial information selectively interfered with responses to relevant verbal information. The key difference between the studies appears to be the response method: in Shor’s experiment, participants responded vocally, and in Palef and Olsen’s study, they responded by pressing buttons.

There is an explanation for this asymmetry which has received some interest: the interaction of response type and task focus. Tentative support for this notion can be gleaned from an early study conducted by White (1969), wherein participants were asked only to respond to where a word appeared within a printed display, either by verbalizing its location or by moving a lever into one of four corresponding positions (this response manipulation was divided into blocks). White found that

there was less interference when the participants had to respond manually with the lever than when they had to verbalize the spatial information to speak it out loud. In White's paper, the data begins to document an interaction between congruence (in their case, comparing incongruent stimuli with verbally neutral stimuli consisting of nonsense syllables) and response method. Their data show that verbal responses to location judgments were affected to a marginally greater extent by incongruence than the non-verbal responses, though they reported that this did not reach significance (without giving exact values). White cautioned in this paper that it was possible that the results were skewed somewhat by participants' unfamiliarity with the non-verbal response apparatus, which may account somewhat for the non-significance of the effect. While this paper does not link response type to task focus explicitly, this demonstrated at least that the interference participants experience within one task varies with the response method they use, and that spoken and lever responses are essentially different somehow. The results from Shor (1970) and Palef and Olson (1975) taken with the above finding from White (1969) strongly indicate that an interaction could be in play.

Since both representation and response formats appear to impact interference strength, both must be considered when explaining spatial Stroop findings. Virzi and Egeth's Translational Model (1985) may be applied. According to this hypothesis, interference is greatest when the nature of the response type and the nature of the information attended are in contrast. For example, when asked to focus on the location of a word, but to respond through speech, the location information must be mapped to its spoken form for the response to be made, so interference occurs. However, when asked to focus on the word and respond through speech, the mapping from representation to response is direct, meaning that there is no or extremely little interference. This is due to a necessary step of 'translation', which manifests in behavior by delaying the correct response when it must be transformed from one information type to another uncomplementary response type in the presence of conflicting information which maps more fluidly to the target response type.

To recap previously discussed findings with respect to this hypothesis, Shor (1970) varied their study's task focus but always asked participants to respond vocally. They found significantly greater interference in their spatial focus task. This

fits with the Translational Model in that task focus (spatial) and response type (verbal) were in contrast. Palef and Olson (1975) also varied their task focus and kept response type fixed, but they utilized a manual response in the form of pressing one or the other response key, which may align better with the spatial processing system. Fitting again with the Translational Model, significant interference was experienced in the verbal focus task, wherein response method (spatial) and task focus (verbal) were contradictory, but not in the spatial focus task. White (1969) conversely held task focus constant and varied response method. In their study, participants focused only on the location of words and were asked to respond in some trials by speaking the location aloud and in other trials by moving a lever into one of four positions corresponding to the four locations at which the words could appear. Again, their results suggest that interference may have been somewhat greater when response type and task focus are in contrast, which is when participants had to respond by speaking aloud, supporting the Translational Model.

Although the evidence from these extant spatial Stroop studies all coincides with predictions of the Translational Model, they provide only tentative support for it because they do not vary both factors in one sample of people. Virzi and Egeth's Experiment 2 (1985) almost accomplished this. Response method was manipulated within-subjects, and task focus was manipulated between-subjects. Their stimuli were the words "LEFT" and "RIGHT" positioned to the left and right of the centralized fixation point. To manipulate response type, Virzi and Egeth asked participants on half of trials to respond spatially by pressing a left or right-hand button and on the other half of trials to respond verbally by speaking the word "left" or "right" aloud. To manipulate task focus, participants were asked either to respond to the location or the meaning of the word depending on whether they were assigned to one between-subjects condition or the other. Their results supported the predictions of the Translational Model: interference was significant in the two conditions where response type and task focus contradicted: verbal task focus with spatial response and spatial task focus with verbal response. No significant interference was detected when response method and task focus matched.

It is sometimes useful to compare theories which aim to explain a single phenomenon of cognition with broader but related models of cognition, to establish where more specialized theories may or may not fit within certain frameworks. On

the broad cognitive topic of working memory, probably the best-known model is Baddeley and Hitch's Multicomponent model (Baddeley, 1986; Baddeley, Hitch & Allen, 2021). This model has undergone plenty of alterations and additions in those years, but its most important characteristic has stayed the same: this, of course, is the model's verbal and visual slave systems, which are tasked with maintaining only verbal and visual information respectively. Drawing this line firmly between (and characterizing independent systems within the mind for) these two modalities marks a very stark similarity to the Translational Model: both theories emphasize the separateness of visual elements like color or space from words. Though the Multicomponent model does not approach the issue of response output, we suggest it would not be such a stretch for the framework to posit that certain methods of response may be connected more inherently to these existing information maintenance modules. The maintenance method for verbal information, known as the *phonological loop*, is already described as an 'internal voice' repeating the to-be-maintained verbal material. Therefore, it stands to reason that it might be more direct to express with one's physical voice the information which is already being articulated by their internal voice than to involve a limb and have to orient it in a semantically corresponding space. Similarly to the Multicomponent model, the Time-Based Resource Sharing model (Barrouillet et al., 2004; Barrouillet & Camos, 2015) suggests that verbal information is maintained using a rehearsal mechanism which is unique from all other information types. Domain-general working memory models, without specific stores or rehearsal mechanisms for different information types, align less strongly with the Translational Model, but if the Translational Model is compellingly demonstrated, it might be integrated into any working memory model to improve its specificity.

In the modern (and post-pandemic) world, online studies are becoming increasingly convenient, both for participants and the researcher. This shift has been validated by important meta-scientific findings such as those from Uittenhove et al. (2023) suggesting that researchers can expect similar data quality when recruiting online populations compared with running experiments on undergraduates, even in the lab. However, in spite of the spike in popularity of online experimental software, it can be difficult to guarantee that asynchronous data collection goes smoothly. This is the case especially when an experimental method requires that participants have

access to specific equipment. While most people browsing the internet can be expected to have access to a computer or laptop with a trackpad or mouse and keyboard on which to participate in online experiments, it is not guaranteed that they will have access to a microphone, which poses problems for moving the spatial Stroop paradigm (with a spoken response manipulation) online. We asked ourselves the question: is it possible to replace the spoken response in this method with something more convenient, but still exceedingly familiar and implicit, as speech production must be? The average person spends a great deal more time typing today than they did when Virzi and Egeth ran their study in 1985 (whether it be composing formally on computers or texting on our smart phones), to the point where the majority of people in the Western world must be implicitly familiar with the locations of letter keys. Logan and Crump (2011) attest to this when they report that, when typing, “our fingers find the correct locations five to six times per second” (p. 13). Perhaps then, familiar letter keys could adequately stand in for a spoken verbal response in this paradigm.

Virzi and Egeth largely refrain from commenting on what constraints may act on the connections between processing nodes and response types in their theory. If we were to find that a variant of a manual response could elicit the same pattern of responding as the previously used verbal response, this would speak to several possible theory constraints or targets for elaboration. First, it suggests that one processing system can feed into multiple different kinds of response (i.e., verbal processing can output without the need for translation into both a spoken response and a keypress response node). This would imply that the previously linear and parallel systems from sensory input to response output demonstrated in Virzi and Egeth’s 1985 paper may actually be better represented by more complex, flexible, web-like structures with multiple response offshoots. Second, this finding would confirm that the same physical kind of response (here, keypresses) can be fed into by different information processing systems. Finally, this would suggest that it is not the physical nature of the response which solely determines the information type to which it aligns. If the physical action of key presses can complement verbal information as well as spatial information, we can ascertain that the semantic associations attached to the response are at least somewhat responsible for how it aligns with the various information processing systems.

Experiment 1

The first of our experiments aims to replicate the findings of the Virzi and Egeth experiment detailed above (see Figure 1 below for a demonstration of the expected pattern), with some alterations that aim to test whether the paradigm is robust to changes that will enable more convenient online hypothesis testing with it. The current study used four verbal items: the letters “N”, “E”, “S” and “W”, representing the four cardinal directions of North, East, South and West. These appeared randomly at four locations around a simple line art compass image in the center of the computer screen: immediately above, below, to the left and the right. In half of trials, participants were asked to respond to the location at which the letters appeared, and in the other half they were instructed to respond to the meaning of the letter with regards to the direction it represents. In half of the trials for each task focus, participants responded using the arrow keys on their computer keyboard, and in the other half using the N, E, S and W keys. Trials within these blocks could be incongruent or congruent, meaning that the irrelevant information type was either contradictory or complementary to the focus information type, respectively.

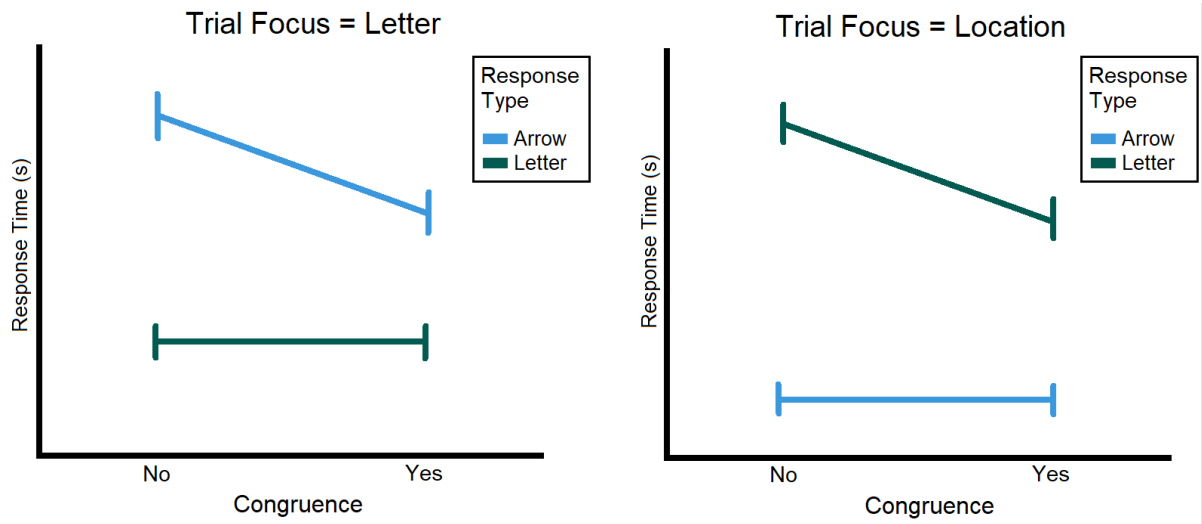

Figure 1 demonstrates the expected pattern of results in-line with the Translational Model.

In addition to nearly replicating Virzi and Egeth’s study, this study also aims to determine whether two manual responses can be used in place of a manual and vocal response (as were used in Virzi & Egeth, 1985) if one of the manual responses is explicitly associated with verbal information and the other associated with direction. If pressing letter keys on a keyboard can replicate findings from the more intuitively word-complementary response of speaking aloud, this will have

implications for the Translational Model as discussed above. This will also be promising for online experiments using this paradigm going forward.

The hypotheses of this experiment are as follows:

1. Overall, congruent trials will be responded to more quickly than incongruent trials.
2. Trials where the response type matches the focus (letter keys with letter focus and arrow keys with location focus) will be responded to more quickly than when these factors mismatch.
3. The interference effect (measured by the difference in response time between the congruent and incongruent trials) will be more pronounced in trials where the response type and the focus do not match (e.g., arrow key response to verbal information).

Method

Participants

Participants were undergraduate students from Cardiff University's School of Psychology with normal or corrected-to-normal vision, recruited opportunistically through the University's Experimental Management System, and rewarded with partial credit towards their course requirement. Demographic information was not collected from the participants, which limits the conclusions which could be drawn from this data regarding generalizability. The student population from which the sample came were mostly female, mostly aged 18-23 years and all fluent (but not all native) speakers of English as required of undergraduate students on a course taught in the medium of English. Initially, 57 participants took part, but several were excluded on the grounds of exclusion criteria which are detailed in the Results section. This resulted in 34 participants being included in the analysis.

Materials and Apparatus

This experiment was designed in PsychoPy (Peirce et al., 2019) and delivered online on participants' personal computers within their internet browsers, delivered through Pavlovia (pavlovia.org). Participants were requested only to sign up to participate if they carried out the experiment on a US or UK layout QWERTY keyboard (as other layouts may not have the important keys in the same locations).

The stimuli for experimental trials consisted of the letters “N”, “E”, “S” and “W” (presented in uppercase), which appeared one at a time in one of four positions around a central image depicting a compass. These positions were immediately above, below, to the left and to the right of the compass image, to emulate the positions of the letters which represent the four cardinal directions on a real compass face (see Figure 2 for an illustration). The letters in all trials were presented in white against a medium grey background. The central compass image was also primarily white, with some greyscale details.

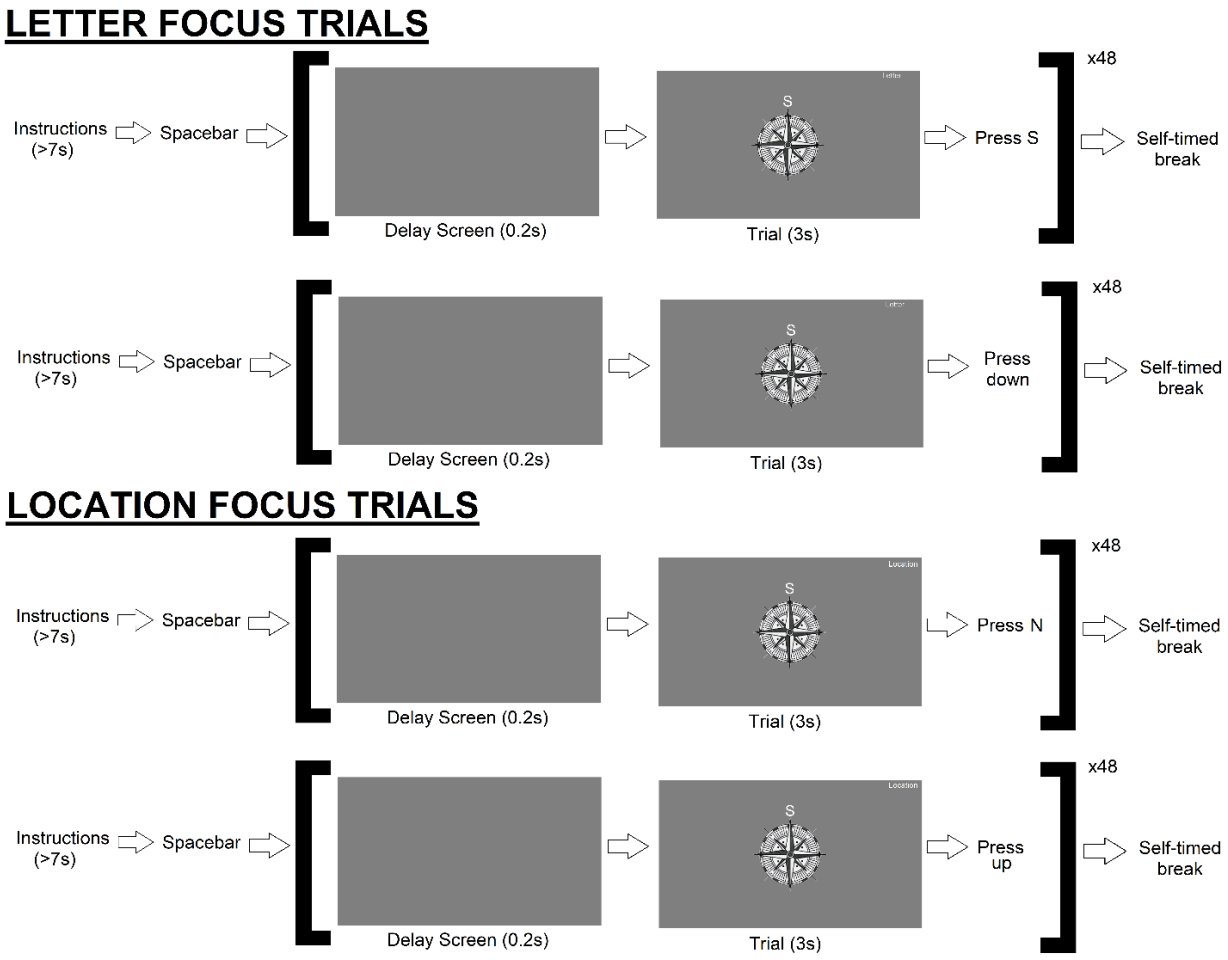

Figure 2 shows the structure of all four experimental blocks in Experiment 1.

To attempt to disambiguate between facilitatory and inhibitory processes, the former of which is not explained by the Translational Model, control trials were also implemented. However, later analysis revealed that this was not a successfully designed control task. For the sake of brevity, please find methodological and

analytical details of these trials in the *Supplementary Materials* folder of this project's OSF page (https://osf.io/hdrq5/).

Design

The study was within-subjects, with block order determined randomly in each case to prevent confounding order effects. Trial order was also randomized within blocks. The dependent variables of interest were participant accuracy and response times on correctly answered trials. We expect that response times will be a more sensitive measure than response accuracy, due to the tendency of undergraduate participants to perform incredibly accurately in basic cognitive tasks such as this. The independent variables were trial focus (letter or location information), response type (NESW or arrow keys) and congruency (irrelevant information type was congruent or incongruent).

Procedure

Participants were given information about what the study would entail and asked for informed consent, which was taken by their continuing to take part in the experiment. Participants were then asked to read two pages of general instructions, informing them about the images that would be used and that they would first be undergoing some practice trials. The consent and instructions screens were all programmed to not advance on keypresses for a set time (lengthier text screens like the pre-block instruction screens detailed in Figure 2 lasted for at least 7s, and brief text screens for 3.5s), to encourage participants to fully read the information. Next, participants were given written instructions in parts, explaining how each of the four experimental blocks would need to be carried out, followed by four practice trials for each block type (totaling 16 practice trials) with accuracy feedback after each trial. Participants were also made aware at this point of an occasional secondary task which required them to respond by pressing the spacebar when an asterisk (" * ") appeared on a trial instead of a letter. The aim of these asterisk trials was to check whether participants were paying proper attention rather than pressing keys repetitively or randomly. Following this, the instructions declared that the real trials would begin. Six blocks of trials then took place, four of which were experimental and the remaining two were control blocks, which are discussed in the relevant document in the project's OSF page (https://osf.io/hdrq5/). Every block consisted of 50 trials, 2 of

which were the aforementioned asterisk trials. The remaining trials in the experimental blocks were split evenly between congruent and incongruent trial types. Within each block, each letter or location appeared an equal number of times and trials ended automatically if a response was not provided within 3s of stimulus onset. After each block, participants were prompted to take a short break, the length of which they could determine themselves by delaying progressing the screen until they felt ready to proceed. Upon the completion of all six blocks, participants were asked to divulge honestly whether they were familiar with the locations of the relevant letters on a compass before they began the experiment. It was stressed that their honest answer was really valued for the sake of the data's validity and would not affect their anticipated reward for completion. Participants were then thanked and debriefed through on-screen messages. This study was approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Exclusion Criteria

Participants were excluded if they: attained less than 85% accuracy in all trials (N = 14); did not respond to at least half of the trials correctly in each cell (N = 8 additional participants); did not respond correctly to at least one 'catch-out trial' (trials which had an asterisk in place of a letter required the participant to respond by pressing the spacebar; N = 1 additional participants); responded "no" to the question asking whether they were familiar with the directions on a compass before they took part in the experiment (N = 0 additional participants). These exclusions resulted in 34 participants' data being used in analysis. Participant exclusion was done before the individual datasheets were compiled.

The response time data were further filtered through single trial exclusions, log transformations were performed on raw response times to combat skew, and averaging over multiple trials of the same type to convert the data into a one-row-per-subject format. These processes were carried out in RStudio (R Core Team, 2021). Within RStudio, the package 'Tidyverse' (Wickham et al., 2019) was used for several steps of these processes. For single trial exclusions in the accuracy analyses, catch-out trials were removed, and participants' accuracy scores were

arcsine square root transformed to combat skew. For single trial exclusions in the reaction time analyses, the criteria were as follows: catch-out trials and trials where the participant did not respond or responded incorrectly were removed from the compiled data sheet for analysis, so that only the reaction times of correctly answered trials were analyzed. Further, for both analysis types, trials with impossibly short reaction times were excluded, as these were reasoned to be mis-presses. Each trial has a delay of 0.2s before the stimulus appears (see Figure 2), so in line with common consideration in cognitive research that responses of less than 0.2s are too quick to accept as a genuine response (e.g., see Whelan, 2008), a filter was applied to remove responses of less than 0.4s (this resulted in the removal of only 0.1% of data). Very slow responses could not be given by participants due to the trial timing out 3s after stimulus onset.

Main Analyses

A 3-way repeated measures ANOVA was conducted on the accuracy data, which revealed a significant main effect of Congruence, with congruent trials responded to more accurately in general than incongruent trials ($F(1,33)=25.725$, $p<.001$, $\eta_p^2=.438$). The two-way interaction between Focus and Response Type was significant ($F(1,33)=11.740$, $p=.002$, $\eta_p^2=.262$). Finally, the key three-way interaction between all factors was significant ($F(1,33)=44.148$, $p<.001$, $\eta_p^2=.572$).

A 3-way repeated measures ANOVA was conducted on the response time data, which revealed significant main effects of all three factors: Focus ($F(1,33)=78.288$, $p<.001$, $\eta_p^2=.703$), Response Type ($F(1,33)=160.398$, $p<.001$, $\eta_p^2=.829$) and Congruence ($F(1,33)=120.935$, $p<.001$, $\eta_p^2=.786$). Two of the two-way interactions, Focus by Response Type ($F(1,33)=416.114$, $p<.001$, $\eta_p^2=.927$), and Focus by Congruence ($F(1,33)=7.406$, $p=.010$, $\eta_p^2=.183$), were also significant. Most importantly, the key three-way interaction (as illustrated in Figure 3 below) of Focus, Response Type and Congruence was significant ($F(1,33)=90.337$, $p<.001$, $\eta_p^2=.732$),

suggesting that interference occurred to different extents depending on the combination of both trial focus and method of response.

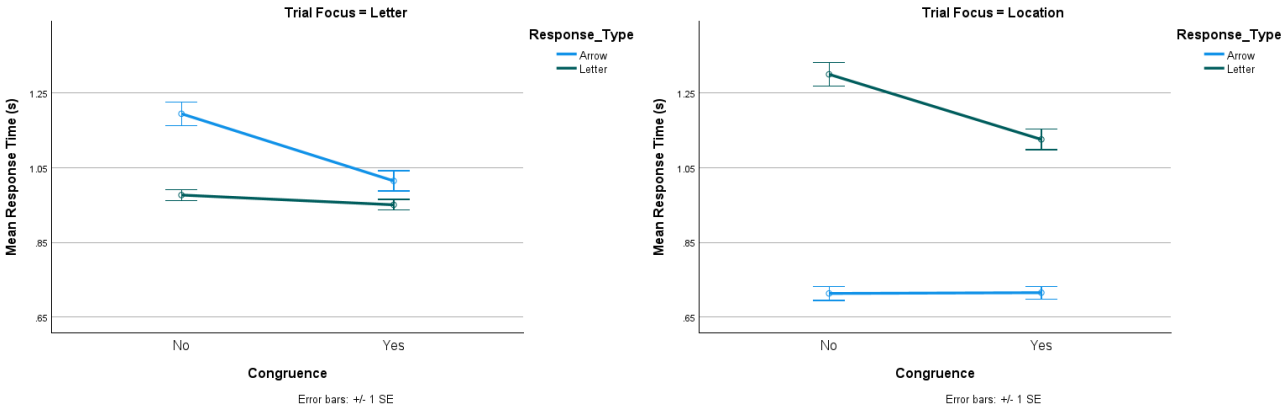

Figure 3 shows the effect of congruence as a function of response type on participant response times within the letter-focused trials (left) and location-focused trials (right).

Discussion

In summary, all three main hypotheses of this experiment are supported by the data here: congruent trials were responded to more quickly overall than incongruent trials, response times overall were faster when the response method and trial focus matched, and interference was greater in conditions where the trial focus and response method did not match. This third finding supports the stipulations of the Translational Model and provides a modernized near replication of Virzi and Egeth's Experiment 2 (1985), despite superficial differences in methods. The significant main effect of Focus suggests that generally, location information was responded to more quickly than letter information, which is consistent with previous findings (Shor, 1970; Palef & Olson, 1975; DeSoto et al., 2001).

The main effect of Response Type suggests that across all trials, participants respond more quickly when using the spatial response than the verbal response, which we expect reflects the highly intuitive layout of the arrow keys compared to the letter keys. Snyder et al. (2014) report that while typing, expert typists can accomplish an average rate of between six and seven keystrokes per second with near perfect accuracy, reflecting a brilliant implicit knowledge of keys' locations on a keyboard. However, these experts' explicit knowledge is surprisingly limited: their findings suggest that experts can only accurately report approximately half of the keys' locations when asked explicitly. Our method may have been tapping into our

participants' explicit knowledge which may have slowed their response times. However, the main interest in this experiment for drawing conclusions about the Translational Model was the difference in response times that trial congruence elicited and how this difference varied with the conjunction of response type and task focus, which does not necessitate the comparison of raw response times across response types. We expect that congruent trials would be hindered by this to the same extent as incongruent trials would be, hence it seems that the detection of an interaction was not hindered.

The first difference between the current study and Experiment 2 of Virzi and Egeth (1985) is the considerable update of the experimental technology. The current study was conducted online, a product of its time following the COVID-19 pandemic. Due to the commonly held conception of higher levels of noise in data collected online, methods that can be conducted reliably online may prove good candidates for future research, especially with many researchers turning to online research after seeing its time and cost effectiveness. It is also advantageous for recruiting more demographically diverse samples from outside university populations. Relatedly, the use of two manual responses in this experiment, with varying degrees of spatial and verbal associations, makes it a much more convenient methodological choice whether the study is conducted online or in-person than the alternative: collecting and scoring spoken responses. If the current methodology can be shown to reliably work online and in-person, perhaps this may encourage further study in this particular research area.

The 'catch-out' trials detailed in the methods section were included in the study as a measure of participant attention to allow post-hoc participant exclusion. In hindsight, these trials are made redundant with the application of the 85% accuracy inclusion criterion. It is also possible, but we think unlikely, that this secondary task could have affected the results because enacting dual tasks can impact on cognitive resources. On the basis that this measure did not exclude any participants after the performance quota was applied, we will not include these trials again in the future for exclusion purposes. Another minor difference between this replication and the original study is the stimuli used. Virzi and Egeth's (1985) study used only two words as their verbal stimuli "LEFT" and "RIGHT", whereas the current study used four single letters to represent the cardinal directions: "N," "E", "S", and "W". A study

which utilized four possible location and verbal stimuli each found that the difference between response times on congruent and incongruent trials was greater (Shor, 1970) than when only a two-item distinction was made. Further, Virzi and Egeth provided trial-by-trial feedback whereas the current experiment did not, in the interest of timesaving. Instead, the current study implemented a short training session at the beginning of the experiment with feedback to teach participants how to respond to each condition. The length of trials in the current experiment is also slightly shorter than those in the original experiment. These are very small changes which did not result in any tangible differences in outcomes but should nonetheless be noted.

The more noteworthy difference between our study and Virzi and Egeth's (1985) is of course that we used key presses for "verbal" responses, relying on participants' prior familiarity with the locations of letter keys. This successful replication, even with our manual-verbal responses also provides important new detail for the theory. This experiment has provided evidence to suggest that more than one response can be aligned with one information processing system, and that the same physical type of response can have variants which align with different information processing systems.

Experiment 2

Following the successful online replication of the key findings of Virzi and Egeth (1985) and the promising evidence that the Translation Hypothesis's 'verbal response' node might be more flexible than initially suggested, we decided to run another version in the lab in the hopes of replicating the findings in-person and further attesting to the robustness of the effect using our adapted manual-verbal response (Experiment 2's preregistration can be found here https://osf.io/xzg7u). Following data collection in Experiment 1, we noticed that on a QWERTY keyboard, the S, W, and E keys are situated relative to one another just like the arrow keys referring to those directions, with S lying below, W to the left, and E to the right. We reasoned that it is possible that this correspondence could influence how participants think of the letter response task, for example, by providing a spatial shortcut. We think that this is unlikely to have altered the results in Experiment 1, because if participants had utilized this spatial layout of the letter keys and therefore minimized differences between the letter and arrow key response types, we would not have

observed the strong interaction between congruence and response type. If this had been the case, we would expect to see the same pattern replicated from one response type to the other, instead of the response types mirroring one another. Nonetheless, we decided that it was important to replicate the study in the lab so we could fully control this factor. This has been directly addressed in Experiment 2 by having participants respond on individual NUM pads to emulate response boxes, one for letter keys and one for arrow keys. On the letter keys NUM pad, the keys were labelled with small stickers and arranged in a horizontal line reading left to right "N-E-S-W". Since neither the control nor the catch-out trials used in Experiment 1 added value to interpreting the data, these have been removed entirely from Experiment 2.

Finally, we were interested in whether the presence of the compass image was affecting the participants' responses somehow. We reasoned that it was possible that the compass image, which is new to this paradigm compared to previous spatial Stroop tasks that inspired this work, might be necessary for producing the observed effect, or otherwise might be boosting the effect sizes. Therefore, in Experiment 2 we manipulated the presence of the compass image to assess this.

Method

Participants

The program G*Power (Faul et al., 2007) was used to run a power analysis to calculate sufficient sample sizes for each of the groups in Experiment 2. It is recommended to use the effect size of the smallest effect one cares to detect when calculating sample sizes: in this case, this is the three-way interaction which had an effect size of η_p^2=.698. When inserted into the a priori ANOVA setting on the power calculator with an alpha error probability of 0.05, it indicates that at least 29 participants should be tested to detect this effect. This sample size estimation was extrapolated to apply to the Compass Absent condition for lack of better data on which to base such an assumption or power calculation. Participants were recruited in the same way as in Experiment 1 with the additional criterion that they had not participated previously and assigned to the between-subjects condition of Compass Present/Compass Absent randomly. Again, demographic information was not collected from these participants, which limits the conclusions which could be drawn

from this data regarding generalizability. As in the previous experiment, the student population from which the sample came were mostly female, mostly aged 18-23 years and all fluent (but not all native) speakers of English as required of undergraduate students on a course taught in the medium of English. Initially, 77 participants took part, but 18 were excluded on the grounds of several exclusion criteria which are detailed in the Results section. This resulted in 59 participants being included in the analysis.

Materials and Apparatus

This experiment was delivered in-person on a computer (using an iiyama ProLite XUB2294HSU, 21.5-inch monitor with a maximum resolution of 1920x1080 pixels), via PsychoPy (Peirce et al., 2019). Participants were asked to respond on the two detachable NUM pad keyboards which can be seen in Figure 4, one which was spatially oriented to mirror the compass display and one wherein the keys were intended to be spatially neutral, marked with the letters “N”, “E”, “S” and “W” (all in a horizontal line). The stimuli for experimental trials were largely the same as in Experiment 1, with the only change that the Compass Absent group did not see the compass image, nor any central fixation dot or cross, in any of the trials they experienced.

Figure 4 shows the custom-labelled NUM pads on which participants responded to trials in Experiment 2.

Design

The study had one additional between-subjects variable compared to Experiment 1, which was whether or not the compass image was present throughout the experiment (Compass Present vs. Compass Absent). The dependent variables of interest were again, participant accuracy and response time in seconds on correctly answered trials.

Procedure

The following details highlight changes in the procedure compared to Experiment 1. When participants arrived in the lab, they were verbally given information about what the study would entail and asked for informed consent, which was taken by their continuing to take part in the experiment. The consent screen was programmed to not advance on keypresses for a set time of 7s, to encourage participants to fully read the information. Before the experiment began, participants were given 120 simple trials wherein a letter (N, E, S or W) appeared in white text on-screen, presented centrally, and were tasked with pressing the correspondingly labelled letter key on the letter NUM pad (shown on the left of Figure 4). The aim of this short training task was to familiarize the participants with the locations of the letter keys before the experiment began, so that data from early trials in this response type were not confounded by confusion regarding which button is which. When participants completed this training, they were offered the chance to run through the training again if they felt they were still not familiar with the keys' locations. This training data was not used in analysis.

The remaining parts of Experiment 2 were identical to those in Experiment 1, except the following changes: the removal of all control and asterisk trials, an enforced 1-minute break between blocks instead of an optional one, and participants were prompted after completing the practice trials to indicate this to the experimenter in the room, who then asked them to clarify whether they understood the task from the practice trials before moving on to the real trials. At the end, participants were thanked and debriefed verbally in addition to the on-screen text used previously and given the opportunity to ask questions about the experiment. This study was approved by the Cardiff University School of Psychology Research Ethics

Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Exclusion Criteria

Participants were excluded if they met the same criteria as in Experiment 1 (except the criterion concerning the catch-out trials, which were removed from this experiment): less than 85% accuracy in all trials (N = 13); did not respond correctly to at least half of the trials correctly in each cell (N = 1); responded “no” to a question asking whether they were familiar with the directions on a compass before they took part in the experiment (N = 4). Fifty-nine participants’ data were thus included in the analysis. Participant exclusion was again done manually before the individual datasheets were compiled. The data were transformed in the same way here as in Experiment 1.

Main Analysis

Compass Present Data

A 3-way repeated measures ANOVA was conducted on the accuracy data from the participants in the Compass Present condition. In this data set, the main effect of Congruence was significant ($F(1,29)=8.533$, $p=.007$, $\eta_p^2=.227$). The two-way interaction between Focus and Response Type was significant ($F(1,29)=10.973$, $p=.002$, $\eta_p^2=.275$). Finally, the key three-way interaction between all factors was significant, ($F(1,29)=4.360$, $p=.046$, $\eta_p^2=.131$).

A 3-way repeated measures ANOVA was conducted on the response time data from the participants in the Compass Present condition. In this data set, the main effects of Focus ($F(1,29)=79.780$, $p<.001$, $\eta_p^2=.733$), Response Type ($F(1,29)=142.094$, $p<.001$, $\eta_p^2=.831$) and Congruence ($F(1,29)=44.031$, $p<.001$ $\eta_p^2=.603$) were all significant. The Focus by Response Type interaction ($F(1,29)=398.570$, $p<.001$, $\eta_p^2=.932$) was again significant, as was the 3-way interaction ($F(1,29)=37.620$, $p<.001$, $\eta_p^2=.565$) between all three factors (see Figure 5), which is an indicator that the data conforms to the pattern set out by the Translational Model.

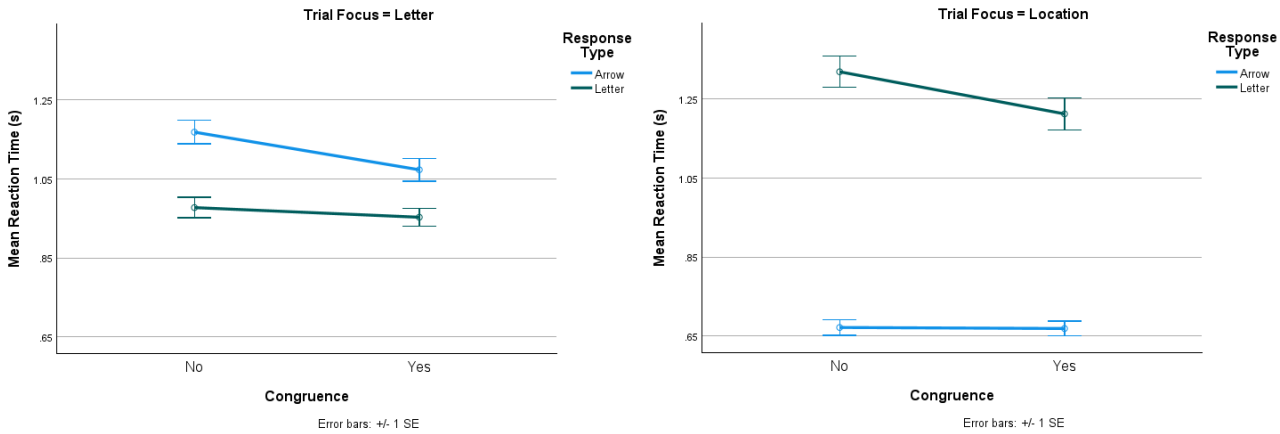

Figure 5 shows the effect of congruence as a function of response type on participant response times within the letter-focused (left) and location-focused (right) trials of the Compass Present group's data.

Compass Absent Data

The results were largely the same in the 3-way repeated measure ANOVAs run on the Compass Absent condition data set. In the accuracy data, both the Response Type ($F(1,28)=6.820$, $p=.014$, $\eta_p^2=.196$) and Congruence ($F(1,28)=18.739$, $p<.001$, $\eta_p^2=.401$) main effects were significant. Again the interaction between Focus and Response Type was significant ($F(1,28)=13.603$, $p<.001$, $\eta_p^2=.327$), as was the key three-way interaction between all factors ($F(1,28)=11.105$, $p=.002$, $\eta_p^2=.284$).

In the response time data, all three main effects were significant: Focus ($F(1,28)=30.870$, $p<.001$ $\eta_p^2=.524$), Response Type ($F(1,28)=117.727$, $p<.001$ $\eta_p^2=.808$) and Congruence ($F(1,28)=49.661$, $p<.001$ $\eta_p^2=.639$) were all significant. The Focus by Response Type interaction ($F(1,28)=311.502$, $p<.001$, $\eta_p^2=.918$) was significant and finally, the key 3-way interaction ($F(1,28)=12.485$, $p=.001$, $\eta_p^2=.308$) was also significant (see Figure 6). Again, this 3-way interaction being significant indicates that the data conforms to the Translational Model pattern that we expected.

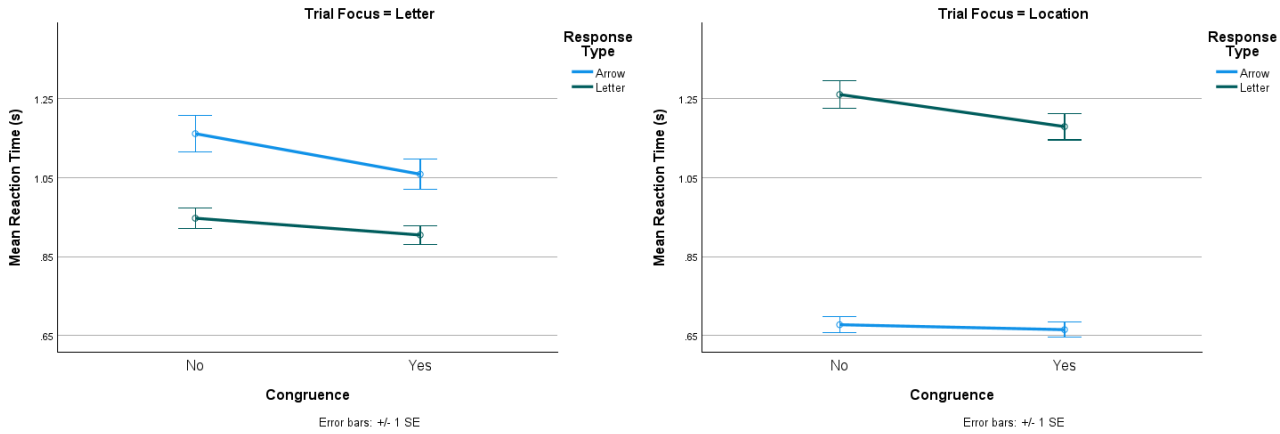

Figure 6 shows the effect of congruence as a function of response type on participant response times within the letter-focused (left) and location-focused (right) trials of the Compass Absent group's data.

Combined Data

We also ran a mixed 4-way ANOVA on both data sets combined to test the hypothesis that compass presence (the between-subjects variable) would have an influence on the extent that Stroop interference occurred. However, there was no evidence from the response time data to support either the prediction that the between-subjects condition interacted with congruence ($F(1,57)=0.560$, $p=0.457$, $\eta_p^2=.010$), nor the prediction that there may be a significant 4-way interaction between all four factors ($F(1,57)=.686$, $p=0.411$, $\eta_p^2=.012$). In fact, there was no evidence to support the existence of any difference in reaction times as a result of compass presence, with even the between-subjects main effect reported as not significant ($F(1,57)=1.080$, $p=.303$, $\eta_p^2=.019$). The equivalent comparisons in the accuracy data were also non-significant. These results suggest therefore that the presence or absence of the compass diagram had no significant impact on participants' experience of Stroop-like interference in this experiment, neither through response slowing nor accuracy reduction.

Follow-up Tests

Follow-up tests were conducted to determine whether Stroop-like interference was absent or just reduced in the letter focus-letter key and location focus-arrow key cells (trials requiring no translation). These analyses were conducted on the combined data set (both the Compass Present and Compass Absent data) for improved statistical power associated with a greater *N*, given that there appeared to

be no influence of the presence or absence of the compass on participants' data. These analyses took the form of 2-way and 1-way repeated measures ANOVAs.

Location-focus trials

In the accuracy data for only the location-focus trials, the main effects of Response Type ($F(1,58)=30.884$, $p<.001$, $\eta_p^2=.347$) and Congruence ($F(1,58)=13.838$, $p<.001$, $\eta_p^2=.193$) were both significant, as was the interaction between them ($F(1,58)=6.595$, $p=.013$, $\eta_p^2=.102$). In 1-way ANOVAs on each level of the response type variable, the results differed: Congruence was not a significant factor at the Arrow Key level of the response factor, ($F(1,58)=1.911$, $p=.172$, $\eta_p^2=.032$), whereas it was a significant factor at the Letter Key level of the response factor ($F(1,58)=13.410$, $p<.001$, $\eta_p^2=.188$).

In the response time data for only the location-focus trials, the main effects of Response Type ($F(1,58)=804.793$, $p<.001$, $\eta_p^2=.933$) and Congruence ($F(1,58)=38.589$, $p<.001$, $\eta_p^2=.400$) were both significant, as was the interaction between them ($F(1,58)=20.747$, $p<.001$, $\eta_p^2=.263$). In the results of 1-way ANOVAs conducted on this data subset, Congruence was not a significant factor at the Arrow Key level of Response Type ($F(1,58)=1.274$, $p=.264$, $\eta_p^2=.021$), but was at the Letter Key level ($F(1,58)=37.437$, $p<.001$, $\eta_p^2=.392$). These results indicate that Stroop-like interference does not occur in both Response Types when participants respond to location: there is only evidence to support the claim that when the focus is location information, Stroop-like interference occurs when participants are responding using Letter Keys, which is when translation must occur between information type and response type.

Letter-focus trials

In the accuracy data for only the letter-focus trials, the main effect of Congruence was significant ($F(1,58)=20.447$, $p<.001$, $\eta_p^2=.261$), as was the two-way interaction between Congruence and Response Type ($F(1,58)=6.276$, $p=.015$, $\eta_p^2=.098$). In 1-way ANOVAs, Congruence was a significant factor at the Arrow Key level of response type ($F(1,58)=30.294$, $p<.001$, $\eta_p^2=.343$), but not at the Letter Key level ($F(1,58)=3.015$, $p=.088$, $\eta_p^2=.049$), in-line with the results above on location-focus trials.

In the response time data for only the letter-focus trials, the main effects of Response Type ($F(1,58)=65.318$, $p<.001$, $\eta_p^2=.530$) and Congruence ($F(1,58)=72.369$, $p<.001$, $\eta_p^2=.555$) were both significant, as was the interaction between them ($F(1,58)=24.476$, $p<.001$, $\eta_p^2=.297$). In 1-way ANOVAs, Congruence was a significant factor at both levels of response type (letter key: [$F(1,58)=14.602$, $p<.001$, $\eta_p^2=.201$]; arrow key: [$F(1,58)=74.277$, $p<.001$, $\eta_p^2=.562$), but to a significantly greater extent when a translation was required (Arrow Key response level).

These results indicate that when the focus of a Spatial Stroop trial is verbal (a letter), Stroop-like interference in the form of response slowing occurs when participants respond with either letter or arrow keys, but to a larger extent with the arrow keys (when translation between information type and response type must be done). Contrastingly, Stroop-like interference manifesting as increased response errors to verbal items only occurs when a translation must occur, at the Arrow Key level of response.

Discussion

The aims of this experiment were two-fold: first, to replicate our online findings in the lab and with greater control over responses; second, to assess whether the presence or absence of a visual compass stimulus during trials had any quantitative influence on participants' experience of interference. With regards to the aim of replication, analyses revealed that in both conditions' data sets, all results were the same as Experiment 1. These findings all specifically align with the predictions made by the Translational Model that Stroop-like interference occurs to a greater extent when the trial Focus and trial Response Type are contradictory than when they are complementary. With regards to the second experimental aim, the prediction that Compass Presence or Absence would influence reaction times in some way was not at all supported.

To summarise the findings, the main effects of Congruence, Response Type and Focus were all significant, as were the 2 and 3-way interactions which we expected to occur in-line with the Translational Model. This provides replications in two new samples of the pattern of results associated with the Translational Model and supports the notion that within this paradigm, a manual response with verbal

connotations can stand-in for a spoken response to the same effect. Follow-up 1- and 2-way ANOVAs revealed that the effect of congruence was significantly greater at both levels of Focus when participants responded with the contradictory Response Type method. A hard interpretation of the Translational Model would predict no Stroop interference whatsoever when focus and response type are complementary. This was found to be the case in Location focus trials; however, there was not evidence to suggest that this was the case in our Letter focus trials, wherein a small amount of Stroop interference was still observed when participants used the complementary verbal response type. Therefore, the theory is only partially supported by these findings. This will be considered further in the General Discussion below.

General Discussion

Here, we report two experiments, one online and one in-lab, which both replicate the pattern of results predicted by the Translational Model: significantly greater effect of congruence when response type and task focus are in contrast compared to when they are complementary. This interaction is detected in both verbal and spatial versions of the task. The novel finding with consequence to theory that these experiments contribute is that a manual response with strong verbal associations, such as letter keys on a keyboard or response pad, can stand-in for the previously studied spoken word response to largely the same effect. Our results unambiguously support the Translational Model because they demonstrate that the need to translate causes Stroop interference in both Location and Letter focus trials.

Throughout the experiments reported here, effect sizes have been consistently larger, and significant effects have been more widely detected in the response time data than in the accuracy data, corresponding with our intuition that response time may be a more sensitive measure of interference. It is interesting that there is a discrepancy between the accuracy and response time findings in the follow-up analyses of Experiment 2. The accuracy results align perfectly with the predictions laid out in the translational model: interference is only detected when a translation must be made from stimulus type to response type. However, the analyses of the response time data indicate the existence of interference in the letter-focus trials where no translation is required and therefore according to the

translational model, no interference should occur, indicating that some other factor is responsible for that small amount of interference. It will be pertinent to assess whether this finding will be replicated in the further assessments undertaken in this area and to revisit the question of whether response times are a stronger measurement of Stroop interference.

We suggest that it is possible that this discrepancy between the trial foci is due to some innate difference in cognitive processing of verbal and spatial information. There is some reason to believe that these information types are not treated equivalently by the mind. Verbal-spatial binding asymmetries (Campo et al., 2010; Elsley & Parmentier, 2015) for instance, suggest that spatial information is necessarily recalled when attending to verbal information, but the same does not occur in these data for verbal information when spatial information is being attended to (though see Delooze et al., 2022, for a discussion on this). Further, the phenomenon of visuo-spatial bootstrapping (e.g., Darling & Havelka, 2010) shows that the familiar spatial layout of a numbered response pad in long-term memory can be advantageous for the accuracy of recall of digits compared to a novel spatial layout. However, to the best of our knowledge, no such phenomenon exists to document the reverse: evidence that verbal information in long-term memory might aid in the recall of spatial information in the short-term. Additionally, articulatory suppression has been shown to be very consistent in its selective “knocking out” of verbal working memory, whereas its counterpart, spatial tapping, is less reliable in selectively disabling spatial working memory (see Zimmer, Speiser & Seidler, 2003 for an example). For these reasons, we suspect that maintenance of verbal and spatial information is not equivalent in working memory. Possibly, the influence of multiple sources of interference in letter-focused spatial Stroop constitutes one more example of different treatment of verbal and spatial representations in working memory. Our findings suggest consideration of the Translational Model by working memory models, allowing that different kinds of representation will differ in the ease with which they might be translated into a response format.

Alternatively, this asymmetry could simply be attributed to our ‘stand-in’ verbal response being slightly imperfect. The data make a very strong case that the letter keys used here *work* as a verbal response, but perhaps pressing keys in response to reading letters or words will never be as natural as voicing them aloud. Support for

this suggestion can be gleaned from Virzi and Egeth's results: they found no interference whatsoever in their spoken responses to word information. Regardless of the comparative strength of the connection between the verbal system processing stage and our manual-verbal response, it is clear that the Translational Model requires updating to accommodate this new finding. This change could be as simple as adding secondary response stage nodes which are connected to their decision stage nodes with dashed lines representing a marginally weaker connection, as demonstrated in Figure 7 below.

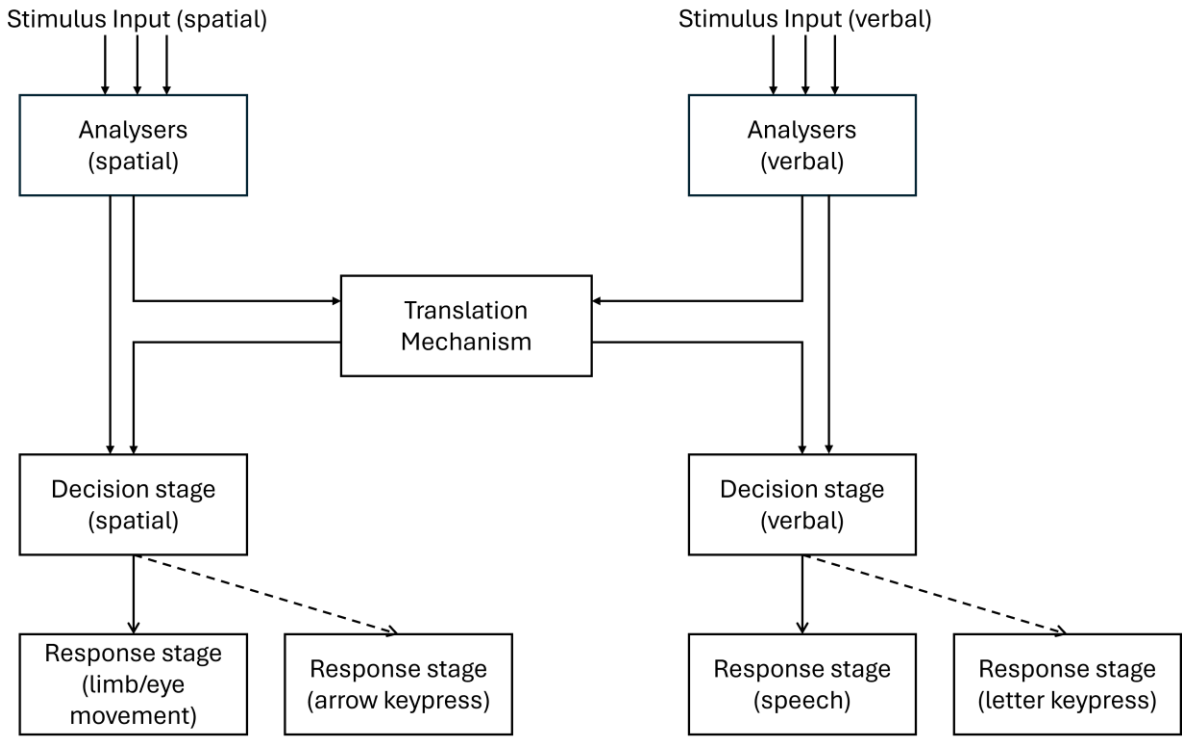

Figure 7 shows an adapted version of the translational model with added response stage nodes.

From the apparent success of using two manual responses in lieu of a spoken and manual response, as has most commonly been conducted previously in this area, one can start to ascertain something about the stimulus-to-response mapping which is only vaguely described in the Translational Model. It seems that it is not something inherent about the actions of speaking and button-pressing which align with verbal and spatial processing respectively, but more important is how the participant *thinks* about the response they are required to give. A button press can be verbal or spatial (and likely many other information types too) depending on the buttons' associated representations. As food for thought, these data pose the

question of which factors determine whether a connection exists between a response type and a given stimulus type. A straightforward suggestion for response to this is 'expertise' and it seems there is evidence to support it. Returning to the original Stroop paper (Stroop, 1935), the study demonstrates that with training, a reverse Stroop effect (incongruent ink color interfering with word reading) *can* be elicited after extensive training in out-loud color naming. This may be early evidence that expertise can connect a previously un-aligned response to a new information processing system.

To conclude, these experiments corroborated previous findings in support of the Translational Model in a Stroop task carried out on verbal-spatial stimuli. Our experiments additionally extend the literature by successfully using a new verbal response type, and we suggest alterations that might improve the model in light of this. The effect of congruence is very clearly bi-directional here, but the size of the effect of congruence, and its mediation by the interaction between focus and response type is not a perfect reflection in letter focus trials as it is in location focus trials. It is still unclear why verbal and spatial features of a stimulus are treated unequally in many examples across cognition, however, these studies suggest that the spatial Stroop task is a good candidate for future work trying to unravel this mystery, and by extension, argues for integration of the Translational Model into wider working memory theory.

Data Accessibility

The stimuli, program code, original (anonymised) data and scripts used to prepare the data for analysis are available within the project's associated OSF page: https://osf.io/xzg7u.

References

Baddeley, A. D. (1986). *Working memory*. Oxford University Press.

Baddeley, A. D., Hitch, G. J., & Allen, R. (2021). A multicomponent model of working memory. *Working memory: State of the science*, 10-43. https://doi.org/10.1093/oso/9780198842286.003.0002

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of experimental psychology: General, 133*(1), 83-100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., & Camos, V. (2015). *Working memory: Loss and reconstruction*. Psychology Press. https://doi.org/10.4324/9781315755854

Campo, P., Poch, C., Parmentier, F. B., Moratti, S., Elsley, J. V., Castellanos, N. P., ... & Maestú, F. (2010). Oscillatory activity in prefrontal and posterior regions during implicit letter-location binding. *Neuroimage*, *49*(3), 2807-2815. https://doi.org/10.1016/j.neuroimage.2009.10.024

Darling, S., & Havelka, J. (2010). Visuospatial bootstrapping: Evidence for binding of verbal and spatial information in working memory. *Quarterly Journal of Experimental Psychology*, *63*(2), 239-245. https://doi.org/10.1080/17470210903348605

Delooze, M. A., Langerock, N., Macy, R., Vergauwe, E., & Morey, C. C. (2022). Encode a letter and get its location for free? Assessing incidental binding of verbal and spatial features. *Brain Sciences*, *12*(6), 685. https://doi.org/10.3390/brainsci12060685

DeSoto, M. C., Fabiani, M., Geary, D. C., & Gratton, G. (2001). When in doubt, do it both ways: brain evidence of the simultaneous activation of conflicting motor responses in a spatial stroop task. *Journal of Cognitive Neuroscience*, *13*(4), 523-536. https://doi.org/10.1162/08989290152001934

Elsley, J. V., & Parmentier, F. B. R. (2015). Rapid Communication: The Asymmetry and Temporal Dynamics of Incidental Letter–Location Bindings in Working Memory. *Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://doi.org/10.1080/17470218.2014.982137

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp.

Logan, G. D.,& Crump, M. J. C. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. In B. H. Ross (Ed.*), The psychology of learning and motivation: Advances in research and theory* (Vol. 54, pp. 1–27). Burlington: Academic Press. https://doi.org/10.1016/B978-0-12-385527-5.00001-2

Palef, S. R., & Olson, D. R. (1975). Spatial and verbal rivalry in a Stroop-like task. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *29*(3), 201. https://doi.org/10.1037/h0082026

Pavlovia, https://pavlovia.org Open Science Tools, Nottingham, UK.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods, 51,* 195-203. https://doi.org/10.3758/s13428-018-01193-y

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shor, R. E. (1970). The processing of conceptual information on spatial directions from pictorial and linguistic symbols. *Acta Psychologica*, *32*, 346-365. https://doi.org/10.1016/0001-6918(70)90109-5

Snyder, K. M., Ashitaka, Y., Shimada, H., Ulrich, J. E., & Logan, G. D. (2014). What skilled typists don't know about the QWERTY keyboard. *Attention, Perception, & Psychophysics*, *76*, 162-171. https://doi.org/10.3758/s13414-013-0548-4

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. https://doi.org/10.1037/h0054651

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From Lab-Testing to Web-Testing in Cognitive Research: Who You Test is More Important than how You Test. *Journal of Cognition, 6*(1): 13. DOI: https://doi.org/10.5334/joc.259

Virzi, R. A. & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13(4),* 304-319. https://doi.org/10.3758/BF03202499

Whelan, R. (2008). Effective analysis of reaction time data. *The psychological record*, *58*, 475-482. https://doi.org/10.1007/BF03395630

Wickham, H. et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4(43),* 1686, URL https://doi.org/10.21105/joss.01686 .

White, B. W. (1969). Interference in identifying attributes and attribute names. *Perception & Psychophysics*, *6*(3), 166-168. https://doi.org/10.3758/BF03210086

Zimmer, H. D., Speiser, H. R., & Seidler, B. (2003). Spatio-temporal working-memory and short-term object-location tasks use different memory mechanisms. *Acta Psychologica*, *114*(1), 41-65. https://doi.org/10.1016/S0001-6918(03)00049-0

3. The Compass Task 2: A Working Memory-Stroop Hybrid

Introduction

The Stroop effect (Stroop, 1935) is a well-known phenomenon, characterised by slowed responses when indicating the printed colour of a word's text when the semantic meaning and printed colour differ compared to when they match. Variants of the task exist using other types of stimuli as well. For instance, in the verbal-spatial Stroop task the stimuli are words or icons indicating locations or directions which appear at different locations around the screen. Like in the colour Stroop task, the word or icon's meaning can match or mismatch with the location in which they appear, and this affects how quickly participants are able to respond. An interesting way in which spatial Stroop findings tend to differ from colour Stroop findings is the ease with which interference can usually be detected in both directions (e.g., DeSoto et al., 2001; Shor, 1970; Virzi & Egeth, 1985; Delooze & Morey, 2024, also Chapter 2 of this thesis). In colour-word Stroop tasks, if participants are tasked instead with reading the words as quickly as possible while ignoring the colour they are printed in, it has generally been documented that the incongruence of the coloured text interferes with the task of word reading to a smaller degree, if it is found to do so at all (e.g., Stroop, 1935; Gumenik & Glass, 1970; Glaser & Glaser, 1982; Chmiel, 1984).

Researchers have tried various methods to obtain a reverse Stroop effect in the colour Stroop paradigm. Stroop (1935) found that participants needed to undergo many days of training in the task of naming the colour of the text to elicit a reverse Stroop effect. Nealis (1974) presented participants with a to-be-read Stroop stimulus followed immediately by coloured Xs of the interfering colour ink and was able to elicit a reverse Stroop effect through retroactive interference. Glaser and Glaser (1982) manipulated both stimulus onset asynchronies (also referred to as pre-exposure times) and the proportion of trials participants encountered which were congruent to find a specific conjunction at which reverse Stroop interference could be observed. Gumenik and Glass (1970) and Dyer and Severance (1972) both elicited a reverse Stroop effect by reducing the readability of the colour words. In the

study by Gumenik and Glass, the reverse Stroop effect was only one sixth of the size of the forward Stroop effect before readability-reducing masks were applied. Dyer and Severence's study is limited in that it did not investigate both reverse and regular Stroop interference in the same instance, thus it is impossible to directly compare the size of the effects (MacLeod, 1991). However, their largest reported reverse Stroop effect is only an increase in response time of 0.067s per word, which equals some of the smallest per-word forward Stroop effects detected in colour-word stimuli around that time (for instance, 0.069s in Glaser & Glaser, 1982; 0.088s in Dyer, 1971), and is dwarfed by others (for example, 0.375s in Gumenik & Glass, 1970; 0.468s in Klein, 1964; 0.47s in Stroop, 1935). These varied methods and small effect sizes suggest that our understanding of the reverse Stroop effect and how it compares to the regular Stroop effect is still incomplete. Perhaps this difference in effect sizes indicates that there is a fundamental difference between the task of naming colours while ignoring word meaning (as in the regular Stroop effect) and the task of reading words aloud while ignoring colour information (as in the reverse Stroop effect).

The reverse Stroop effect seems to be relatively common in scanning task (Uleman & Reeves, 1971) and card sort task (e.g., Martin, 1981; Chmiel, 1984) versions of the Stroop task. A scanning Stroop task is a method wherein participants are presented with a large display of many colour-word Stroop stimuli, and their task is to find and mark all the stimuli which match a particular criterion, ignoring the other feature. For example, *find all instances of the word "BLUE" (regardless of its ink colour)* or *find all items printed in green ink (ignoring what the words say)*. Card sort Stroop tasks are those in which colour-word stimuli printed onto slips of card must be sorted into piles or bins by different criteria, either the colour communicated by the meaning of the word, or the colour of the ink in which the word is printed. Findings from these tasks are not perfectly comparable to the classic Stroop paradigm due to differences in the methods, and again, it is hard to even roughly compare the size of the reverse Stroop to the regular Stroop effect when both are not tested within the same sample and method. It is noteworthy that in the card sorting experiment by Chmiel (1984) which did assess both the standard and the reverse Stroop effect, that the extent of the reverse Stroop effect is only ever as large as their regular Stroop effect under the double influence of articulatory suppression and the requirement of a translation. Articulatory suppression is the process of repeatedly uttering nonsense

verbalisations with the goal of occupying the verbal working memory rehearsal mechanism, and a translation generally refers to generating a response which is in a different domain to that of the stimulus. To elaborate on this for card sort tasks specifically, the bins could be labelled using labels which are the same type as the feature that participants are searching for, such as labelling bins with colour words (e.g. a “GREEN” bin, and a “BLUE” bin, etc.) if participants are sorting by word identity. Alternatively, bins or piles could be marked with labels which require a translation of some kind, for example, if participants are searching for cards based on the word’s meaning but the bins were labelled with coloured squares. Unless both of these manipulations were in play, Chmiel (1984) saw considerably smaller reverse than regular Stroop effect sizes.

Meanwhile, spatial Stroop tasks appear not to require any additional training, stimulus obscuring, articulatory suppression or very precise cueing to elicit strong bi-directional interference; it suffices to simply to elicit a translation by varying what kind of response participants are required to make. Virzi and Egeth (1985) and Delooze and Morey (2024, also Chapter 2 of this thesis) ran Stroop tasks using verbal-spatial stimuli, asking participants to judge either the location of the stimuli or their semantic meaning. These experiments also varied the response method with which participants must respond to the target information, with one response type aligning more closely with the verbal information (speech or letter key presses) and the other aligning more closely with the spatial information (directional key presses). The interference that participants experienced was strongest when (or only present when, as in Virzi & Egeth, 1985) the nature of the response which they gave did not match the nature of the stimulus they judged. This supports Virzi and Egeth’s (1985) translational model of Stroop interference, which posits that the delay and detriment to accuracy which is characteristic of the Stroop effect is caused by the need to translate incoming information from one domain into a response type that better complements the to-be-ignored information’s domain. For example: if a participant is judging a location (which is spatial) while ignoring a written word (which is verbal) by outputting a vocal response, they are required to translate spatial information into a verbal format where it then contradicts the written word information in the same domain.

It logically follows that the spatial Stroop paradigm in particular is a straightforward method of observing this effect due to the fact that both the verbal and the spatial elements of the stimuli have complementary response types. Written verbal items such as letters or words have an obvious output method in speech and movements of the limbs in space for a keypress or joystick response seem intuitively to map well to spatial representations. Meanwhile, it is harder to envision a response that humans can make which maps so smoothly and intrinsically to colour representations, which may be the cause of the difficulty in eliciting a strong reverse colour-word Stroop effect which has been observed by researchers in this field in the past.

As mentioned above, one way in which the Stroop task has been modified is through the introduction of stimulus onset asynchronies, essentially dividing up the elements which make up the stimuli and presenting them separately in time. Experimenters have used this method to test whether the Stroop effect may be a result of differential speeds of processing for verbal compared to colour information, a competing theory of Stroop-like interference. The basic outline of the hypothesis is that word information is processed more quickly than colour information, and therefore it is available for response earlier, causing interference. In line with this idea, Dyer (1971) hypothesised that at a particular pre-exposure of word information, interference on the colour naming task would reach its maximum peak, and that at sufficiently long pre-exposures, the information would be processed separately in time and therefore interference would dissipate. Their results suggested a peak of interference with pre-exposure times of between 40-60ms, but even at their longest pre-exposure times of 500ms, interference persisted. In a further study, Dyer tested much longer pre-exposure times, finding that interference reached its minimum at 2,000ms but then, intriguingly, increased again at pre-exposure times beyond that (Dyer, 1974). MacLeod (1991, p. 18) argued that “such a pattern is not easily reconciled with any simple relative speed-of-processing interpretation” and that the lack of convincing evidence that manipulation of stimulus onset asynchronies can elicit a reverse Stroop effect is further ammunition against the viability of the expectation that text colour can provoke robust interference on word reading. A follow-up study using spatial rather than colour stimuli revealed a slightly different pattern of results, indicating that the relationship between stimulus onset asynchrony

and extent of Stroop-like interference may not be identical across different stimulus types. With verbal-spatial stimuli, there is still a decrease in interference as pre-exposure time becomes longer, as was seen with colour-word stimuli. However, differently than data collected using colour-word stimuli, the data from these verbal-spatial stimuli do not show a peak of interference at particularly short pre-exposure times (Dyer, 1972). The important takeaway from these findings is that interference endures a surprisingly long time but appears not to do so in a linear fashion, and that this pattern may differ depending on the stimuli which are used.

An interesting paper by Kiyonaga and Egner (2014) reported that in their modified version of the Stroop test, holding colour words in working memory was a sufficient substitute for witnessing them in real time. The study involved presenting participants with a colour word in black font which the participants would need to remember, then after a 2,000ms delay, showing a coloured rectangle which the participants must respond to immediately with one of four computer keys based on its colour. Finally, participants in the study were shown another colour word in black ink (either the same as before or different) and asked to indicate if the word was the same or different as the word they were maintaining. They found that participants were significantly slower to correctly respond to coloured rectangles when they were holding an incongruent colour word in mind than when the colour word was congruent. In this experiment, the verbal information temporarily residing in working memory seemed to be acting comparably to when the word was presented simultaneously, as is the case in classic Stroop tasks. This is a strong demonstration of the mechanistic difference between instructing participants to maintain the information compared to allowing them to inhibit the information. Dyer (1974) showed that when participants were able to inhibit the irrelevant word information, interference was at its very lowest at a 2,000ms stimulus onset asynchrony. However, in Kiyonaga and Egner's unique method requiring participants to maintain the information, they implemented delays of 2,000ms between presentation of the to-be-remembered and the to-be-judged items, and the interference in the reaction times that they witnessed was just as strong as when the items were presented simultaneously. Another interesting difference in Kiyonaga and Egner's method compared to those pre-cueing studies discussed earlier is that Kiyonaga and Egner additionally found an interfering effect of incongruent colour patch stimuli in their

memory measures for the identity of the words. This direction of effect is reminiscent of the reverse Stroop effect. This suggests that not only do the contents of working memory affect perceptual cognition by making immediate judgment responses slower, but also that this retention-free cognition (participants responded immediately and therefore did not need to maintain the colour patch) affects the contents of memory in that it can hamper our ability to respond to recognition prompts quickly and correctly, even when just a single word is to be remembered.

Beyond these principal findings, Kiyonaga and Egner (2014) went on to test whether this temporally separate Stroop effect is subject to some of the same factors that influence the simultaneous Stroop effect. They tested proportion congruence effects, wherein higher proportions of congruent trials cause participants to display exaggerated Stroop effects on the incongruent trials. They also assessed the effects of stimulus-response congruence, wherein additional interference can be measured in response times when both the stimuli presented and the necessary response to be enacted have their own individual incongruence. Response incongruence here is due to multiple targets requiring the same response, for instance, if responses to blue and yellow colour patches required the pressing of the same key. Their results indicated that the working memory-Stroop effect which they describe is indeed affected by these manipulations. Given these successes, Kiyonaga and Egner concluded that holding interfering information in working memory is essentially the same as observing it, as far as eliciting Stroop-like interference is concerned.

The translational model (Virzi & Egeth, 1985) suggests that we automatically process information that is seen in the world around us, and that plans are made by a dedicated, domain-relevant information processing system to output the information as a response. When translation occurs, it can be thought of as the process of moving information from one system (e.g., the spatial system, if the item is a location) into another (e.g., the verbal system, if the output must be vocal). As discussed earlier, for written words, the most relevant response plan might be to articulate them aloud. For observed colour patches, it seems intuitively that no such convenient complementary response method exists. We know that the classic colour-word Stroop tasks are limited for the reason explored earlier: without the addition of various manipulations, they usually only elicit Stroop-like interference in one direction: participants struggle to vocally express a colour (which requires a

translation), but when participants are asked to vocally express a written word (which does not require a translation), effect sizes are much smaller and the effect is less consistently detected, even with the many creative methods which have been used. Perhaps this is due to this suggested physical inability to make a truly complementary “colour response”. If humans were capable of a response similar to the capabilities of camouflaging animals projecting colours onto their skin, perhaps this would allow experimenters to more consistently elicit a strong reverse Stroop effect. In the absence of such a step in human evolution, we are limited to using inefficient colour-aligned stand-in responses, such as pressing colour-marked keys or clicking on a colour-wheel (both of which have spatial connotations involved) or vocalising the word which represents the colour aloud. In this way, colour Stroop stimuli seem to be quite unique, and therefore we suggest that it would be pertinent to utilise this working memory Stroop paradigm developed by Kiyonaga and Egner (2014) to test the translational model on verbal-spatial Stroop stimuli, which are preferable in that they allow for clean manipulation of conflict of response formats. This method secondarily has the benefit of allowing us to determine whether the major findings from Kiyonaga and Egner (2014) endure over other domains. As we have shown in previous experiments, spatial Stroop tasks can easily elicit both the usual Stroop effect and the reverse Stroop effect (Delooze & Morey, 2024, also Chapter 2 of this thesis), using inherently complementary response methods which humans are capable of enacting. This makes the spatial Stroop task a good candidate for further testing in this area.

Following two successful experiments using a compass version of the verbal-spatial Stroop task, wherein single letters (N, E, S and W, representing the cardinal directions) appeared in four locations (above, below, left and right) around a compass image (Delooze & Morey, 2024, also Chapter 2 of this thesis), we turned our attention to the interplay between working memory and Stroop interference. The current study aims to utilise the working memory Stroop hybrid task from Kiyonaga and Egner (2014) to test the following hypotheses: first, that Stroop-like interference will be observed using verbal-spatial stimuli (extending their major finding into a new domain); second, that the interference observed will be bi-directional (occur in both the letter judging and the location-judging task); and third, in-line with the translational model (Virzi & Egeth, 1985), interference will occur only when a

translation is required. It is reasonable to suggest that all three of these hypotheses should be supported if, as concluded by Kiyonaga and Egner (2014), information held in working memory has the same power of interference in the working memory version of the Stroop task as simultaneously presented interfering information does in the classic Stroop task. If these hypotheses are not supported, the data reported could be taken as evidence against the claim that information held in working memory elicits Stroop interference in the same way as information which is simultaneously presented, given that all three of these hypotheses have been shown to be supported in a simultaneous Stroop version of this task (Delooze & Morey, 2024, also Chapter 2 of this thesis).

Experiment 1

Method

Participants

Seventy-one participants from Cardiff University's Undergraduate Psychology program took part in the experiment in exchange for either course credit or monetary payment. Paid participants received £10 for their time (approximately 45 minutes). Participants were recruited opportunistically via the School of Psychology's Experiment Management System. Sixty-one of the participants took part online, and the remaining 10 participants were recruited to take part in the lab in-person. The pre-registered target sample size was 45 participants in total, with the in-lab proportion aimed to contribute approximately 20% of the sample (pre-registration is accessible at https://osf.io/dkyfh). No demographic information was collected due to there being no reason to suspect any confounding effects of any particular participant attribute, but this limits the conclusions which could be drawn from this data regarding generalizability. The student population from which participants were taken was mostly female, mostly aged 18-24 years, and all were fluent speakers of English as required of undergraduate students on a course taught in the medium of English.

Materials and Apparatus

The experiment was built using PsychoPy (Peirce et al., 2019) and hosted for online distribution on https://pavlovia.org/. All participants, regardless of taking part online or in-lab, completed the experiment in a computer browser served by pavlovia.org. All participants were required to use either a PC or laptop to complete the task, as the experiment was not tablet or smartphone compatible. Due to participants using their personal devices, it could not be ensured that they would have a keyboard with a NUM pad, so the program was coded to accept only responses using the number keys above the QWERTY letters.

The experimental stimuli can be divided into two groups: letters (“N”, “E”, “S” and “W”) which were always presented in the centre of the screen, and locations (above, below, left and right) around a centralised compass diagram, all of which were marked out using an asterisk. Within a trial, stimuli were always paired so that one part of the trial utilised letter stimuli and the other location stimuli. See the Procedure section below for more information about the structure of a trial.

When stimuli were targeted for the participant to commit them to memory, they appeared in red font (in the case of location stimuli, the asterisk was red, but the compass diagram remained black and white). In all other instances, they were presented in white font. This difference in colour was implemented to increase the ease with which participants would be able to distinguish between the different phases of a trial.

Design

The experiment was a 2(Congruence: Congruent/Incongruent) x 2(Response Type: Arrow keys/NESW keys) x 2(Focus: Letter information/Location information) within-subjects design. Blocks were presented to the participants in a random order uniquely generated by the experiment program. Trials were deemed to be congruent if the to-be-remembered stimulus and the to-be-judged stimulus were semantically the same (e.g., the letter “N” and the location above the compass, which both symbolise the concept of ‘North’) and incongruent if they were not (e.g., the letter “N” and the location left of the compass, which symbolise the concepts of ‘North’ and ‘West’ respectively). All mentions of ‘congruence’ in this experiment relate to this relationship, *not* the relationship between the to-be-remembered word and the

recognition response probe (essentially whether the correct response to the memory question is yes or no), which is not a variable of interest. Response Type was manipulated between blocks, so that participants were instructed whether they were to respond to stimuli using the arrow keys or the NESW keys at the start of each new block. The same applies to Focus: trials were blocked so that the type of stimulus which participants had to make judgments about was always the same in each block. It is important to remember that the type of stimulus which participants made immediate judgments about was always opposite to the stimulus type which they were remembering. So, when judgment Focus was letter, memory Focus was necessarily location and vice versa.

Procedure

In-lab participants were not given any verbal task instructions by the experimenter, to equate their experience with the online participants. They were told that they were permitted to ask the experimenter questions if they felt confused, but none of them chose to do so. Otherwise, both the online and in-lab participants had the same structure of experience.

First, participants were shown a consent form screen which instructed them that by continuing to take part, they consented to the terms of the experiment. Next, participants were shown a series of instruction screens which explained the structure of the trial and the fact that the keys with which they must respond would change between blocks. Before the real trials began, participants were given a small number of practise trials (a minimum of 20) which were almost identical to the real trials which are detailed and illustrated in Figures 1 and 2 below. Unique instructions appeared before each practise block to explain the task for that block. Five practise trials were given for each experimental block to expose participants to each task type. Participants were required to respond correctly to at least four of the five judgment phase questions in each practise block before the program would move on to the next practise block. No accuracy quota was imposed for the memory phase responses. The five trials in each practise block were the same for all participants and were recycled in the case of the participant needing to take part in more than five trials for one practise block to attain four correct answers. The differences between these practise trials and the real trials were as follows: the practise trials gave feedback on participants' responses to both the judgment and memory

response phases, which real trials did not; and at stages where responses were required (judgment phase and working memory probe phase), practise trials did not time out. However, the stimuli did disappear at the same times as they would in the real trials, and after 6s of no response, a message appeared on-screen to prompt the participant to give a response. In the practise trials, participants still needed to give responses after this message appeared for the trials to progress.

After participants completed the practise trials, they were instructed that the real trials would now begin and warned that these would be considerably faster-paced (this was in response to pilot feedback which expressed that the change from lenient practise trials to fast-paced real trials was off-putting). As was the case in the practise trials, specific instructions presented before each block of trials informed participants which keys were the appropriate response type for each block.

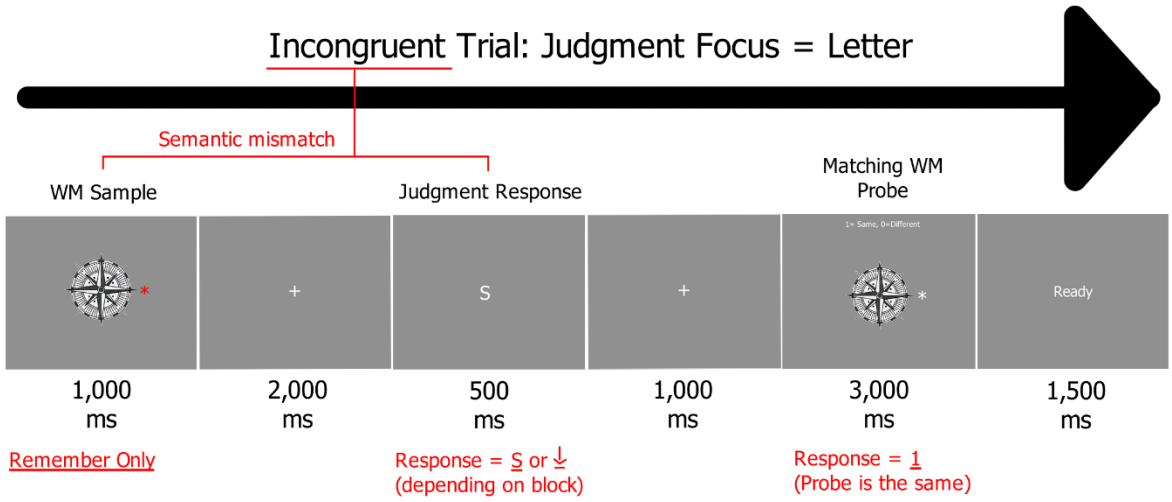

Figure 1 shows the structure of an incongruent trial wherein the Focus of the judgment phase is letter information, and the Focus of the memory phases is location information.

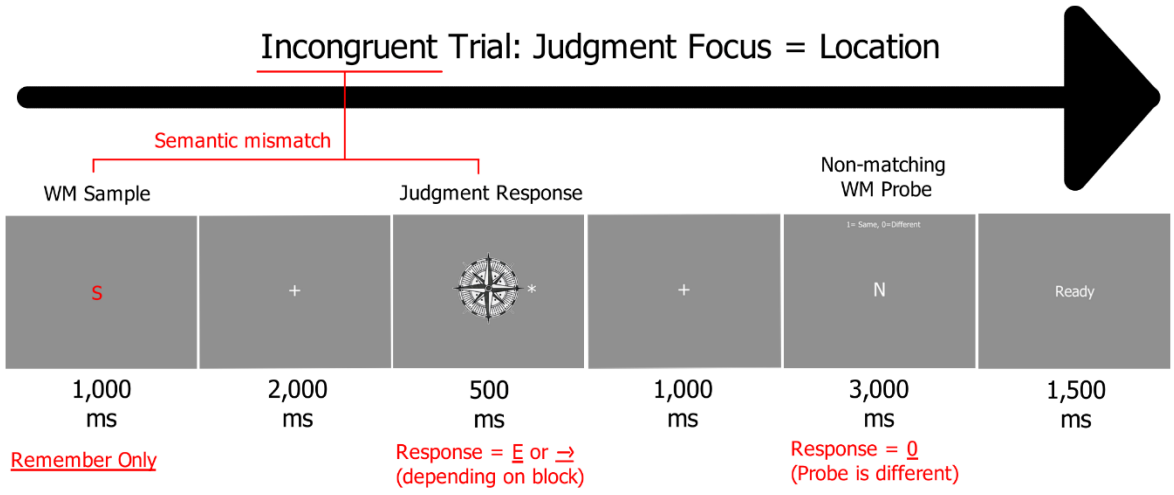

Figure 2 shows the structure of an incongruent trial wherein the Focus of the judgment phase is location information, and the Focus of the memory phases is letter information.

The trials in this experiment consisted of three phases in order: the working memory (WM) sample phase, the judgment phase, and the WM probe phase.

In the WM sample phase, participants were shown either a letter or a location presented in red font for 1,000ms, which they knew from instructions and the practise trials that they were tasked with remembering. This was followed by a fixation cross presented in the centre of the screen for 2,000ms. No response was required in this phase.

In the following judgement phase, participants were shown a stimulus for 500ms which was of the opposite kind to that shown in the WM sample phase (e.g., if shown a letter to remember, they were shown a location to judge). Judgment phase stimuli could be semantically congruent or incongruent in relation to the WM sample stimuli. Figures 1 and 2 above show two examples of incongruent trials. Here, participants were required to make a response on their keyboards: either by pressing the arrow keys, or the letter keys “N”, “E”, “S” or “W”. The diagrams below illustrate this mapping, but the responses are designed to be intuitive: if the judgment stimulus is “S”, participants must respond by pressing either the S key, or the down arrow key (which corresponds with the direction on a compass of ‘South’). If participants did not respond to this phase within the 500ms of stimulus presentation or the following 1,000ms of fixation cross time, they were shown a message in red text lasting 3,000ms which prompted them to respond more quickly. Responses

during or after this message were not recorded and these trials were considered non-responses.

The final phase of each trial was the WM probe phase, which presented participants with a stimulus lasting 3,000ms of the same type (letter or location) as the WM sample and prompted them with on-screen instructions to press 1 if this item matched the one that they previously committed to memory for this trial, or 0 if it did not match (the exact instruction prompt text was “1= Same, 0= Different”). It is perhaps important to note that participants were not able to respond during this phase using the 1 and 0 keys found in a NUM pad to the right of some computer keyboards. Responses were only accepted if participants responded using the 1 and 0 key located above the letter keys in the traditional location on a QWERTY keyboard.

At the very end of each trial (and for a duration of 3,000ms after each instruction screen), a ‘gap’ screen containing the text “Ready” was presented for 1,500ms to give participants a small break between each trial and an opportunity to move their hands from the 1 or 0 key after their WM probe response back to the relevant keys (arrow or NESW) in preparation for the judgment phase.

Each block consisted of 48 trials. Within each block, there were equal numbers of congruent and incongruent trials. All locations and letters were also equivalently presented (the letter “S” appeared 12 times, as did “W”, as did the asterisk above the compass, and below, etc.) and in the WM probe phase, the correct answer was 0 as often as it was 1.

At the end of the experiment, all participants were thanked and debriefed. This study was approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Participants’ data were included in analyses if they scored greater than 66% overall (this was calculated using their combined judgment and memory trials accuracy scores, as was done in the study by Kiyonaga and Egner, 2014) and answered at least eight judgment trials correctly for each cell (e.g., the incongruent

trials of the arrow key response block of the letter focused trials). The application of these criteria resulted in an *N* of 53 participants (9 from the lab, the rest took part online). This is a slightly larger sample size than was pre-registered due to the sampling method: early batches of online participants had very high exclusion rates, and therefore later batches were made larger to account for that but had comparably low exclusion rates. The data used in the following response time analyses were the log transformed response times (s) for the correct responses to the judgment and memory questions. The data used in the following accuracy analyses were the arcsine square root transformed proportions of correct judgment and memory responses. These transformations were applied to attempt to counter skew and align the data with the assumptions required by the ANOVA analyses.

The independent variables were the Focus of the task (what kind of information participants were remembering and which they were making judgments about), the Response Type (whether participants responded to judgment items using the Arrow keys or the letter keys N, E, S & W) and Congruence (whether the memory item presented and the to-be-judged item have the same meaning: e.g. the letter N and the location at the top of the compass). The analyses below addressed each of the four dependent variables of transformed response time and accuracy data for both the judgment and memory responses. We include judgment accuracy here for completeness, even though Kiyonaga and Egner (2014) did not find to be sensitive to the effect of congruence in their study. In-line with their finding, we pre-registered analysis of that measure as “secondary analysis”, however, it is reported here on equal footing with the other measures to ensure consistency with later experiments. Additionally, follow-up analyses were performed on the data which visually conformed to the expected pattern of results, to unpick in finer detail where the data did and did not align with our tested model with regards to interactions. These were composed of 2-way and 1-way ANOVAs on different sub-sections of the data (e.g., only the Location Focus trials). The measure in which we are most interested is the judgment response time measure, as accuracy is typically very high in tasks such as this and therefore the accuracy data can sometimes be uninformative. We anticipate that the findings in the other measures may sometimes give support to our conclusions drawn from the judgment response time measure, but we do not wish to rely on their sensitivity. The data were transformed in R Statistical software (v4.1.2,

R Core Team 2021) and Microsoft Excel (Microsoft Corporation, 2018), the participant exclusions were done manually, and the analyses were run in IBM SPSS Statistics version 29 (IBM Corp, 2023). Axis scales have been altered so that graphs reporting accuracy measures have the same y axis for all three experiments reported here. This is to better enable comparison of sizes of effects, and to avoid over-interpreting small but statistically significant outcomes.

Judgment Response Times

In the 3-way ANOVA conducted on judgment response time data, the key three-way interaction between Focus, Response Type and Congruence was significant ($F(1,52)=10.226$, $p=.002$, $\eta_p^2=.164$). This indicates that Stroop-like interference (measured by the difference in response time between congruent and incongruent trials) occurs to different extents depending on the conjunction of Focus and Response Type. On the surface, this interaction seems to support our hypothesis that the translational model also applies to the delayed Stroop task. However, closer inspection of Figure 3 below shows that while the Location Focus data appear to adhere to the expected pattern (a visibly smaller effect of congruence when the response type complements the task focus; right panel of Figure 3), the Letter Focus data do not even show the characteristic Response Type preference of faster responses for NESW keys, the complementary response type.

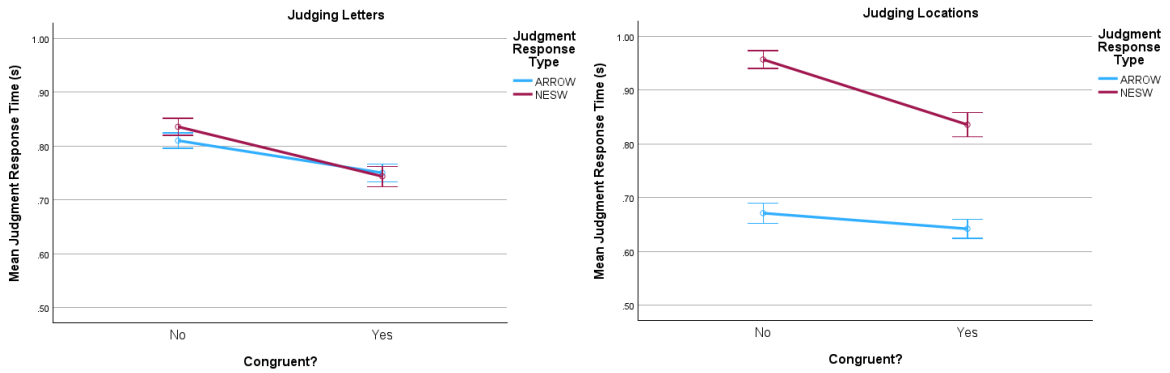

Figure 3 shows the mean average of participants' response times in seconds to the judgment items as a function of Congruence (on the x-axis), Response Type (the legend), and judgment Focus (one graph for each). Error bars represent +/- 1 standard error.

To further investigate the nature of the significant three-way interaction, we followed up with separate 2-way ANOVAs. A 2-way ANOVA was conducted on the letter focus trials only, including the factors of response type and congruence. These letter judgment findings do not replicate the story of previous research done in

simultaneous spatial Stroop tasks, as the effect of Congruence is of a significantly greater magnitude in NESW key response trials than in Arrow key response trials, as demonstrated by the two-way interaction ($F(1,52)=5.719$, $p=.020$, $\eta_p^2=.099$), which is the opposite of the expected result when responding to letters, since the NESW trials should be where no translation need occur.

A 2-way ANOVA was also conducted on the Location Focus trials only, including the factors of Response Type and Congruence. The main effects of Response Type ($F(1,52)=154.126$, $p<.001$, $\eta_p^2=.748$) and Congruence were significant ($F(1,52)=53.998$, $p<.001$, $\eta_p^2=.509$), as was the interaction between them ($F(1,52)=35.083$, $p<.001$, $\eta_p^2=.096$). To further investigate whether this significant interaction indicates that the effect of Congruence is absent when participants responded using Arrow keys, or just that the extent of the significant effect of Congruence depends on the level of Response Type, a further 1-way ANOVA was conducted on each Response Type data set with Congruence as the only factor. This analysis confirmed that Congruence produced a significant effect in both cases but to significantly different extents at each level of Response Type (Arrow key ($F(1, 52)=13.517$, $p<.001$, $\eta_p^2=.206$) and NESW key ($F(1, 52)=60.977$, $p<.001$, $\eta_p^2=.540$)).

Other findings from the 3-way ANOVA are as follows: Response Type ($F(1,52)=106.523$, $p<.001$, $\eta_p^2=.672$) and Congruence ($F(1,52)=72.480$, $p<.001$, $\eta_p^2=.582$) were significant main effects, whereas Focus was not significant at the $p<.05$ level. The two-way interactions between Focus and Response Type ($F(1,52)=124.012$, $p<.001$, $\eta_p^2=.705$) and between Response Type and Congruence ($F(1,52)=33.009$, $p<.001$, $\eta_p^2=.388$) were significant, whereas the interaction between Focus and Congruence was not. The results concerning the main effects indicate that participants were faster at responding to congruent than incongruent trials and to respond with arrow keys than letter keys, but that participants were no faster overall at responding to letter or location information. Critically, the significant main effect of Congruence replicates the findings from Kiyonaga and Egner (2014) and also supports the notion that their delayed Stroop task, which was previously conducted with colour-word stimuli, can also be successfully undertaken with spatial stimuli.

Judgment accuracy

A 3-way ANOVA conducted on the judgment accuracy measure revealed that the key three-way interaction between focus, response type and congruence was not significant, though visually the data reproduces the pattern of results expected by the translational model: the graphs in Figure 4 below suggest better accuracy when response type and focus correspond, but also that this interaction may extend to the size of the effect of congruence. However, as stated the evidence for this interaction did not reach significance. Further ANOVAs were conducted to determine whether the data from the location focus trials conformed to any of the predictions set out by the translational model. A 2-way ANOVA revealed significant effects of both response type ($F(1,52)=64.832$, $p<.001$, $\eta_p^2=.555$) and congruence ($F(1,52)=17.597$, $p<.001$, $\eta_p^2=.253$), plus a significant interaction between them ($F(1,52)=4.477$, $p=.039$, $\eta_p^2=.079$), though this effect was quite small. An ANOVA on just the arrow key responses showed that this response type is not susceptible to interference in this circumstance, as the effect of congruence was not significant. Further in line with the prediction, congruence was a significant main effect in an ANOVA on just the NESW trials ($F(1,52)=21.309$, $p<.001$, $\eta_p^2=.291$).

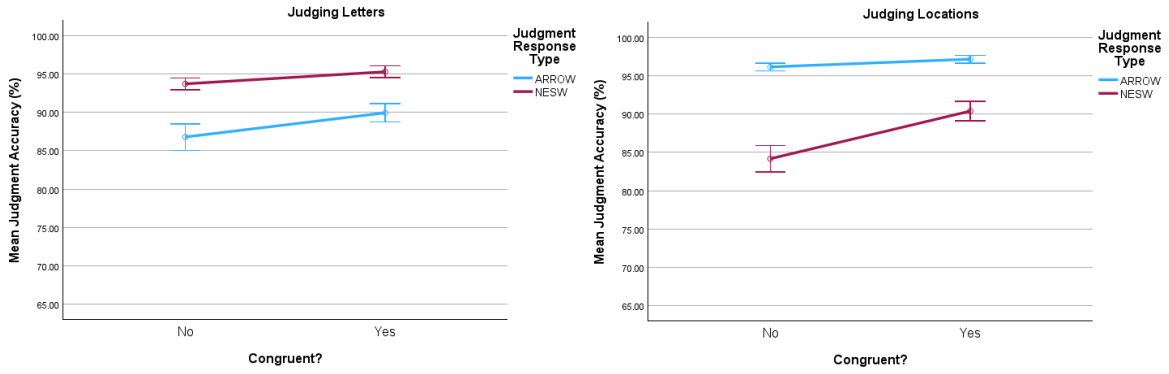

Figure 4 shows the mean accuracy of participants' responses as a percentage for the judgment items as a function of Congruence (on the x-axis), Response Type (the legend), and judgment Focus (one graph for each). Error bars represent +/- 1 standard error.

With regard to other findings from the 3-way ANOVA, we found two significant main effects: response type ($F(1,52)=4.734$, $p=.034$, $\eta_p^2=.083$) and congruence ($F(1,52)=16.663$, $p<.001$, $\eta_p^2=.243$). This moderately-sized main effect of congruence is a contradiction to our expectation following Kiyonaga & Egner's null result in this measure. For two-way interactions, only focus by response type

($F(1,52)=92.945$, $p<.001$, $\eta_p^2=.641$) was significant, reflecting that judgment accuracy was better when the response type complemented the task focus.

Memory Response Times

In a 3-way ANOVA on the correctly answered memory item response time data, the three-way interaction was significant ($F(1,52)=7.571$, $p=.008$, $\eta_p^2=.127$), indicating that Congruence impacted response times to different extents depending on the conjunction of both Response Type and Focus, however, as Figure 5 reveals, the data did not conform to the expected pattern of less interference when a translation was necessary. These findings corroborate one of the findings from Kiyonaga and Egner (2014), who reported that Congruence had a significant effect on memory item response times, but do not support the predictions put forward by the translational model.

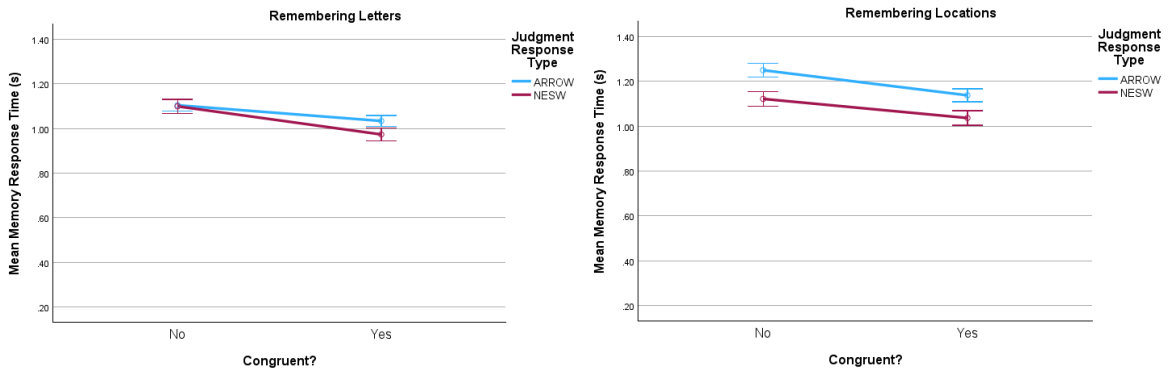

Figure 5 shows the mean average of participants' response times in seconds to the memory items as a function of Congruence (on the x-axis), Response Type (the legend), and judgment Focus (one graph for each). Error bars represent +/- 1 standard error.

With regard to the other findings, all three main effects were significant. Congruence had a significant effect ($F(1,52)=94.414$, $p<.001$, $\eta_p^2=.645$): trials where the intervening judgment items were incongruent with the to-be-remembered memory item tended to elicit slower recognition response times. Focus was a significant main effect ($F(1,52)=25.693$, $p<.001$, $\eta_p^2=.331$), meaning that participants were faster to respond to letter memory probes than location memory probes overall, perhaps suggesting that letters are processed faster than compass directions/points. Response Type was significant ($F(1,52)=17.713$, $p<.001$, $\eta_p^2=.254$), meaning that participants were faster to respond to memory probes following a NESW key judgment response than an arrow key judgment response. Finally, the interaction

between Focus and Response Type had a significant influence on response times ($F(1,52)=8.448$, $p=.005$, $\eta_p^2=.140$).

Memory Accuracy

A 3-way ANOVA conducted on memory accuracy data revealed that the key three-way interaction between task focus, response type and congruence was not significant at the p<.05 level, indicating that the extent of the effect of congruence did not vary as a result of the factors of response type and focus. With regard to the other findings, the 3-way ANOVA first revealed a significant main effect on memory performance was Congruence ($F(1,52)=10.712$, $p=.002$, $\eta_p^2=.171$). This replicates the findings of Kiyonaga and Egner (2014): in trials where the WM item and Judgment item were semantically congruent (e.g., the letter 'N' and the location at the top of the compass), the memory items were answered more accurately than trials where the WM item and Judgment item were semantically incongruent (e.g., the letter 'E' and the location at the top of the compass). The interaction between Focus and Response Type was also significant ($F(1,52)=13.506$, $p<.001$, $\eta_p^2=.206$), which is visible in Figure 6 in the reversal of preference for Response Type depending on Focus In the Remembering Letters graph (left panel): when participants had responded to the location judgment questions with the arrow keys (a response not requiring a translation), they were better at accurately recognising the following letter memory probe than when they had used NESW keys (and vice versa in the Remembering Locations graph). This finding is encouraging for the Translational Hypothesis, but incomplete as evidence, as the theory would predict a three-way interaction also including Congruence, which was not found. All other factors and interactions were non-significant at the $p<.05$ level.

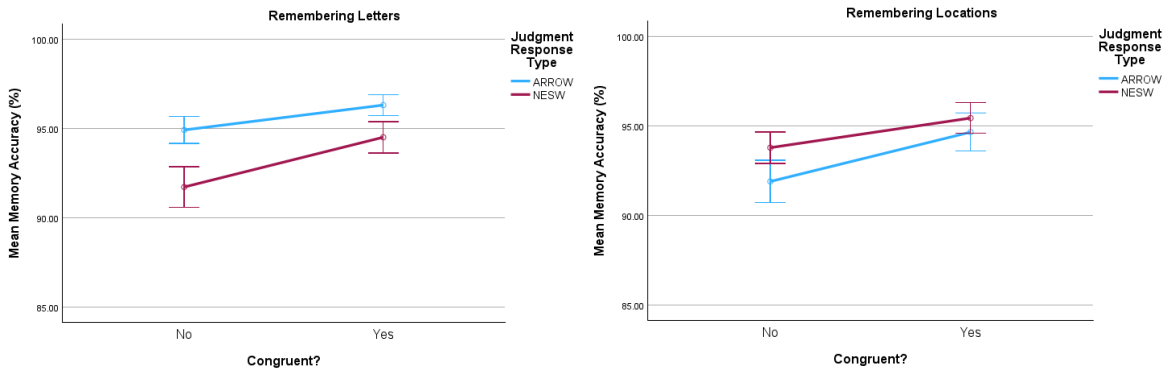

Figure 6 shows the mean accuracy of participants' responses as a percentage for the memory items as a function of Congruence (on the x-axis), Response Type (the legend), and judgment Focus (one graph for each). Error bars represent +/- 1 standard error.

Discussion

In every measure, Experiment 1 showed that Stroop-like interference was manifesting and affecting both accuracy and response times for both the judgment task and the memory task. Performance also typically varied with response type, with only the memory accuracy measure not being significantly influenced by whether participants responded to the judgment task with arrow keys or NESW keys. In the judgment measures, arrow key responses were faster and more accurate, perhaps due to their well-known relative locations and separateness from non-response keys. However, in the memory task, it was the NESW response block which had the significantly faster responses, perhaps an artefact of the proximity of the NESW keys to the memory response keys of 0 and 1 on a keyboard. In the accuracy measures, participants were more accurate in the judgment task when responding to letters when they used the letter keys and to locations when they used the arrow keys, and then they were also more accurate in the memory task following these complementary judgment responses. The measure in which we were most interested, judgment response time, demonstrated an interaction between all of the variables, however it is clear from the graphs that the nature of this interaction does not conform to the expectations set out by the translational model. We would expect to see a slowing of responses due to incongruent stimuli to a much greater extent when the to-be-judged stimulus domain is in contrast with the planned response domain, which was detected in the location-focus trials, but not in the letter-focus trials.

The findings of this experiment are very encouraging for part of the conclusions set out by Kiyonaga & Egner: performance was better and faster for congruent trials than incongruent ones in both the judgment task and the memory task. This includes a measure in which their method was not able to capture the effect, judgment accuracy. Thus, our Experiment 1 definitely supported Kiyonaga & Egner's primary finding by demonstrating it within a second sub-domain – Stroop-like interference is happening in both verbal-colour and verbal-spatial stimulus sets when the interfering item is held in working memory. However, their extended conclusion

that holding an item in working memory elicits Stroop interference *exactly* like seeing it in real time is not well-supported by this data. Our previous studies (Delooze & Morey, 2024, also Chapter 2 of this thesis) show that using these stimuli and these response methods in a simultaneous presentation version of this task (the more classic Stroop formula) did elicit the expected pattern detailed by the translational model. The only difference from these methods to those is that the interfering item here was being held in working memory instead of appearing alongside the to-be-judged item, which appears to prevent the pattern from manifesting. This suggests that contrary to Kiyonaga and Egner's findings regarding response level conflict and proportion of incongruence effects, this 'working memory Stroop' method does not elicit Stroop interference in the exact same way as the traditional method, since not all effects which are shown to moderate Stroop interference in the simultaneous method are observed here.

These results are not entirely encouraging for the translational model because our data did not demonstrate the exact pattern of effects we expected. The reversal of preference for response types depending on task focus that happens in both memory and judgment accuracy measures is encouraging, but the amount of interference experienced was not impacted by that interplay. There is a glimmer of hope for the translational model to be found in the judgment accuracy data, which visibly approaches the expected pattern, though the three-way interaction did not reach significance. Specifically in the location focus trials, the pattern is produced perfectly: no interference in the arrow key response type trials (the response not requiring a translation), but considerable interference in the verbal response type trials (the response requiring a translation). In the judgment response time data, which we value as the most useful measure, the story is similar, with the data from location focus trials adhering to our expectations better than those from the letter focus trials. Success in the spatial dimension but not in the verbal dimension may imply that the issue could be just with the nature of the 'verbal' response type, and not necessarily the whole method. Perhaps, while the NESW key response type used here is sufficiently aligned with the verbal system that it can emulate a spoken response when the interfering item is present (as demonstrated in Delooze & Morey, 2024, also Chapter 2 of this thesis), it is not sufficiently strongly aligned that it can emulate a spoken response when the interfering item is only held in memory. A

caveat to this idea is that *here*, the expected pattern is successfully depicted in the location judgment response time data, but in the simultaneous Stroop data (Delooze & Morey, 2024, also Chapter 2 of this thesis), it was in the letter focus data that this manifested as expected, and in the location focus data that the pattern was flawed. This inconsistency between which stimuli (and thus which response type) suffer from this incomplete translational model effect would seem to suggest that the issue is not inherent to the response methods, but elsewhere.

Alternatively, it may be possible that this wider failure to detect the expected pattern of results is an issue inherent with separating the verbal and spatial stimuli in time and that under other circumstances, the expected pattern would be observed in both task foci. The literature discussed earlier on the topic of stimulus onset asynchronies in Stroop research suggests that the strength of an interfering item varies as a result of time elapsed since presentation, and not in a way that we wholly understand. It also suggests that these patterns of interference strength may vary with different stimulus types. In fact, there is a plethora of evidence that different types of information are not treated equivalently by the working memory system. Take first the discussed research showing that the reverse Stroop effect does not usually happen naturally or is very small in colour Stroop paradigms, but that this can be witnessed very easily in spatial Stroop paradigms. Additionally, cognitive researchers have developed a pretty consistent method of interrupting the verbal rehearsal system (articulatory suppression), but spatial memory is unreliably affected by various tasks designed to inhibit it, such as spatial tapping (e.g., Zimmer et al., 2003). Finally, there is evidence to suggest that specifically in verbal-spatial memory, associated spatial information may be retained ‘for free’ over a short period when we make an effort to commit verbal information to memory, but that the reverse does not occur (Campo et al., 2010; Elsley & Parmentier, 2015; Chapter 5 of this thesis; though see Delooze et al., 2022 for a collection of studies wherein this was not the case). If the replication issue here is not inherent to the verbal response method used, perhaps it is instead due to the item held in memory not having sufficient power of interference due to natural differences across stimulus types. To address these two possibilities, Experiments 2 and 3 were conducted. Experiment 2 maintained the letter key ‘verbal’ response but endeavoured to strengthen the interfering power of the memory item by changing the memory task from recognition

to recall, and if this led to the anticipated pattern, to determine whether this was due to improved memory item strength or the nature of the planned recall response. Experiment 3 replaced the letter key responses in Experiment 2 with vocal responses to assess whether vocal responses were indeed better aligned with verbal input than our stand-in verbal key response.

Experiment 2

Cognitive research has historically viewed recognition as different than recall. People approach item encoding in memory tasks significantly differently if they think that they will be required to recognise versus when they will be required to recall the items (e.g., Carey & Lockhart, 1973; Tversky, 1973; Hall et al., 1976, Uittenhove et al., 2019). Hall et al. (1976) concluded that the difference in encoding which they detected in their method was very likely quantitative rather than qualitative, ergo when stimuli are encoded for recall, the strength of the representation may be greater than when they are encoded for recognition. This is in-line with Postman et al.'s (1948) idea that items can only be recalled if their representational strength exceeds a certain threshold, whereas items may be recognised with much weaker representations. If in our experiment, the memory stimuli are being encoded with less-than-optimal strength, perhaps this may not be sufficient to elicit full Stroop-like interference in the same way that currently viewed stimuli do in simultaneous Stroop methods. As discussed above, different types of information are often treated differently by our cognitive systems, which may result in the finding from Experiment 1 that location-focus trials aligned somewhat with our expectations, but letter-focus trials totally deviated from them. It is possible that lower levels of representational strength are necessary for remembered letters to interfere with judgments of locations than the reverse, which might manifest as the observed pattern of results in Experiment 1.

Since the nature of the response type very much matters in the simultaneous Stroop paradigm, we posited that it may also matter in this working memory Stroop paradigm, specifically that by implementing a recall response which better suits one stimulus type compared to the other, we may see differing patterns of results in the judgment performance data. There is some work in the working memory literature to support this notion that the nature of a planned response impacts perceptual

cognition. First, it is relevant to establish that, in-line with the findings above regarding differential encoding for recognition versus recall memory, there is evidence from neuro-imaging studies that memory for an item is stored in a way that complements the to-be-enacted response action if this is known. For instance, Henderson et al. (2022) studied BOLD signals in visual and sensorimotor regions of interest during an item localisation task. They found that when participants could plan their response in advance using an informative response cue, they stored the presented information as a motor code representing the to-be-enacted response in the sensorimotor areas. Contrastingly, when they could not plan their response in advance due to the presentation of an uninformative response cue, the information from the trial was much more likely to be stored as a spatial code in the visual areas of interest. As for how these differentially stored planned responses influence perceptual cognition, Fagioli et al. (2007) found that participants were more accurate in a visual discrimination task when the target attribute (size or location) that they were attending to was complemented by the response they planned to use to identify it (grasping or pointing, respectively). This is an example of a complementary planned response boosting visual processing of the to-be-responded-to item, which may map onto Stroop tasks very closely.

Similarly, and more relevant to our current method where two tasks are being enacted in each trial, Heuer and Schubö (2017) found that the planning of an unrelated complementary response (grasping or pointing) to be enacted later (not on the item in question) boosted participants' accuracy in recognising whether a probe matched or mismatched the previously shown display on the relevant attribute (size or colour, respectively). This work demonstrates that a held action plan can also influence the processing of an unrelated item in an interleaved task, similar to the effect that we expect that planned memory responses may have on the interleaved judgment task in the current studies. The translational model emphasises the connection between the to-be-inhibited information and the required response method: when they align, it takes extra time and effort to inhibit an automatically planned response concerning the interfering item and instead output the response tied to that system using the information which requires translation. Thus, one reason as to why the pattern expected by the translational hypothesis was not observed in the letter focus trials in the method run in Experiment 1 could be that

when participants planned a recognition response (pressing 1 or 0), they overwrote the usually automatic response planning which goes on as a result of the parallel input-to-output systems (in this case, movement in the target direction) and causes the need for inhibition in simultaneous Stroop tasks.

The findings discussed above lead us to suspect that if participants were to be required to recall the location rather than to recognise it, we may see a clearer effect of translation in the letter focus trials which previously failed to demonstrate the expected pattern, in which strong Stroop-like interference is observed when letters are judged using arrow responses, but little (or no) interference is observed when letter responses are used. If the replacement of the recognition task with a recall task successfully elicits the pattern characteristic of the translational model, we expect that this is due to one or the other of the following factors: the nature of the action plan that is held, or the nature of the encoding. If the to-be-remembered item is not simply *boosted* by the intention to recall, but instead there is an interfering effect of the nature of the planned recall response, we expect to see a different pattern of interference in the judgment task when participants are holding in mind a recall action plan with the complementary arrow key response than one using the non-complementary letter keys.

Contrastingly, if the nature of the memory item's encoding – stronger for recall than for recognition – is the key, then the nature of the recall plan (memory response type) should not affect the judgment. In that case, we would expect to see an interaction (which is not further mediated by memory response type) between judgment response type and congruence in the judgment measures. To be specific, this would consist of no (or significantly less) effect of congruence in the response type which complements the stimuli and requires no translation (NESW keys), compared to a significant effect of congruence in the response type which does not complement the stimuli and thus requires a translation (arrow keys). Here, we would believe that eliciting the characteristic pattern of the translational model is made possible by the stronger representations of the interfering location item which result from the intention to recall it.

Method

There are two major methodological changes in Experiment 2 compared to Experiment 1. First, in Experiment 2 the memory phase of every trial requires participants to *recall* the item they committed to memory instead of indicating whether they recognise a probe, as participants did in Experiment 1. In half of trials, this was done with the arrow keys, and in the other half this was done with keys marked N, E, S and W. Second, to shorten the duration of the experiment and reduce the likelihood of fatigue effects, the location focus trials were removed: this meant that in all trials in this experiment, participants were tasked with remembering locations and judging letters. Since the expected pattern was already seen in the location focus trials using the recognition method, testing the effect of recall in these trials was considered less of a priority. Some less important changes were also made regarding the way in which the experiment was hosted (Experiment 1 was hosted online while Experiment 2 was conducted in the lab) and the verbal response method (running Experiment 2 in the lab allowed us to re-use the USB NUM pads from earlier experiments with simultaneous spatial Stroop methods).

Participants

Sixty-two students were opportunistically recruited from Cardiff University's undergraduate psychology course through the Experiment Management System and took part in exchange for course credit. As in the previous experiment, no demographic information was collected due to there being no reason to suspect any confounding effects of any particular participant attribute. However, this limits the conclusions which could be drawn from this data regarding generalizability. The student population from which participants were taken was mostly female, mostly aged 18-24 years, and all were fluent speakers of English, as required of students taking part in a course in the medium of English. All participants attended a lab session to take part in the experiment. Exclusion criteria are detailed in the Results section below.

Materials and Apparatus

For Experiment 2, all participants took part in-lab using PsychoPy (Peirce et al., 2019) hosted locally using an iiyama ProLite XUB2294HSU, 21.5-inch monitor with a maximum resolution of 1920x1080 pixels. Participants were asked to respond

to both the judgment probe and the recall probe using the two detachable NUM pad keyboards which can be seen in Figure 7, one which was spatially oriented to mirror the compass display and one wherein the keys were intended to be spatially neutral, marked with the letters “N”, “E”, “S” and “W” (all in a horizontal line). The stimuli in Experiment 2 were identical to those used in Experiment 1.

Figure 7 shows the USB NUM pads on which participants responded when taking part in Experiment 2. The left-most NUM pad has four keys, each one marked with one of the following letters: N, E, S and W. This NUM pad is the ‘verbal’ response pad. The right-most NUM pad is marked with four directional arrows representing the cardinal directions of North, East, South and West.

Design

The design of this experiment differed from Experiment 1 in that the independent variable of task focus was removed: the only trials which were included were those where the to-be-remembered items were locations, and the to-be-judged items were letters. A different two-level independent variable was introduced in this experiment, which was the planned recall response type: NESW keys or arrow keys. This resulted in a three-way within-subjects design of 2 (Judgment Response Type: NESW keys/Arrow keys) x 2 (Memory Response Type: NESW keys/Arrow keys) x 2 (Congruence: Congruent/Incongruent). The only independent variable which was not

blocked was Congruence, with both congruent and incongruent trials occurring in every block.

Procedure

Before participants were given any instructions regarding the primary task, they were first familiarised with the locations of the N, E, S and W keys on the relevant NUM pad (see Figure 7). This was achieved with a short and simple familiarisation task consisting of 120 simple trials (number of trials was chosen to approximately reach the 128 trials used by MacLeod, 2005) wherein a letter (N, E, S or W) appeared in white text on-screen, presented centrally, and participants were tasked with pressing the correspondingly labelled letter key on the letter NUM pad. When these were finished, participants were asked by the experimenter if they felt confident that they knew the locations of the keys, and if their answer was that they did not, the experimenter would re-run the familiarization task until the participant responded that they felt confident.

Once familiarity was established, the real task would begin. This consisted of four blocks with two requiring NESW key judgment responses and the other two requiring arrow key judgment responses, as in Experiment 1, but instead of task focus varying with block, memory response type varied. In this way, the four blocks are as follows: Judgment: NESW, Recall: NESW; Judgment: NESW, Recall: Arrow; Judgment: Arrow, Recall: NESW; Judgment: Arrow, Recall: Arrow.

These blocks could occur in any order. Before each block, participants were given task instructions and six practise trials with correct/incorrect feedback for each response, specific to that block. The re-arrangement of practise blocks to occur before each block instead of all together at the start was in response to poor performance in Experiment 1. An additional change also made in response to this was that during the time out message which appeared following a lack of response during the 1.5s judgment period, participants were able still to make their response. For the analyses reported here, these responses were not used as the in-lab setting of Experiment 2 resulted in considerably fewer exclusions, but they could be obtained from the data if they were required. The structure of trials was identical to Experiment 1, except that at the end of the trial, instead of seeing a probe for a recognition response, participants saw text indicating that they should recall the item

they were maintaining in working memory. This text read, for example, *"Which direction do you remember? Please respond with: NESW"*, and was presented across three lines, centrally aligned, in the same font as the other text within the experiment, all in white except the response type text ("NESW" or "arrow"), which was presented in red for emphasis. At the very end of the experiment, a language check question was asked of participants, which was worded as follows.

"Thank you for your time and effort! Before you finish the experiment, please respond to the question below. Your response will not affect your credit payment.

Is English your first language?

1= YES, 3= NO, 7= PREFER NOT TO SAY"

Participants responded on either NUM pad, as all three numbered options were clear of stickers on both pads. Following Experiment 1, a colleague suggested that participants whose first language was not English may struggle to associate the letters N, E, S and W with the directions on a compass, since it is not a structure that is commonly used in day-to-day life, and they would likely have learned different letter associations dictated by the equivalent directional words in their language. This first-language data was collected with the intention to only use it to rule out that the issue was not due to participants' lack of letter-direction associations if again our results were not in-line with what we expected. This proved to be unnecessary. This study was approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Participants were excluded from analyses if they failed to meet the following criteria: scored 66% accuracy or more overall, scored six or more correct in each trial type (e.g., NESW judgment, arrow recall, incongruent). This resulted in the exclusion of 19 participants, leaving a sample size of 43. Individual trials were excluded if participants responded more quickly than 0.2s, which is reasoned to be an error rather than a true response. Excessively slow responses were not possible, as the program moved on automatically after 1.5s for the judgment response and 3s for the memory response. For accuracy analyses, data was arcsine square root transformed

and for analysis of response time data, incorrect responses were excluded and all included data points were log transformed to better comply with ANOVA requirements. For all measures, first a 3-way ANOVA was conducted to determine whether there was a significant three-way interaction which would support our 'nature of action plan' hypothesis. Where relevant, follow-up ANOVAs were then conducted to assess the finer elements of the data. Any additional significant results separate from the three-way interaction were then reported.

Judgment Response Times

Importantly, the three-way interaction was significant ($F(1,42)=54.212$, $p<.001$, $\eta_p^2=.563$). The effect of the planned memory response is particularly impressive in the graphs within Figure 8, showing a complete reversal of the pattern from one memory response type to the other. When the planned memory response is complementary to the stimulus it relates to, in the Memory Response Type = ARROW graph, we see that the judgment response time is susceptible to interference when it requires a translation ($F(1,42)=64.972$, $p<.001$, $\eta_p^2=.607$). Meanwhile, when the judgment response requires no translation, the effect of congruence is still significant ($F(1,42)=4.936$, $p=.036$, $\eta_p^2=.105$), but the effect size is significantly smaller, as indicated by the two-way interaction ($F(1,42)=43.944$, $p<.001$, $\eta_p^2=.511$). When the planned memory response is not complementary to the stimulus it relates to and a translation is required, in the Memory Response Type = NESW graph, the arrow key response (which is still significant, ($F(1,42)=35.515$, $p<.001$, $\eta_p^2=.458$)) is the one that is less affected by interference, as indicated by the two-way interaction ($F(1,42)=26.344$, $p<.001$, $\eta_p^2=.385$).

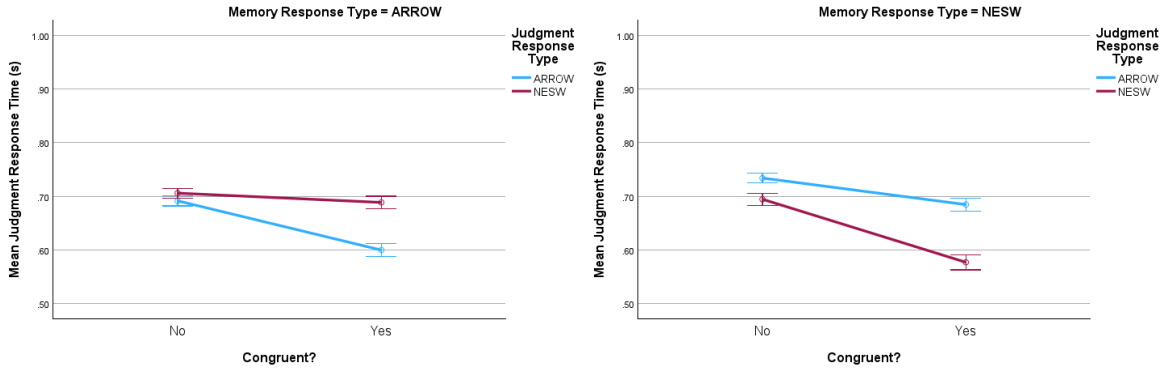

Figure 8 shows mean response times for the letter judgment task as a function of Congruence (x-axis), judgment response type (legend) and planned memory response type (one graph for each). The error bars represent +/- 1 standard error.

Only one main effect was significant, congruence ($F(1,42)=107.044$, $p<.001$, $\eta_p^2=.632$), which also interacted significantly with memory response type ($F(1,42)=10.432$, $p=.002$, $\eta_p^2=.199$). One other two-way interaction was significant, judgment response type by memory response type ($F(1,42)=72.252$, $p<.001$, $\eta_p^2=.632$), which illustrates that there is a notable boost to response time associated with making a judgment response with the same method as the participant intends to make their recall response.

Judgment Accuracy

The key three-way interaction was significant ($F(1,42)=21.334$, $p<.001$, $\eta_p^2=.337$). In the Memory Response Type = ARROW graph of Figure 9 below, it is clear to see the expected pattern of results, with the complementary judgment response type of NESW keys suffering from no interference as a result of stimulus incongruence ($F(1,42)=.007$, $p>.05$), but the non-complementary response of arrow keys which require a translation seeing a moderate effect of congruence ($F(1,42)=20.525$, $p<.001$, $\eta_p^2=.328$). In the other graph in Figure 9, Memory Response Type = NESW, a small sized interaction between response type and congruence is still present ($F(1,42)=7.627$, $p=.008$, $\eta_p^2=.154$), but with the significant effect of congruence being slightly but significantly smaller in the arrow key response type ($F(1,42)=16.473$, $p<.001$, $\eta_p^2=.282$) than in the letter key response ($F(1,42)=52.028$, $p<.001$, $\eta_p^2=.553$). This is evidence for the notion that not only does held information affect our processing and response to irrelevant items in an interleaved task, so do held action plans.

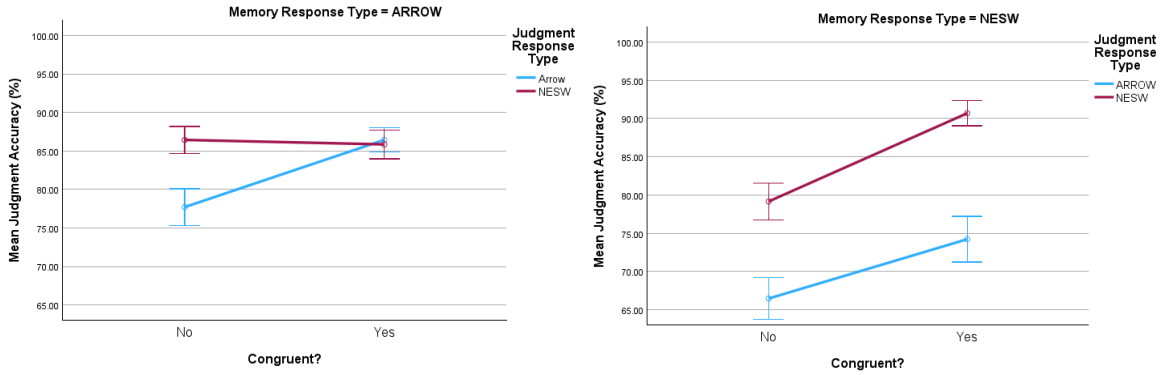

Figure 9 shows the mean accuracy as a percentage for the letter judgment task as a function of Congruence (x-axis), judgment response type (legend) and planned memory response type (one graph for each). The error bars represent +/- 1 standard error.

All three main effects were significant: judgment response type ($F(1,42)=39.336$, $p<.001$, $\eta_p^2=.484$), memory response type ($F(1,42)=16.405$, $p<.001$, $\eta_p^2=.281$) and congruence ($F(1,42)=47.883$, $p<.001$, $\eta_p^2=.533$). The two-way interactions of judgment response type by memory response type ($F(1,42)=19.561$, $p<.001$, $\eta_p^2=.318$) and memory response type by congruence ($F(1,42)=8.530$, $p=.006$, $\eta_p^2=.169$) were also significant.

Memory Response Times

The three-way interaction was significant ($F(1,42)=47.505$, $p<.001$, $\eta_p^2=.531$). The graphs in Figure 10 below again demonstrate the stark difference in the effect of interference experienced in the memory task as a result of the preceding judgment response type. When no translation was necessary in the judgment task, seen in the Judgment Response Type = NESW graph, the data show the characteristic pattern of the translational model: no effect of congruence on the complementary response - in this case, participants reported directions using arrow keys - ($F(1,42)=.358$, $p>.05$) but a significant effect of congruence ($F(1,42)=55.696$, $p<.001$, $\eta_p^2=.570$) in the trials using the non-complementary response type, where a translation was required. Contrastingly, if the judgment response preceding the recall response in question required a translation, as in the Judgment Response Type = ARROW graph, there is no discernible difference in the extent of the effect of congruence for the two memory response types.

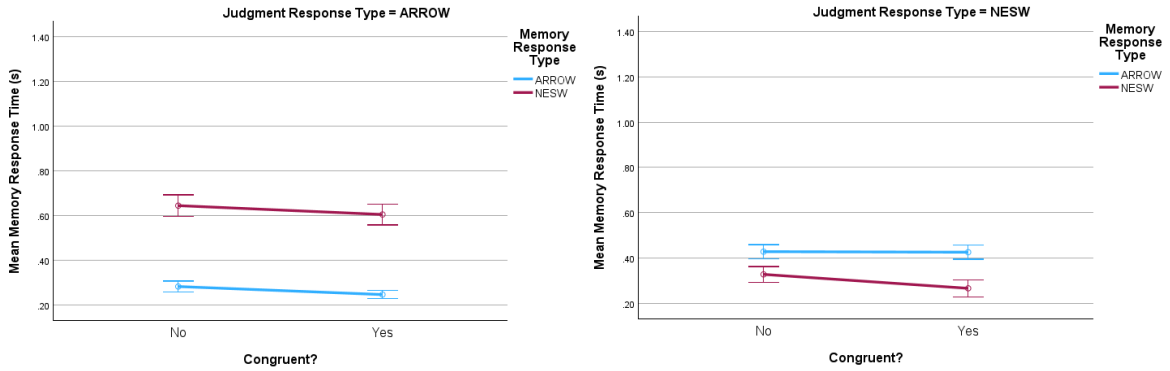

Figure 10 shows mean response times for the location recall task as a function of Congruence (x-axis), judgment response type (legend) and preceding judgment response type (one graph for each). The error bars represent +/- 1 standard error.

Additionally, all three main effects were significant: memory response type ($F(1,42)=9.077$, $p=.004$, $\eta_p^2=.178$), judgment response type ($F(1,42)=23.128$, $p<.001$, $\eta_p^2=.355$), and congruence ($F(1,42)=66.576$, $p<.001$, $\eta_p^2=.613$). All three of

the two-way interactions were significant: memory response type by judgment response type ($F(1,42)=116.321$, $p<.001$, $\eta_p^2=.735$), memory response type by congruence ($F(1,42)=13.110$, $p<.001$, $\eta_p^2=.238$) and judgment response type by congruence ($F(1,42)=5.621$, $p=.022$, $\eta_p^2=.118$).

Memory Accuracy

A 3-way ANOVA conducted on the memory accuracy data revealed that importantly, the three-way interaction ($F(1,42)=5.517$, $p=.024$, $\eta_p^2=.116$) was significant. In the Judgment Response Type = NESW graph in Figure 11 below, we can see the pattern characteristic of the Translational model, with the memory response type requiring no translation, the arrow keys, being on average more accurate and also less susceptible to interference than the response type requiring translation, the NESW keys. Conducting further ANOVAs on the data for the arrow key response alone revealed no effect of congruence ($F(1,42)=.091$, $p>.05$), whereas there was a significant effect of congruence for the NESW key data alone ($F(1,42)=5.358$, $=.026$, $\eta_p^2=.113$). This same characteristic pattern is not seen in the Judgment Response Type = ARROW graph in Figure 11, perhaps a result of the translation which was required of participants in the interleaved judgment task to output a seen letter as a spatial response. Following a translation for the judgment response, arrow key memory responses do not benefit from the protection against interference that is characteristic of the translational model.

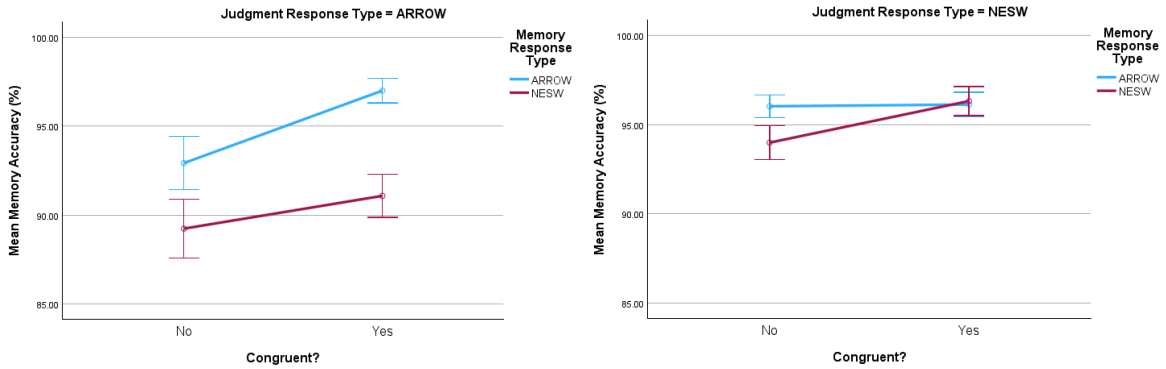

Figure 11 shows the mean accuracy as a percentage for the location recall task as a function of Congruence (x-axis), judgment response type (legend) and preceding judgment response type (one graph for each). The error bars represent +/- 1 standard error.

Additionally, all three main effects were significant: memory response type ($F(1,42)=16.395$, $p<.001$, $\eta_p^2=.281$), judgment response type ($F(1,42)=11.309$, $p=.002$, $\eta_p^2=.212$) and congruence ($F(1,42)=13.257$, $p<.001$, $\eta_p^2=.240$). The

interaction between memory response type and judgment response type was significant ($F(1,42)=10.631$, $p=.002$, $\eta_p^2=.202$), primarily illustrating a penalty for memory accuracy when both judgment and memory responses required a translation.

Discussion

To reiterate the most straightforward results, stimulus incongruence was a detriment to participant performance and response speed in both tasks, again corroborating Kiyonaga and Egner's finding of the occurrence of Stroop-like interference using this working memory Stroop paradigm. Stroop interference can clearly be observed when the interfering item is being held in working memory. The findings presented here also confirm that under some circumstances, the pattern predicted by the translational model can be witnessed in this working memory Stroop paradigm, adding support for Kiyonaga and Egner's specific claim that items held in working memory cause interference in the same way as those witnessed in real time. The current results build on their finding by starting to suggest and test mechanisms for how the held item elicits interference, and which factors affect the strength of that interference.

Importantly for our hypotheses specific to Experiment 2, analyses in all four dependent variable measures detected a difference in the extent of interference as a result of both judgment response type and memory response type. In all cases, this reflected that when one response type variable was complementary to the task it reflected, the extent of interference in the other task differed more drastically depending on whether response type for that task was complementary or not. When the response type was not complementary in one task, the sizes of the interference effects observed in the other task were much more similar regardless of the task-relevance of response type. To illustrate with an example, when participants held an arrow key response plan to the location memory stimuli, no translation was required for the memory task. This resulted in the characteristic translational model pattern of results manifesting in the letter judgment task: little or no interference effect when no translation was required for that task (participant used NESW keys), and a large amount of interference when a translation was required (participant used arrow keys). However, when the opposite was true, and a translation was required for the

planned memory response (NESW keys), this changed the pattern of results in the judgment task so that the pattern which we expected could not be observed. Instead, the sizes of the effects of interference were typically much more similar. These results support our first hypothesis, that the nature of the held action plan (the kind of response which will be made) influences the extent of Stroop-like interference that participants experience.

These results confirm our suspicion that in Experiment 1, we did not detect the predicted interaction between response type and congruence in the letter focus trials due to the recognition nature of the memory task. From these results we may suggest that at least part of the reason that the translational effect occurs is because of planning responses to the stimuli we see in the world around us, even if we do not consciously intend to enact them. We can speculate that by replacing the automatically planned response to the memory item with a plan for a recognition response, the method in Experiment 1 did not wholly recreate the influences experienced in simultaneous Stroop paradigms. This is because the maintained response to the memory item (akin to an automatically planned response in simultaneous Stroop methods) was a neutral yes/no response instead of the usual stimulus-complementary response type aligned with the relevant processing system. In this recall version of the paradigm, the planned response to the interfering stimuli was again made complementary to the stimuli, and it paved the way for the expected interaction to occur. It appears that in Experiment 1, this overwriting of response plan with a recognition response only occurred for the trials wherein participants had to remember locations, since the trials wherein they had to remember letters produced the expected pattern of results. Is there something about letter stimuli compared to locations which preserves their automatic response plans even in the face of planning a recognition response? In further support of this idea of automatic response planning and beginning to suggest that there may be hierarchies of response planning automaticity across different stimulus types, Uleman and Reeves (1971) reported a significant positive correlation between one of their measures of participants' "habit strength" and the extent of the interference that participants experienced in a Stroop-style scanning task. They take this finding to suggest that the greater a participant's strength of habit to find words compared to their strength of habit to find colours, the more susceptible they are to interference in the Stroop-

like colour-and-word scanning task. This individual difference measure of habit strength resonates nicely with our idea that participants habitually and automatically plan complementary responses, and that this tendency is linked to interference in the Stroop task.

The judgment response time data is different than the other measures because the patterns are close to perfect mirrors in one memory response type compared to the other. When the planned memory response type does not require a translation (arrow keys), we see almost the same interaction as we see in the other measures: strong interference when the judgment response type contradicts the stimuli (arrow keys) and much less interference when they are complementary (letter keys). However, when the planned memory response type requires a translation, this measure diverges from the others. Rather than the sizes of the interference effects of the two response types becoming less distinct from one another, as is the case in the other dependent variables, the data looks like a flipped version of the expected pattern, now with the judgment response type which requires a translation (arrow keys) being less susceptible to interference than the one not requiring a translation (letter keys). Across both halves of the data, interference occurs in all cases: there is interference when no translations occur, when only a memory translation occurs, when only a judgment translation occurs and also when both items are translated. The most interference occurs when only one response requires a translation, and interference is actually weaker in this measure when a translation must occur for both responses. This indicates that interference is not just caused by the necessity of enacting a translation, but specifically that two items inhabiting the same parallel processing system causes additional interference.

We could suppose that there are many layers of interference to be observed within this version of the task, perhaps indicating that it was an oversight to expect that our method in Experiment 2 would rule out one hypothesis if evidence was gleaned to support the other. It seems feasible therefore, that there may be both interference due to enhanced encoding and interference due to the nature of response plan, which co-exist. Undeniably, this potential interplay of interferences is an interesting topic of study and has important repercussions for the translational model and our understanding of how Stroop-like interference works when one source of interference is remembered rather than perceived. However, it would be

useful to also have a clear understanding of this paradigm without this extra factor of nature of the recall response plan interfering with results and making the picture more complicated. Therefore, we were primarily interested next to see what the effect on judgment measures would be if participants could *not* plan the exact nature of their recall response. Would we still see a translation model-like pattern and provide evidence in support of the enhancing encoding theory, thus building a multi-layer picture of Stroop-like interference in the working memory Stroop task? To address this, the two recall response types were interspersed within blocks in Experiment 3 so that participants could not anticipate which response type they would have to make until the time at which they were prompted to make it.

The fact that the judgment response time data is always affected by congruence regardless of the response types used may reflect that the other measures are less sensitive to the effects in which we are interested and justify our decision to focus primarily on the judgment response time measure. As discussed earlier, a perfect adherence to the predictions set out by the translational model would require that no interference occurs when no translation is required, but in the judgment response time data, congruence is a significant effect at every conjunctive level of both response types. A finding similar to this was also observed in our Experiment 1 and the simultaneous Stroop paradigm using these stimuli and response types (Delooze & Morey, 2024, also Chapter 2 of this thesis), however, this was not found in Virzi and Egeth's (1985) results using a vocal response. With this in mind, Experiment 3 was also conducted partly to determine whether a 'purer' verbal response, speaking aloud, would give a more complete recreation of the pattern expected by the translational model. If this prediction were to be confirmed, it would further elaborate on our understanding of the systems put forward in the translational model. It seems clear from our many experiments using this response method so far that a manual response *can* be mapped onto the verbal system. However, perhaps this mapping is imperfect and may still require some small amount of translation. From the quite different pictures in judgment response time data compared to the other measures, which all conformed to the expectations of the translational model (when no translation was required for the opposite response, e.g., if concerned with memory accuracy, this is the case when judgment response type is NESW), we are

also interested in finding out where this extra step of translation is observable – i.e., perhaps this only manifests in judgment response times.

Experiment 3

Following the findings from Experiment 2 which confirmed our suspicion that the effect of planned response type would mediate the extent of Stroop-like interference even in a delayed Stroop task, we decided to assess whether our stand-in verbal response is in fact as strong as the traditionally used spoken verbal response. Recent evidence using a colour Stroop task from Augustinova et al. (2019) assessed the extent of interference in vocal and manual responses. They found stronger interference and facilitation (concepts that are combined in our experiments) using vocal responses than keypress responses (keys marked with coloured stickers) when focusing on font colour and ignoring word meaning. Using a more varied battery of Stroop variations than will be applied here, they attempted to break down this difference into the relative contributions of task, response and semantic conflicts, finding evidence to support the notion that different combinations of these forces are at play in different circumstances. This idea of layers, or different sources, of Stroop interference resonates with one of our motivations for running this third experiment: a desire to understand the baser mechanisms in play during the working memory Stroop task using stripped back conditions (at least compared to Experiment 2). Their work utilised the simultaneous colour-word Stroop format, and only tested interference of word meaning on colour naming, so the aims of our current experiment are quite different, but still this work provides a good basis on which to suppose that there might be many layers of Stroop interference and that the pattern of data collected through vocal responses in our paradigm might be different than that observed in our Experiment 2 with verbal-manual responses.

To briefly describe the method of Experiment 3, letter focus trials were again included in a version of Experiment 2 (maintaining the recall version of the memory task) wherein instead of a letter key (NESW) response, a microphone was employed to collect vocal responses. If a vocal response is perfectly mapped to the verbal processing system, we should see perfect adherence to the patterns laid out in Virzi and Egeth's translational model (and demonstrated in their data, Virzi & Egeth, 1985). This will be satisfied if no interference from the location item held in memory

is detected when participants judge letters using a vocal response. If such a pattern is still not observed and instead this experiment replicates the pattern of results from Experiment 2, it could reflect that this method (including the use of the letter keys as a verbal response) is sensitive to detecting effects which have not been previously observed in spatial Stroop studies. Additionally, we were interested in whether the ability to plan the memory response would mediate this interaction, or whether knowing that it would need to be recalled would be sufficient to elicit the expected pattern of results, in-line with the enhanced encoding theory put forward in Experiment 2.

Method

The major methodological changes from Experiment 2 are the removal of the NESW key response type for both judgment and memory responses, which has been replaced with vocal responses collected with a microphone, and the interspersal of the different memory response types within blocks. These changes were made to address whether a vocal response would elicit a stronger interaction between response type and congruence, and to eliminate the possibility for participants to plan the exact nature of the memory response they would enact.

Participants

Forty-four students enrolled at Cardiff University elected to take part. In this experiment, the majority of participants were recruited in the summer break via advertisements on social media and in building-wide newsletters. When the target sample size could not be reached in this way, data collection continued into the academic year using the undergraduate Psychology cohort and recruiting opportunistically via the Experiment Management System. Participants were reimbursed £7 for their time, or accepted course credit, depending on the time period in which they took part. Again, no demographic data was collected, which limits the conclusions which could be drawn from this data regarding generalizability. However, the student population from which participants were taken was mostly female, mostly aged 18-24 years, and all were fluent speakers of English, as is required of those participating in an undergraduate course in the medium of English.

Materials and Apparatus

The experimental stimuli and experimental program remained the same compared to Experiment 2. In Experiment 3, only one USB NUM pad was provided to participants, the one marked with the arrow keys. Additionally, participants spoke into a tabletop microphone (either a PreSonus M7 or an AKG C1000S) to collect vocal responses. Two additional images were used within the experiment, one was a simple line art image of a microphone, which appeared on-screen whenever participants needed to give a vocal response, and the other was a simple depiction of arrow keys which appeared on-screen whenever participants needed to give an arrow key response. One or the other of these icons appeared in the top right corner for the judgment phase and centrally for the memory phase.

Design

The experiment was again a 2 (Judgment Response Type: vocal/arrow key) x 2 (Memory Response Type: vocal/arrow key) x 2 (Congruence: congruent/incongruent) within-subjects design, but this time, the only independent variable which was blocked was the Judgment Response Type. All blocks contained congruent and incongruent trials, and trials which required a vocal and an arrow key memory response. The dependent variables of interest were the accuracy (%) of responses in both the judgment and memory phases, and the time taken to make each correct response. For arrow key responses, this was measured as the time taken to press the correct key, and for vocal responses, this was measured as the length of the whole correct utterance. A simple proxy for this was to use the duration of the voice clips which contained correct responses. Participants were asked to press the enter key as soon as they had finished speaking to submit their vocal responses, and in the vast majority of cases, this was accomplished correctly, meaning that these could be used as acceptable response time measures. One of the many reasons which motivated the use of the NESW key response type in the first place was the increased ease with which their response times could theoretically be compared to arrow key presses. The task of pressing an arrow key should be much faster than the task of speaking a letter and then the additional step of pressing the enter key (Virzi & Egeth, 1985 found that vocal responses were significantly slower than manual responses in their Experiment 2), thus we do not expect that our response time results will exactly recreate previous results in the

regard that, even when the vocal response is complementary and the trial is congruent, we still expect that the response will take longer than arrow key responses. There is no reason to suspect that these increased response times should negatively impact the expression of the predicted null effect of congruence on response times when response type is complementary to the stimulus. So, while the responses' relative positions on the response time graphs may be different, we still expect to see the variance in the key element which is the extent of interference.

Procedure

The procedure was the same as Experiment 2 with a few exceptions. First chronologically is the removal of the familiarisation task at the beginning of the experiment, as no NESW key response was ever required. The next difference is that, in place of NESW key responses, participants were required to give vocal responses by speaking one letter at a time into the microphone, followed by pressing the enter key on the NUM pad in front of them to log their response. The final difference is in the mixing of memory response type within blocks, which resulted in participants being unable to wholly plan their memory response until the time at which they needed to enact it. This study was approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Participants were excluded from analyses based on failing to meet the following criteria: scored 66% accuracy or more overall, scored 6 or more correct in each trial type (e.g., vocal judgment response, vocal memory response, incongruent). This resulted in the exclusion of eight participants, leaving a sample size of 36. Individual trials were excluded if participants responded more quickly than 0.2s, which is reasoned to be an error rather than a true response. To maximise data collection, there was no time limit imposed by the program for participants' responses. Of all correct responses, only the extremely small proportions of 0.4% (for judgment responses) and 0.5% (for memory responses) exceeded 5s in duration, and the longest correct response took only 8.68s. Therefore, no responses were excluded based on excessive length. For accuracy analyses, the data was arcsine square root transformed to attempt to counter skew as a result of a ceiling

effect. For analysis of response time data, incorrect responses were excluded, and all included data points were log transformed to better comply with ANOVA requirements. Since the judgment responses occurred before participants had any knowledge of the memory response type they would enact on that trial and they should have no way to intuit it, it was not feasible that the variable of memory response type would have any influence on this data, and therefore only 2-way ANOVAs were conducted. For the memory measures, it was feasible that there may be a carryover effect of the judgment response type which had just occurred, so 3-way ANOVAs were conducted, as in previous experiments.

Judgment Response Times

A 2-way ANOVA was conducted on the judgment response time data (because memory response type was not yet known and could not have had a systematic influence on the judgment performance), which revealed that the interaction between the factors of judgment response type and congruence reached significance ($F(1,35)=53.019$, $p<.001$, $\eta_p^2=.602$), with, as predicted, the vocal response being less susceptible to interference than the arrow key response which requires translation. This interaction is demonstrated in Figure 12 below. Again, further ANOVAs were conducted to more closely examine the interaction, which revealed that the effect of congruence was significant in the arrow key response data ($F(1,35)=76.159$, $p<.001$, $\eta_p^2=.685$) where translation occurred, but also in the vocal response data ($F(1,35)=4.197$, $p=.048$, $\eta_p^2=.107$) wherein no translation occurred. This does not provide perfect support for our suggestion that the strength of the mapping from verbal processing to letter key response output may be weaker than the mapping to the spoken response output, because here we see the same unexpected effect of interference when no translation is required as we did when participants used a manual verbal response. The data adheres to the expected pattern of results imperfectly: a very small but still significant effect of congruence when the response type and stimulus type are complementary.

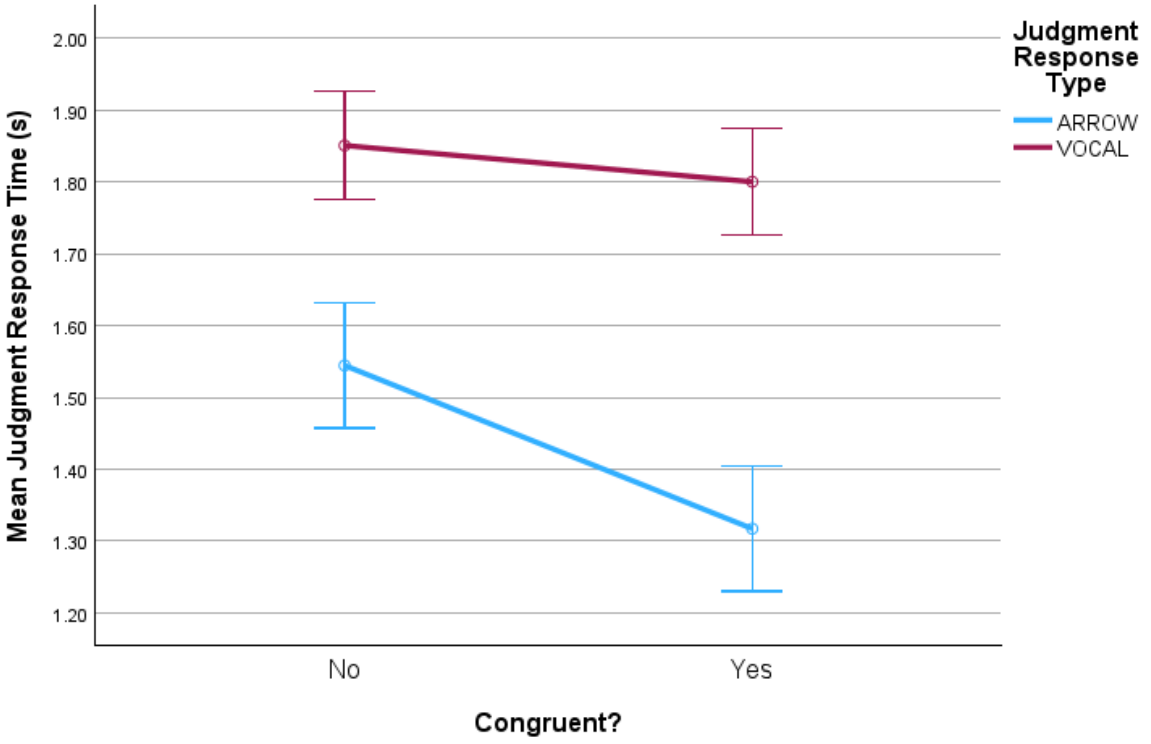

Figure 12 shows mean response time in seconds for the letter judgment task as a function of congruence (x-axis) and judgment response type (legend). The error bars represent +/- 1 standard error.

The analysis also revealed two significant main effects: judgment response type ($F(1,35)=54.115$, $p<.001$, $\eta_p^2=.607$) which reflects considerably slower vocal responses than arrow key responses overall, and congruence ($F(1,35)=64.456$, $p<.001$, $\eta_p^2=.648$) which reflects faster responses to congruent than incongruent trials overall.

Judgment Accuracy

A 2-way ANOVA conducted on the accuracy data for judgment responses revealed a significant interaction between judgment response type and congruence ($F(1,35)=11.122$, $p=.002$, $\eta_p^2=.241$), which aligns with the prediction set forward by the translational model, since when a translation is required in the arrow key response trials, participants are more susceptible to interference than when no translation is required and a vocal response is given. This is demonstrated in Figure 13 below. To verify whether participants are just less susceptible to or not at all influenced by congruence in the non-translation trials, further ANOVAs were conducted. These tests showed that where a translation was required in the arrow key response data, there was a significant effect of congruence ($F(1,35)=16.798$,

p<.001, η_p^2=.324), and in the vocal response data, where no translation was required, there was no significant effect of congruence (F(1,35)=.020, p>.05).

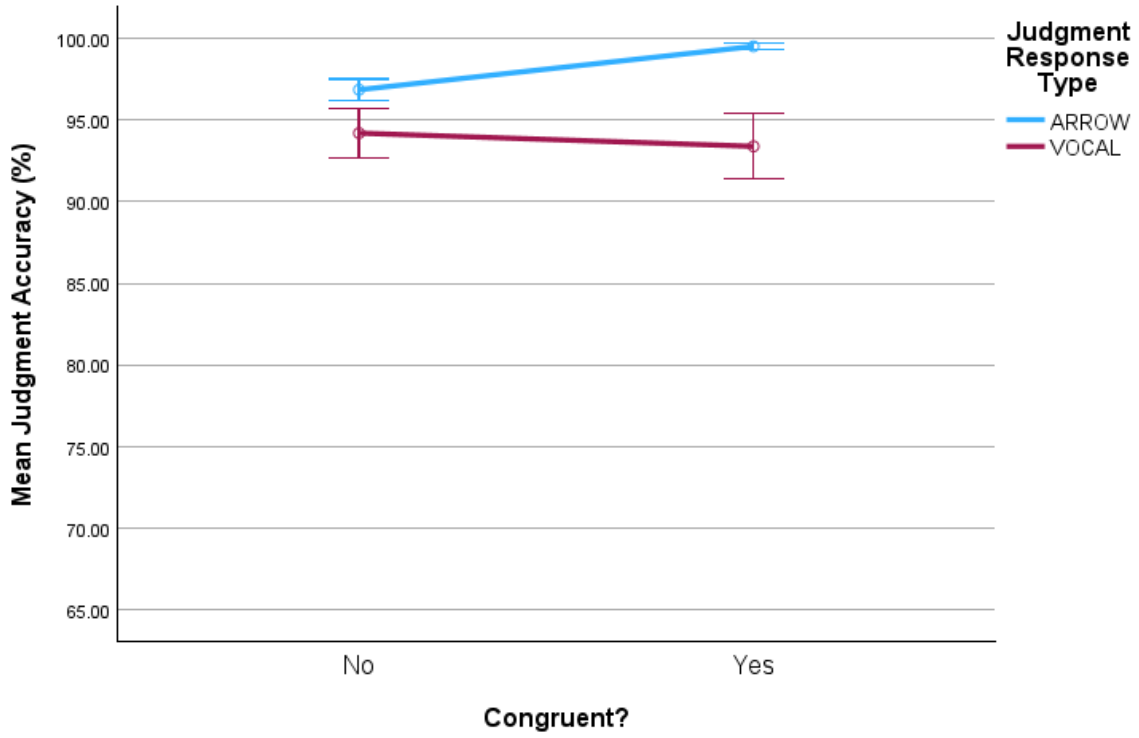

Figure 13 shows mean accuracy as a percentage for the letter judgment task as a function of congruence (x-axis) and judgment response type (legend). The error bars represent +/- 1 standard error.

The analysis also revealed a significant main effect of congruence (F(1,35)=8.055, p=.008, η_p^2=.187) reflecting better accuracy on congruent trials, and a significant main effect of judgment response type (F(1,35)=5.247, p=.028, η_p^2=.130).

Memory Response Times

A 3-way ANOVA conducted on the response time data for correct memory responses does not suggest an interaction between the factors of memory response type and congruence. See Figure 14 below for a demonstration of this. However, it did reveal two significant main effects of these factors: memory response type (F(1,35)=1159.841, p<.001, η_p^2=.971) with arrow key responses being radically faster than vocal responses, and congruence (F(1,35)=36.642, p<.001, η_p^2=.511) with congruent trials eliciting faster responses overall. This suggests that both congruence of stimuli and the enacted memory response type affect the time taken

for participants to respond, but not in a way that one factor's influence depends on the other factor.

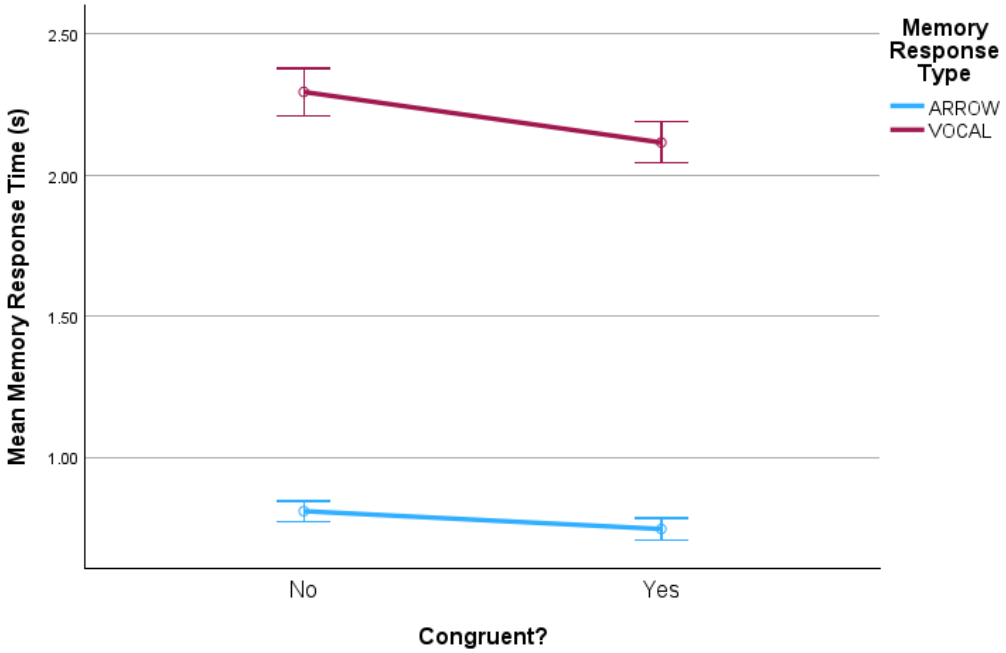

Figure 14 shows mean response time in seconds for the location recall task as a function of congruence (x-axis) and memory response type (legend). The error bars represent +/- 1 standard error.

Judgment response type alone was not a significant factor influencing memory response times, but the interaction between judgment response type and memory response type was significant ($F(1,35)=24.648$, $p<.001$, $\eta_p^2=.413$), reflecting a marginal speed benefit when participants responded to the memory probe using the same method as which they used to respond to the judgment item (a non-switch benefit). Finally, the interaction between judgment response type and congruence reached significance ($F(1,35)=5.713$, $p=.022$, $\eta_p^2=.140$), which is not explicitly predicted by the translational model but does not seem to be incompatible. This interaction, illustrated in Figure 15 below, appears to reveal an exacerbatory effect of translation during the judgment phase: when participants had to translate the to-be-judged letters into arrow key responses, they were then more affected by incongruence in their memory responses compared to when they had to respond to the letters with the complementary response of speaking them aloud (though even the non-translation judgment response elicited some effect of congruence on the

memory response time: $F(1,35)=7.690$, $p=.009$, $\eta_p^2=.180$). No other interactions were significant.

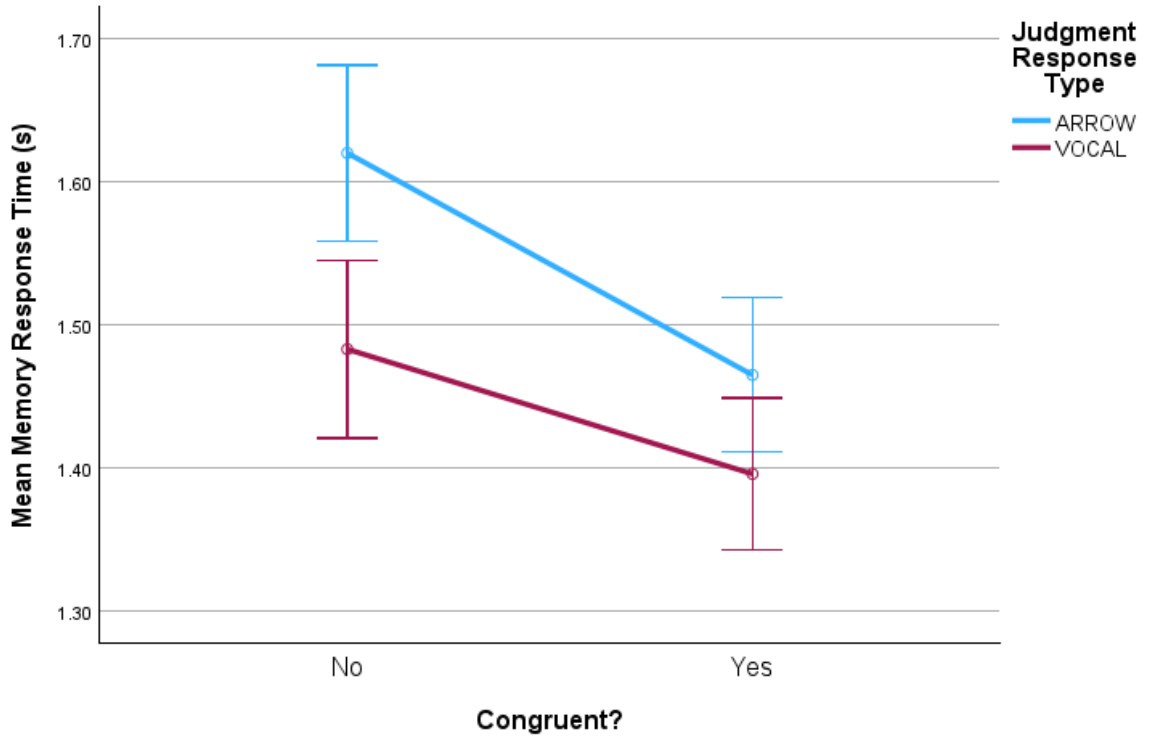

Figure 15 shows mean response time in seconds for the location recall task as a function of congruence (x-axis) and the preceding judgment response type (legend). The error bars represent +/- 1 standard error.

Memory Accuracy

A 3-way ANOVA conducted on the memory accuracy scores revealed that one interaction reached significance, which is the memory response type by congruence interaction ($F(1,35)=9.036$, $p=.005$, $\eta_p^2=.205$), illustrated in Figure 16 below. This pattern is predicted by the translational model, as the recalled stimuli are locations, meaning that the spatial response of arrow key presses should be less vulnerable to interference than the non-complementary vocal response which requires a translation to be made. Further analyses into this interaction revealed a significant effect of congruence within the vocal response data only ($F(1,35)=6.804$, $p=.013$, $\eta_p^2=.163$), but no significant effect within the arrow key response data only ($F(1,35)=1.885$, $p>.05$). This aligns perfectly with the predictions made by the translational model, which suggests that there should be no effect of incongruence in the response type which best complements the concerned stimulus. It is interesting that the preceding judgment response type had no effect on memory accuracy in this

version of the experiment, whereas it did interact with memory response type and congruence in Experiment 2. There were no other significant main effects on the data overall.

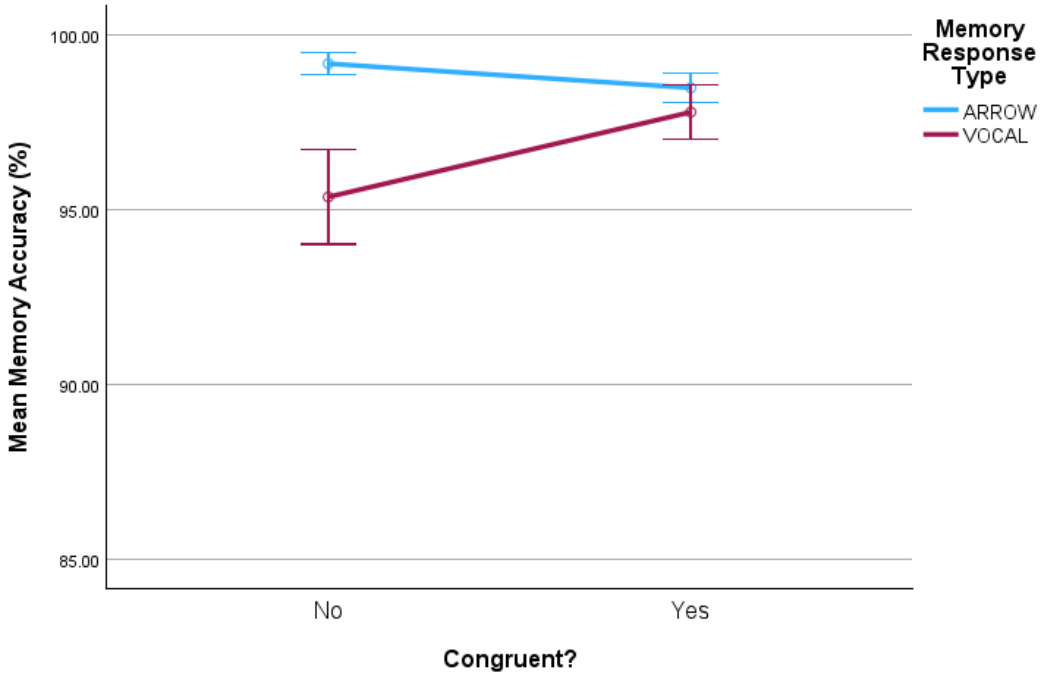

Figure 16 shows mean accuracy as a percentage for the location recall task as a function of congruence (x-axis) and memory response type (legend). The error bars represent +/- 1 standard error.

Discussion

To recap the major finding of our third experiment, three of the four dependent variables were differently susceptible to the effects of interference according to the task-relevant response type (memory response time was the exception). These patterns of data conform to the translational model's predictions that no (or in some cases only significantly less) Stroop-like interference occurs when the nature of the response is complementary to the nature of the stimuli, compared to when the nature of the response is contradictory to the nature of the stimuli. The data here firmly demonstrate that even when participants cannot plan the exact nature of the recall response they must enact, just the intention to recall the information enhances its ability to elicit Stroop-like interference on the interleaved judgment task compared to a recognition response (as in Experiment 1). This finding supports our notion that there are layers to the Stroop-like interference which occurred in Experiment 2 and that the complex pattern of results observed there reflect an influence of the nature

of the planned memory response, but do not necessarily rule out the enhanced encoding theory, which is supported by these results.

Our theories surrounding recall versus recognition were mostly focused on the effect that the intention to recall would have on the judgment responses, but it is a boon of this experimental design that we can look on the reciprocal effect of judgment type on memory. The only significant influence on the interference observed in memory accuracy was the memory response type, which, as we would expect given the translational model, reflects the occurrence of interference only when the recall response type required a translation (vocal response), and not when it did not require one (arrow keys). This did not interact further with the judgment response type on the trial, so demonstrated no carryover effects: memory accuracy was affected by Stroop-like interference to the same extent regardless of the nature of the preceding judgment response. The story with memory response times was a little less clear: while memory response type and congruence between the stimuli were both significant influences of response time to the memory question, they did not meaningfully interact. This means that the extent of interference observed was the same regardless of which response type participants needed to use to recall the item, vocal or keypress. Contrastingly, the effect of the preceding judgment response type *did* have a role to play in this measure. This is a carryover effect: the necessity to make a translation in the judgment response exacerbated the effect of congruence in the memory judgment, compared to when no translation was necessary in the preceding judgment task. So, when participants translated letters into arrow key responses during the judgment phase, they were more likely to be slowed by an incongruent stimulus in either recall response method.

We next address our secondary experiment aim, which was to determine whether a vocal response would demonstrate a purer adherence to the predictions of the translational model than the ‘stand-in’ manual verbal response of letter keypresses which had fallen short of this previously. The data here do not support this suggestion: as in Experiment 2 using our ‘stand-in’ verbal response, we still do not see perfect adherence to the assumptions of the translational model in the judgment response time measure. Again in Experiment 3, the data analysed here reveal a very small but significant interference effect when the nature of the judgment response complements the nature of the stimuli (vocal response to letter

stimuli). This finding contradicts Virzi and Egeth's finding of a null effect of congruence in their "vocal-response-to-meaning condition" (Virzi & Egeth, 1985, p. 7). A conservative reading of the translational model suggests that this circumstance should not elicit Stroop-like interference because no translation is necessary. The failure of Experiment 3's data to conform to the pattern laid out by the translational model provides tentative evidence against our expectation that vocal responses are more strongly mapped to the verbal system than the letter keypress response type that we used in previous experiments, and supports the continued use of the letter key response for more convenient data collection in this paradigm.

To reflect on what these findings mean, the non-conformity of the judgment response time measure to the trend of complementary response type (vocal) eliciting no interference could be a reflection of the true state of the world. With sample sizes of 43 and 36 (Experiments 2 and 3, respectively), the two experiments reported here which show this finding are better powered than Virzi and Egeth's (1985) study sample of 24 - perhaps it was due to chance that they did not detect this small effect. The methodological and contextual differences from Virzi and Egeth's finding to ours should also not be understated. This divergence from expectations could be due to the working memory Stroop hybrid nature of our current method, which is undeniably different from the simultaneous Stroop method which Virzi and Egeth used. Or it could be related to our sample – today's undergraduate participants are more able on keyboards through increased computer and smartphone exposure compared to the average undergraduate participant in 1985. Perhaps the reality is that both response types are sufficiently well-mapped to the verbal processing system in our modern participants, and that response time data is a special case wherein with a high-powered method, interference can be detected which is not observable in the other measures considered here. This conjecture is supported by the fact that the letter judgment response time data in Chapter 2's simultaneous Stroop method also showed this pattern and that it is seen in response times for judgments of locations in this chapter's Experiment 1. Further, to properly compare the NESW key response with the vocal response, it would have been more useful to maintain the exact method with only that isolated change from Experiment 2, and to run an additional experiment to assess the effects of foreknowledge of recall response type. Though it was our intuition that Experiment 3's design was cleaner and thus we would likely

see the pattern emerge perfectly here if it were going to at all in this paradigm, perhaps there is something which is still causing a base level of interference which can only be observed in the judgment response time measure.

As expected, in general, vocal responses were slower than arrow key responses. An unexpected quirk of the data in this experiment is that that vocal judgment responses were also significantly less accurate than arrow key judgment responses. This is anticipated in the memory responses, which are always concerning locations and thus require a translation to be output as vocal responses, but in the judgment task, which is always concerning letters, we did not anticipate an effect in this direction. We suggest that this very small effect may be explained by the nature of the measurement rather than the response: a small proportion of vocal trials from otherwise attentive participants were un-transcribable due to difficulties with recordings. Participants were asked to press the enter key to submit their vocal responses only after completing annunciation of the word, but on a very small number of trials they pressed the key too early and too much of their response was lost to determine which letter they intended to say. Contrastingly, it was impossible to lose individual responses in the arrow key trials, which may explain why this response type is more accurate, despite requiring a translation. Detecting an effect as small as this is further evidence that the study was well-powered and gives weight to our other unanticipated findings.

In this experiment, we have focused on a very small part of the predictions laid out in the translational model, largely with the goal of unpicking whether there was fault with the stand-in ‘verbal’ response which we have used throughout these ‘compass task’ experiments. From the results of this final experiment which replicated the same small but significant interference effect even when judging letters with a complementary vocal response, we now believe that there is no such fault. The five experiments reported over these two chapters all provide support for the translational model (less so in this chapter’s Experiment 1, but firmly in the others), so this close dissection is by no means convincing to us that the theory is not useful in understanding Stroop interference. Our aim in conducting these experiments was to provide evidence which may facilitate elaboration of the model, not to disprove it.

General Discussion

The aims of Experiment 1 were: first, to re-purpose Kiyonaga and Egner's (2014) working memory Stroop method and assess whether we would corroborate the finding of an effect of congruence in a spatial Stroop variant of the paradigm. Second, we intended to add to their body of studies assessing whether well-established effects within the simultaneous Stroop literature manifest within the paradigm. The effect we were interested in was the pattern characteristic of the translational model, which posits that Stroop interference is caused by the necessity to translate a stimulus of one type into another domain for the response which is intended. The pattern of interest consists of significantly less (or in a strict interpretation, *no*) interference when the response method is complementary in nature to the stimulus it relates to, compared to a strong effect of interference when these factors contradict (when a translation is required). The data succeeded in providing support for the first of these aims, but evidence was piecemeal for the second hypothesis, with performance from trials wherein participants judged locations adhering to the expected pattern, but performance from letter-judging trials deviating largely from those expectations.

Experiment 2's goal was then to determine whether a small alteration to the method would be all that was required to tease out a more complete recreation of the translational model data pattern, and if the change were successful, to attempt to distinguish between two possible mechanisms which may be responsible. Focusing now on only the location-memory-letter-judgment trials which failed to reveal a translational interaction in Experiment 1, participants were asked at the end of each trial to recall the locations that they committed to memory, instead of being shown a probe to recognise. The method with which participants recalled the location was varied by block, with participants either recalling with a stimulus-complementary arrow key, or a translation-necessitating letter key (letters N, E, S and W representing the cardinal directions). We suggested that, if the necessity to recall were to successfully elicit the pattern we sought because of enhanced encoding of the memory item, that patterns of data would not vary as a result of the type of planned recall response. Contrastingly, if the nature of the planned response was integral, we expected that differences in these data patterns would be detected. This experiment confirmed that data patterns differed when the planned (memory) or

preceding (judgment) task required a translation compared to when it did not. The patterns of data reported in every measure were a good indication that the nature of the response mattered, but signs pointed to a layered effect of interference in this data, and we suspected that our contrasting theories of mechanism of effect may in fact be working in tandem.

So, to determine whether enhanced encoding associated with recall rather than recognition acted independently of the nature of the planned recall response, Experiment 3 was conducted, wherein recall responses still varied, but participants were only cued as to which response type would be required at the time of recalling. In this way, enhanced encoding could theoretically be enacted without any further interference from the nature of the response. This experiment had a secondary aim as well: to assess whether spoken responses instead of letter keypress responses would satisfy more fully the hypotheses laid out by a strict interpretation of the translational model. Specifically, we were looking to find no response slowing when the judgment response did not require a translation. In response to the first experimental aim, the data from Experiment 3 supported the notion that enhanced encoding was partially responsible for the manifestation of the expected pattern when recall is planned. Even when participants could not fully plan their recall response, the interference which was detected in judgment responses was still influenced by the judgment response type used (with much greater interference when a translation occurred because the response type did not complement the stimulus). Regarding the second aim, the data revealed that the judgment response times still showed a very small interference effect when no translation was necessary. This finding suggests that previous failures of the letter key response method to perfectly adhere to the expectations set out in the translational model (Virzi & Egeth, 1985) are not due to its inadequacy as a “verbal” response. We think it is more likely that these results are a reflection of these experiments’ stronger power to detect small effects than those conducted by Virzi and Egeth (1985).

The experiments reported here provide support for the notion put forward by Kiyonaga and Egner (2014) that holding an item in working memory elicits Stroop-like interference. Stroop-like interference was detected in every measure in every experiment detailed here, providing very firm support for their basic conclusion in the sub-domain of a verbal-spatial Stroop effect. Regarding their larger conclusion that

this working memory Stroop interference is enacted in the same way as seeing an item in real-time, some solid support for this is also found here. Having established that data collected using our compass-style spatial Stroop stimuli confirm Virzi and Egeth's (1985) translational model hypotheses when the verbal and spatial stimuli are presented simultaneously (Delooze & Morey, 2024, also Chapter 2 of this thesis), we reasoned that if indeed this working memory Stroop interference demonstrated by Kiyonaga and Egner (2014) worked in the same way as simultaneous Stroop interference, then using the same stimuli and response methods, we should also observe these patterns in their working memory Stroop style. While we were unable to wholly replicate the pattern using their original paradigm involving recognition as the memory task, some elements of the data did line up as expected. With some further probing, we found that by merely changing the nature of their memory task so that recall instead of recognition memory was required, the expected phenomenon was observed very clearly and comparably to more traditional simultaneous Stroop methodologies. This corroborates their claim of the equivalence of Stroop interference elicited by items held in working memory and by items seen in real time.

Our simplest takeaway from these experiments, and one that bodes well for convenient usage of this paradigm going forward, is that a spoken response is not obviously a purer verbal response than a letter keypress. The pattern of data was the same slightly imperfect adherence to the translational model's predictions both when participants used letter keys and when spoken responses were required. A strict interpretation of the translational model suggests that when no translation is required for a stimulus to be output, no interference should occur as a result of an incongruent accompanying stimulus. This was confirmed across many dependent variables in the experiments reported here, but not in the judgment response time measure, where a very small interference effect was detected. Even when participants spoke their answers aloud (which is the verbal response type that Virzi and Egeth used in their experiment), we still see some interference in the judgment response time data for the non-translational trials, mirroring the findings from, and suggesting that the issue is not inherent to the nature of, the letter keypress response. In light of the multiple experimental findings which support a slightly less strict interpretation of the predictions, we do not consider this small divergence to be condemning evidence

against the utility of the theory. Our takeaway about this matter is that it reflects a target for amendment within the theory, but does not detract from its explanatory power, as the experiments reported here have provided a lot of support for the viability of this model.

In Experiment 2's judgment response time data, we observed lower interference when both tasks required a translation than when only one task did. Thus, interference is strongest when both items occupy the same processing system, rather than being an additive result of translation, which would see interference being strongest when two translations occur than one. This supports Virzi and Egeth's notion of competition: "It is hypothesized that this interference occurs at the decision stage of the system used to respond. Two codes are in competition: the translated code to which the subject must ultimately respond and the code for the irrelevant dimension that has arrived directly from the analyzing stage of the system used to respond . . ." (Virzi & Egeth, 1985, p. 7). It is noteworthy that this finding is only observed in the judgment response time data – all other measures from Experiment 2 see much more similar levels of interference in either relevant task response type when the opposite task requires a translation. This indicates that the judgment response time measure is unique in some way, perhaps due to increased sensitivity compared to the other measures.

The experiments reported here aimed to assess the similarity between Stroop interference elicited by simultaneously presented information and Stroop interference caused by items held in working memory. In doing so, we have also drawn conclusions that reflect on the differences between recognition and recall memory, and about the mapping of response types onto processing systems within the translational model. Our conclusion is that items held in working memory do have the capacity to elicit Stroop interference in the same way as currently viewed stimuli, and that the working memory Stroop task will prove to be a useful tool for investigating the Stroop effect.

References

Augustinova, M., Parris, B. A., & Ferrand, L. (2019). The Loci of Stroop Interference and Facilitation Effects With Manual and Vocal Responses. *Frontiers in Psychology, 10,* 1786. https://doi.org/10.3389/fpsyg.2019.01786

Campo, P., Poch, C., Parmentier, F. B., Moratti, S., Elsley, J. V., Castellanos, N. P., ... & Maestú, F. (2010). Oscillatory activity in prefrontal and posterior regions during implicit letter-location binding. *Neuroimage, 49*(3), 2807-2815. https://doi.org/10.1016/j.neuroimage.2009.10.024

Carey, S. T., & Lockhart, R. S. (1973). Encoding differences in recognition and recall. *Memory & Cognition, 1,* 297-300. https://doi.org/10.3758/BF03198112

Chmiel, N. (1984). Phonological recoding for reading: The effect of concurrent articulation in a Stroop task. *British Journal of Psychology, 75,* 213-220. https://doi.org/10.1111/j.2044-8295.1984.tb01894.x

Delooze, M. A., Langerock, N., Macy, R., Vergauwe, E., & Morey, C. C. (2022). Encode a letter and get its location for free? Assessing incidental binding of verbal and spatial features. *Brain Sciences*, *12*(6), 685. https://doi.org/10.3390/brainsci12060685

Delooze, M. A. & Morey, C. C. (2024). *The Compass Task: A New Direction for the Spatial Stroop Paradigm.* OSF.io. https://osf.io/wv9rh

DeSoto, M. C., Fabiani, M., Geary, D. C., & Gratton, G. (2001). When in doubt, do it both ways: brain evidence of the simultaneous activation of conflicting motor responses in a spatial stroop task. *Journal of Cognitive Neuroscience*, *13*(4), 523-536. https://doi.org/10.1162/08989290152001934

Dyer, F.N. (1971). The duration of word meaning responses: Stroop interference for different preexposures of the word. *Psychonomic Science, 25,* 229–231. https://doi.org/10.3758/BF03329102

Dyer, F. N. (1972). Latencies for movement naming with congruent and incongruent word stimuli. *Perception and Psychophysics, 11,* 377-380. https://doi.org/10.3758/BF03206271

Dyer, F. N. (1974). Stroop interference with long preexposures of the word: Comparison of pure and mixed preexposure sequences. *Bulletin of the Psychonomic Society, 3*, 8-10. https://doi.org/10.3758/BF03333373

Dyer, F. N, & Severance, L. J. (1972). Effects of irrelevant colors on reading of color names: A controlled replication of the "reversed Stroop" effect. *Psychonomic Science, 28*, 336-338. https://doi.org/10.3758/BF03328756

Elsley, J. V. & Parmentier, F. B. R. (2015). Rapid Communication: The asymmetry and temporal dynamics of incidental letter–location bindings in working memory. *The Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://doi.org/10.1080/17470218.2014.982137

Fagioli, S., Hommel, B., & Schubotz, R. I. (2007). Intentional control of attention: Action planning primes action-related stimulus dimensions. *Psychological research*, *71*, 22-29. https://doi.org/10.1007/s00426-005-0033-3

Glaser, M. O, & Glaser, W R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 875-894. https://doi.org/10.1037/0096-1523.8.6.875

Gumenik, W E., & Glass, R. (1970). Effects of reducing the readability of the words in the Stroop Color-Word Test. *Psychonomic Science, 20,* 247-248. https://doi.org/10.3758/BF03329047

Hall, J. W., Grossman, L. R, & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4* (5), 507-513. https://doi.org/10.3758/BF03213211

Henderson, M. M., Rademaker, R. L., & Serences, J. T. (2022). Flexible utilization of spatial- and motor- based codes for the storage of visuo-spatial information. *eLife, 11,* e75688. https://doi.org/10.7554/eLife.75688

Heuer, A. & Schubö, A. (2017). Selective weighting of action-related feature dimensions in visual working memory. *Psychonomic Bulletin & Review, 24,* 1129-1134. https://doi.org/10.3758/s13423-016-1209-0

IBM Corp. Released 2023. *IBM SPSS Statistics for Windows,* Version 29.0.2.0 Armonk, NY: IBM Corp

Kiyonaga, A., & Egner, T. (2014). The working memory Stroop effect: When internal representations clash with external stimuli. *Psychological science*, *25*(8), 1619-1629. https://doi.org/10.1177/0956797614536739

Klein, G. S. (1964). Semantic Power Measured through the Interference of Words with Color-Naming. *The American Journal of Psychology, 77*(4), 576-588. https://doi.org/10.2307/1420768

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

MacLeod, C. M. (2005). "The Stroop task in cognitive research," *in Cognitive Methods and Their Application to Clinical Research*, eds A. Wenzel, and D. C. Rubin, (Washington, DC: American Psychological Association), 17–40. https://doi.org/10.1037/10870-002

Martin, M. (1981). Reverse Stroop effect with concurrent tasks. *Bulletin of the Psychonomic Society, 17,* 8-9. https://doi.org/10.3758/BF03333650

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from **https://office.microsoft.com/excel**

Nealis, P. M. (1974). Reversal of Stroop test: Interference in word reading. *Perceptual and Motor Skills, 38,* 379-382. https://doi.org/10.2466/pms.1974.38.2.379

Pavlovia, https://pavlovia.org Open Science Tools, Nottingham, UK.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods, 51,* 195-203. https://doi.org/10.3758/s13428-018-01193-y

Postman, L., Jenkins, W. 0, & Postman, D. L. (1948). An Experimental Comparison of Active Recall and Recognition. *The American Journal of Psychology, 61*(4), 511-519. https://doi.org/10.2307/1418315

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shor, R. E. (1970). The processing of conceptual information on spatial directions from pictorial and linguistic symbols. *Acta Psychologica*, *32*, 346-365. https://doi.org/10.1016/0001-6918(70)90109-5

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. https://doi.org/10.1037/h0054651

Tversky, B. (1973). Encoding Processes in Recognition and Recall. *Cognitive Psychology, 5,* 275-287. https://doi.org/10.1016/0010-0285(73)90037-6

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is Working Memory Storage Intrinsically Domain-Specific? *Journal of Experimental Psychology. General*, *148*(11), 2027–2057. https://doi.org/10.1037/xge0000566

Uleman, J. S. & Reeves, J. (1971). A reversal of the Stroop interference effect, through scanning. *Perception & Psychophysics, 9*, 293-295. https://doi.org/10.3758/BF03212651

Virzi, R. A. & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13(4),* 304-319. https://doi.org/10.3758/BF03202499

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H. et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4(43),* 1686, URL https://doi.org/10.21105/joss.01686 .

Zimmer, H. D., Speiser, H. R., & Seidler, B. (2003). Spatio-temporal working-memory and short-term object-location tasks use different memory mechanisms. *Acta Psychologica*, *114*(1), 41-65. https://doi.org/10.1016/S0001-6918(03)00049-0

4. Rapid Source Forgetting Across Modalities: A Problem for Working Memory Models

Introduction

Working memory is the cognitive system which allows us to store and process a limited amount of information necessary to carry out a wide variety of complex acts (Cowan, 2017). Given the capacity limitation of working memory, which is assumed by most models, forgetting, at least temporarily, is vital. For instance, without forgetting, the mind would quickly become overwhelmed and unable to focus on the information most relevant for our current goal. Some pieces of information must be discarded. Therefore, a large part of understanding memory is understanding the circumstances under which we do not remember: forgetting.

Various models have differing approaches to explaining the flow of information into and out of working memory, including how and when forgetting occurs. The Multicomponent model of working memory (originally presented in Baddeley et al., 1974, but see Baddeley, Hitch & Allen, 2021 for an updated overview) suggests that information is lost from modality-relevant temporary storage systems when we try to exceed their limited storage capacity. The Time-Based Resource-Sharing (TBRS) model (originally presented in Barrouillet, Bernardin & Camos, 2004, but see Barrouillet & Camos, 2021 for an updated overview) is one of many models historically which outlines that forgetting occurs as a result of time-based decay, wherein the probability to recall an item is reduced as a function of time passing (for another example of a decay and rehearsal account, see also Baddeley et al.'s, 1975 account of the Phonological Loop). In the TBRS specifically, this decay occurs only when attention is directed away from the target item. Somewhat similarly, the Embedded Processes model (originally presented in Cowan, 1988, but see Cowan, Morey, & Naveh-Benjamin, 2021 for an updated overview) also states that items may be lost from passive short-term storage through time-based decay, or alternatively by interference from a similar subsequently encoded item. In their Interference model of working memory (Oberauer & Lin, 2017;

2023; see also Oberauer, 2021), Oberauer and Lin consider forgetting to be solely a result of interference: this occurs when the target memory representation is not selected for recall due to competing activation of non-target representations with similar or overlapping context retrieval cues. This theory of forgetting therefore relies on the target having a similar context to the non-targets which are recalled in its place. Popov and Reder's (2020) Resource-Depletion theory of working memory states that we have a limited pool from which to draw resources for cognitive processing and memory encoding. Each processing or encoding action depletes this pool until insufficient resources are available to encode items so that they can be recalled later. Therefore, Popov and Reder (2020) propose that limits in working memory arise at encoding: once the encoding resource has been depleted by encoding some information, further information cannot be encoded until the resource has had time to recover.

Forgetting is especially fascinating for cases wherein intuitively, we would firmly expect to remember. Discrepancies exist in estimates for the maximum duration of working memory persistence, with some sources suggesting that items can endure up to 30s before being transferred to long-term storage (Atkinson & Shiffrin, 1971), others suggesting that the vast majority of items are lost by 18 seconds (Peterson & Peterson, 1959), and still others suggesting that the life of a working memory representation could be as short as four seconds (Sligte, Scholte & Lamme, 2008). Despite these differences, it is safe to say that most researchers would not expect attended information to be lost within one second. These entrenched expectations mean recent findings concerning the phenomena of rapid forgetting known as '*Attribute Amnesia'* are particularly problematic for working memory models. Chen and Wyble (2015; 2016) demonstrated attribute amnesia, the apparent forgetting of features less than one second after they had certainly been attended. In their paradigms, participants very briefly saw an array of colored characters, with the task to find the letter among the numbers and were only asked to report the target's location. After many such trials, Chen and Wyble surprised participants by asking them about the identity or the color of the target, and participants responded poorly on these surprise tests. This finding is particularly surprising because the participants must have attended to the identity of the target to

be able to identify it as the letter among numbers, yet they seem to be very quickly unable to recall which letter it was.

Chen, Carlson, and Wyble (2018) extended this phenomenon to source memory using a variant of the paradigm in which participants were repeatedly asked to give a congruency judgment based on two temporally spaced (their Experiment 2) color-word features: a color word presented in black font, either followed or preceded by a color patch. Here, both items which are presented on a trial have both a 'source' (format: written word or colored square) and a semantic meaning (the color which is represented). Participants completed this congruence task with ease, but when prompted in a surprise trial to choose the color patch they just saw, they could not reliably recall the color that they used to form their judgment (Experiment 2), nor correctly attribute a probe color to its feature source (Experiment 3). Not only were participants unlikely to choose the correct color patch, but they were just as likely to choose the color patch consistent with the color word they had seen. This confirms some intact memory of the recent experience, but loss of key contextual information which would allow the source of a feature to be identified. That is to say that they seem to have intact item memory in that they can recall the semantic representations of the two colors which were presented (which was necessary for the pre-surprise trial task), but no source memory containing information about the format in which each item was presented, hence the term 'source amnesia'.

Curiously, in Chen et al.'s Experiment 2 (on which the current studies are based), this chance-level performance is only witnessed when the item which is probed for recall during the surprise test was the item which was presented second. Participants are much more successful at choosing the color when the probed color patch was presented first. This could be taken to reflect that source information is simply better maintained for the first-presented item than the second-presented item because it must be represented strongly enough to persist until the second item is processed to achieve the task goal. Alternatively, Chen et al. argued that this could be attributed to a sort of primacy effect bias, wherein, in the absence of knowledge concerning the sources of the two semantic color items which are held in memory, the semantic representation of the first-presented item is chosen for recall more often than the second-presented item.

None of the working memory models described above handle this result elegantly. It is difficult for TBRS to explain this finding, since TBRS stipulates that forgetting occurs because attention is occupied with something else across a period during which the forgotten information temporally decays; in this paradigm, attention may no longer be focused on the forgotten feature, but it is lost almost instantly. A further issue this finding poses for decay-based theories is that more time has passed since the first-presented item was encoded, yet this item seemingly remains accessible, or is perhaps prioritized. The Embedded Processes model also outlines that information to which attention is paid should be easily accessible for a short time, before time-based decay can act upon it. Therefore, even if the color is not the most highly activated feature when it is probed, because it has been attended so recently, it should be accessible from activated long-term memory. Possibly, making the congruence judgment and/or interpreting the surprise question degrades the representation of the color, either through time-based decay due to the delay, or through interference of new information, but this again does not account for why the most recently presented item is lost while the first-presented item is preserved (also, see the work by O'Donnell & Wyble, 2023, supporting the idea that attribute amnesia is not solely caused by interference from a surprise question).

The Resource-Depletion theory seems to partially account for the findings of this paradigm, given its strength in explaining the commonly observed primacy effect. However, a limitation on how much can be encoded (Popov & Reder, 2020) does not seem relevant in this paradigm, because so little information is presented for evaluation in the first place: we expect that this model would predict a working memory capacity much greater than one item (as in Popov, So & Reder, 2022; Popov, 2023). Similarly, because the Multicomponent model allows for verbal and visual features to be stored in separate buffers, which would be capable of representing at least one feature at a time, it would not obviously predict that this source information would be lost so quickly and with no competition from more recently presented items. On the subject of competition, the Interference model also seems like it would struggle to explain this loss, as the two 'contexts' (here we call them 'sources' or 'formats') of written word and color patch seem sufficiently distinct to not be cross-activated and cause interference.

Models allowing for removal of information from working memory (e.g., Lewis-Peacock et al., 2018; Oberauer, 2021) may handle these findings marginally more successfully because they include a mechanism, removal, that not only emphasizes the most relevant information in mind but eliminates the no-longer-needed information. Applied here, because the second-presented feature becomes irrelevant for the expected test as soon as a congruency judgment is reached, the detail could be removed from working memory and forgotten. However, under this logic it remains unclear why participants selectively retain the information which was presented first, as the first-presented feature becomes just as irrelevant to the goal.

Given the major challenge that Chen et al.'s (2018) findings pose for working memory, this phenomenon is important to replicate and to understand more fully before theorists consider whether to adapt their models in response. A gap in the Chen et al. (2018) studies is that they did not test participants' memory for the verbal information contributing to the congruency judgments. Such an experiment could speak to the generalizability of the effect, which will be important for theorists to take into consideration. Additionally, the results in all of their studies were consistent with the conclusion that source amnesia may not mean that the color is not represented: consistently, observed errors were misattributions in which participants' choice was consistent with the *word* stimulus that was presented on that trial. These misattributions could indicate, as Chen et al. suggested, that the first-presented feature is more strongly biased for recall, but these findings could also reflect that the verbal feature is more strongly activated, and thus more likely to be selected in surprise tests when the other feature is forgotten.

With the high prevalence of misattributions, which are instances of to-be-ignored information encroaching on target information, it may be useful to draw more explicit parallels between Chen et al.'s paradigm and Stroop interference. Classic Stroop interference occurs when participants struggle to inhibit particularly salient and automatic word-reading tendencies during a color-naming task. In Stroop's (1935) original study, Stroop interference only occurred naturally in this one direction: words interfered with responses to ink color, but not the reverse. Stroop found that participants required considerable training to develop their color naming skills and inhibition of word-reading impulses to a sufficient extent to be able to elicit a 'reverse Stroop effect' wherein performance in a word reading task was impaired by

incongruent text color. This asymmetry of interference is not seen in all variations of the Stroop task: verbal-spatial Stroop tasks, for instance, elicit both the regular (verbal interference on spatial processing) and the reverse (spatial interference on verbal processing) Stroop effect without extensive training (e.g., Virzi & Egeth, 1985), seemingly belying a different relationship between these types of information than between color and word information. It seems that when it comes to interference, a color-word pairing creates quite a unique disparity. This difference in vulnerability to interference suggests that the read word might be more highly activated than the color patch. Drawing a parallel between these two tasks, we suggest that it is possible that the read word would be less susceptible to loss in Chen et al.'s paradigm, whether it is in the first or second position. If greater source amnesia is observed in recall of color information than word information, it would be necessary for models seeking to explain this rapid forgetting to additionally distinguish between the persistence of verbal and visual features somehow.

The working memory models reviewed earlier do not account for the rapid forgetting observed by Chen et al., so it is understandable that they do not necessarily offer explicit insight into what would happen if word, rather than color, were probed in a surprise test. However, using the general assumptions made by each model, we can make suggestions about what potential findings would align with each model. For instance, because the Multicomponent model explicitly distinguishes between verbal and visuospatial storage, we reason that it could predict differential source-related forgetting for visual versus verbal information, due to the different mechanisms and capacities of the different slave systems involved in rehearsing and maintaining information of different types. If word information is not forgotten but color information is, then the Multicomponent model might account for that by expanding on its presumed differences in the durability of representation in these separate, domain-specific stores. Similarly, the TBRS model specifically includes a uniquely verbal memory mechanism, in addition to the domain-general one. Therefore, we expect that TBRS could account for better recall of verbal than for visual source information by appealing to domain-specific resources that are uniquely available for verbal materials. Contrastingly, the Embedded Processes model, the Resource-Depletion theory and Oberauer and Lin's framework are domain-general in nature, and thus they should not predict a discrepancy between

observed source amnesia for verbal or visual information, because the mechanism by which forgetting occurs does not act differently depending on information type. However, it remains the case that, if we observe rapid forgetting of either feature as Chen et al. (2018) observed with color, all models should consider how to explicitly account for those findings.

Here, we address this gap in our knowledge with three experiments: in Experiment 1, we replicated Chen et al.'s Experiment 2 to establish that our method was in line with theirs, in our Experiment 2, we extended the method to test memory for the verbal stimuli, and finally, in Experiment 3, we explored the idea that participants might be encoding a different kind of source than has previously been tested for. Briefly, this method consists of several pre-surprise trials requiring the participant to make a judgment on whether the presented color patch and color word are congruent or incongruent (see Figure 1 for an illustration). These are followed by a surprise trial wherein the participant is instead asked to report the identity of the color patch (our Experiment 1), the word (our Experiment 2), or the first and second items (our Experiment 3) which they were just shown. In line with previous findings, in Experiment 1 below, we expect to find above chance surprise trial performance when the color patch, which is probed, was presented first in the trial, but chance-level performance when it was presented second in the trial.

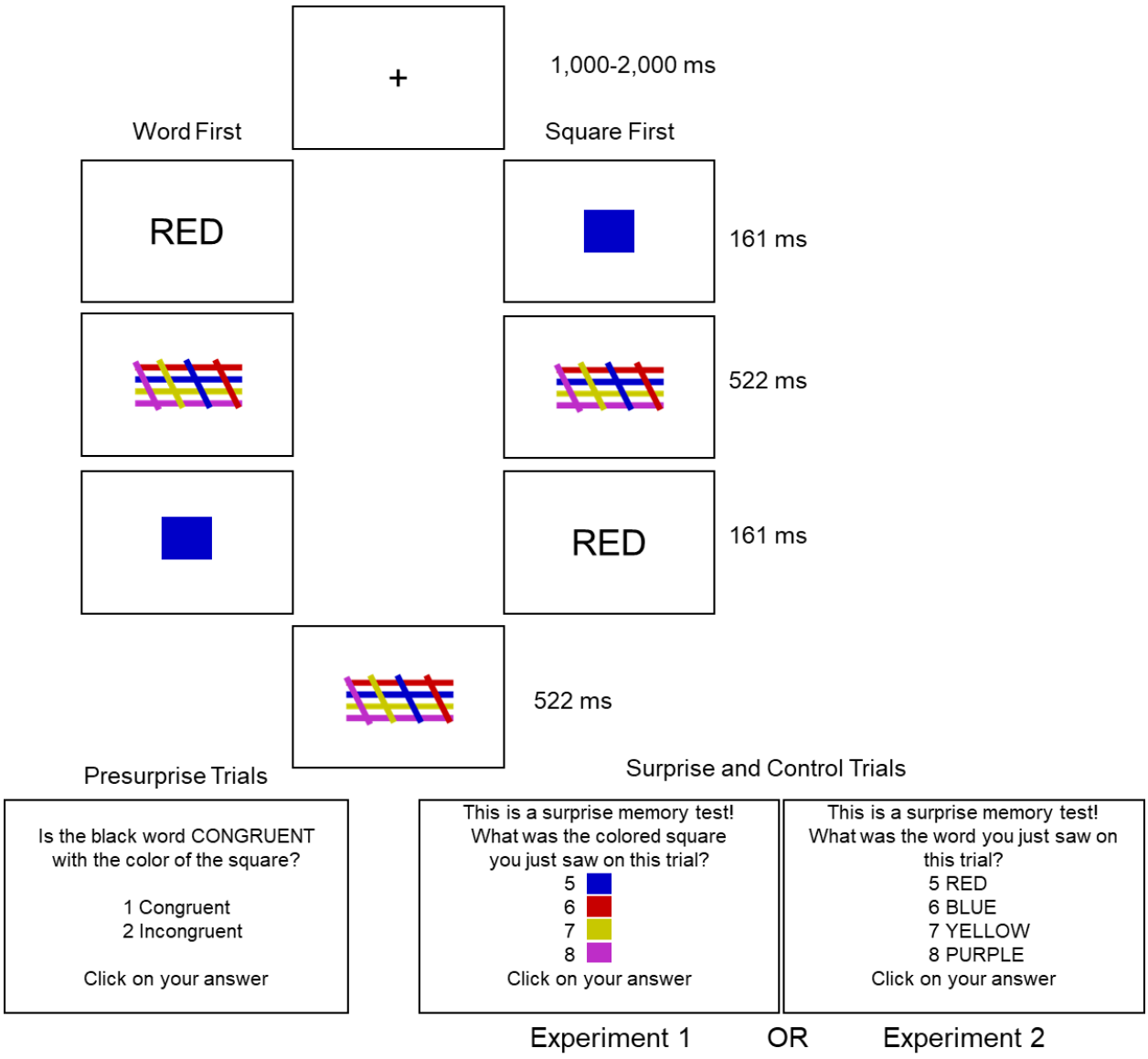

Figure 1 shows an illustration of the procedure in Experiment 1 and Experiment 2.

Experiment 1

Method

In Experiment 1, after many pre-surprise trials consisting of color-word-color-patch congruency judgments, participants were expecting to take part in another congruency test, but instead received an unexpected color memory test. In Chen et al.'s study, participants responded with number keys mapped to response options, whereas in our experiment, participants responded with the mouse by clicking on their chosen answer (both in the congruency judgment pre-surprise trials and in the

surprise trial). This adjustment was made in response to the notion that it may be more straightforward for participants.

Sample size. We selected a sample size of 20 participants for each condition or group in all experiments, aligning with the sample size used by Chen et al. (2018). This decision was made to ensure reliable estimates across our experiments and to guarantee at least an equivalent number of observations compared to those reported in previous experiments.

Participants

In all our experiments, our participants were volunteers recruited via the online data collection agency, Prolific (https://www.prolific.co/). Recruiting via Prolific has been shown to produce comparable data quality in terms of engagement to recruiting university students (whether they take part online or in a lab; Uittenhove, et al., 2023). To participate in our study participants had to meet the following eligibility criteria: (1) native speaker of English, (2) British, American, or Canadian nationality and country of birth, (3) normal or corrected-to-normal vision, (4) no cognitive impairment or dementia, (5) normal color vision, (6) no language-related disorders, (6) aged between 18 and 30 years old at the time of sign-up, and (7) with an approval rating of at least 90% on prior submissions at Prolific. All participants were paid £9 per hour (prorated) for their participation in all experiments, which was approved by Cardiff University's School of Psychology Ethics Committee.

One participant was excluded from analysis due to attaining a pre-surprise trial accuracy of less than 60%. The average age of the participants was 26.5 years (*SD* = 3.01, range 20–31); 46 self-identified as female, 29 as male, three responded that their gender was best represented by the category "other", and one preferred not to specify their gender.

Materials

All experiments were conducted using the online programming software PsyToolkit (Stoet, 2010, 2017). The stimulus design was based on Chen et al.'s Experiment 2 (2018). The verbal stimuli consisted of four different color words displayed in uppercase letters: RED, BLUE, YELLOW, and PURPLE. Verbal stimuli were presented in black, uppercase, 30-point Arial font at the center of the computer

screen on a gray background (RGB values: 150, 150, 150), unless otherwise specified.

Participants were also presented with colored squares measuring 50 pixels by 50 pixels, each displayed in one of four colors: red (RGB values: 200, 0, 0), blue (RGB values: 0, 0, 200), yellow (RGB values: 200, 200, 0), and purple (RGB values: 190, 45, 200). The colored mask was an arrangement of four horizontal lines in each of the four colors, intersected by four diagonal color lines of each of the four colors. The materials and the program are available at the Open Science Framework page associated with this article (https://osf.io/mkwb2/) and the materials described here can be seen illustrated in Figure 1.

Design

The independent variables were as follows: Surprise Trial Congruence (Congruent or Incongruent) and First Stimulus (Word-First or Square-First). The dependent variable was accuracy of color recall, measured using a mouse click. There were 4 groups of 20 participants. Each group was randomly allocated to one of the 4 conditions: word-first congruent surprise test, word-first incongruent surprise test, square-first congruent surprise test, square-first incongruent surprise test (see Figure 1).

Procedure

Each participant took part in a single online experimental session lasting approximately 5 minutes. The procedure (see Figure 1) was based on Chen et al.'s Experiment 2 (2018) with the following modifications. Each trial began with a variable fixation cross lasting between 1,000 ms and 2,000 ms, immediately followed by the presentation of either the word or the color square (depending on the assigned group) for 161 ms. Subsequently, a mask was presented for 522 ms, followed by the second stimulus (word or color square) for 161 ms. Another mask was then displayed for 522 ms before the test phase.

Before the experiment, participants completed two congruency trials (1 congruent, 1 incongruent) as practice trials in which they received feedback, either 'The answer was: Congruent' or 'The answer was: Incongruent'. Feedback was not given during the following pre-surprise trials to ensure consistency with Chen et al.'s (2018) methodology. Participants then completed 48 pre-surprise trials of the same

structure (24 congruent trials, 24 incongruent trials) with an equal number of trials per color arrangements presented in a random order for each participant. In the pre-surprise trials, participants completed a congruency test, wherein they had to click with their mouse to indicate whether the meaning of the color word presented in black matched the color of the square they saw by clicking on either "congruent" or "incongruent".

These were followed by 1 surprise trial which was manipulated to be congruent or incongruent, followed by a further 4 control trials which were randomly selected to be congruent or incongruent. For the surprise and control trials, participants were presented with the following message during the color test: “This is a surprise memory test! What was the colored square you just saw on this trial?” This was followed by the congruency test as they had experienced previously. The order in which the colored squares were displayed during the test phase was randomized. For all tests (congruency and color), participants had up to 1 minute to make their decision. After completing all the trials, participants were asked if they had anticipated the surprise memory test: “Were you expecting the surprise memory test where we inquire about the colored square you recently viewed?”, to which they again responded with the mouse by clicking “Yes” or “No”.

Results and Discussion

In the pre-surprise trials, participants took a mean average of 770.881 ms (SD=1910.536 ms) to respond across all trials. Participants tended to be very accurate in the pre-surprise with a mean score of 45.911 (SD=6.611) out of a maximum total of 48, meaning that the error rate was 4.352%. Participants took understandably longer to respond to the color surprise trials which required new instructions to be read and processed. Here, they had a mean average response time of 5099.987 ms (SD= 3291.218 ms). In the control trials following the surprise trial, wherein participants likely knew that they would need to recall the identity of the colored square, their error rate was 6.013%.

The key comparison for these data is between the incongruent surprise trial error rates and chance performance. These were calculated by dividing the number of participants who made errors by the total number of participants who took part in each surprise trial type. Chen et al. report 60% and 15% error in their word-first and

square-first groups, respectively. In this experiment, our data very closely replicate the findings of Chen et al.'s Experiment 2, with an identical error rate in the word-first and a very similar rate in the square-first trials.

Inferential Analysis

To compare these results to chance, a Chi-Squared Goodness of Fit test was conducted, which found that the Incongruent Square-First results did differ significantly from chance ($c^2(1) = 42.123$, $p < .001$), but the Word-First results did not significantly differ from chance ($c^2(1) = 2.400$, $p > .05$). These inferential results suggest that when the probed item was presented first, its source was remembered, whereas when the probed item was presented second, source information was lost. We decided it would be useful to run these analyses again using participants' performance on the first control trial as the expected data spread to give a more complete picture of the surprise performance, as was done by Chen et al. (2018). A Chi-Squared Goodness of Fit test was conducted, which found that the Incongruent Square-First results did not differ significantly from performance on the first control trial ($c^2(1) = 1.056$, $p > .05$), but the Word-First results did significantly differ from the first control trial ($c^2(1) = 127.368$, $p < .001$). These findings replicate those by Chen et al. (2018) and support that a mouse response is suitable for probing this phenomenon. See Table 1 below for a comparison.

Error Rates	Chen et al.		Current Study	
	Congruent	Incongruent	Congruent	Incongruent
Word-First	N/A	60%	5%	60%
Square-First	N/A	15%	15%	10.5%

Table 1 shows a comparison of the error data from Chen et al.'s Experiment 2 and the current study's error data.

To address the question of misattributions, the number of errors in which the incorrect answer given matched the untested information type for that trial was divided by the total number of errors. Since misattributions were only possible in Incongruent surprise trials, these are the only trials for which data is shown. Our results replicate Chen et al.'s (2018) finding that most errors in the word-first trials were misattributions, but this was the case in much fewer of the errors in the square-first trials. See Table 2 below for a comparison.

Trial Type	Chen et al.		Current Study	
	Errors	Misattributions	Errors	Misattributions
Word-First	60%	40%	60%	50%
Square-First	15%	15%	10.5%	0%

Table 2 shows a comparison of the misattribution data from Chen et al. (2018)'s Experiment 2 and the current study's misattribution data.

These results firmly support the finding from Chen et al. (2018) that source amnesia occurs to a much greater extent when the to-be-recalled item is presented second in a given trial. Our data additionally support their conclusion that misattribution errors attributed to source amnesia are common in this paradigm. This successful replication of previous findings speaks to the robustness of the phenomenon.

Experiment 2

Having established that the source amnesia results for color memory can be replicated, we used the surprise trial in Experiment 2 to instead test participants' memory for word information. In Experiment 1, we asked participants only about the identity of the colored square, so the methodology used so far does not allow us to draw firm conclusions about whether the same pattern of forgetting and misattribution would be observed if memory for words was instead tested. The results in all of Chen et al.'s (2018) studies lead to the conclusion that misattribution is a major contribution to the poor performance thought to demonstrate source amnesia. In a control version of their Experiment 2, Chen et al. (2018) removed the response option which corresponded with the unprobed information on the surprise trial and found that participants' inaccuracy was greatly reduced (down to 10%). Our results from Experiment 1 support this idea, with a huge proportion of the errors made in the Word-First condition, where source amnesia is most common, being misattributions. Misattributions suggest that participants strongly remember the word and are sometimes biased to report it; but do they remember the word information so strongly to the point of commonly misattributing it only because it was presented first, or might they remember the word as strongly regardless of presentation order?

Briefly, as shown in Figure 1, the only difference in Experiment 2 compared to Experiment 1 was during the surprise trial, wherein participants recalled the identity of the word they were shown instead of the identity of the colored square. Replication of this result in a second domain would speak to the generalizability of the rapid forgetting phenomenon and strengthen the argument for theorists to address this it. We expected that word information might be better recalled than color information, given its special status in the Stroop paradigm, and the unique verbal memory mechanisms which are assumed in some working memory models. If memory for words is more persistent than memory for colors, potential explanations involving domain-specific mechanisms might gain support. However, if word information proves to be no better recalled than color patch information during the surprise trial, we would favor modifying domain-general accounts of working memory to account for rapid forgetting.

Method

Our Experiment 2 was identical to our Experiment 1 except for the surprise memory test in which we tested recall of the verbal (word) information instead of colors. This manipulation allowed us to investigate whether the pattern established by Chen et al. (2018) and confirmed in Experiment 1 also generalized to verbal information.

Participants

In Experiment 2, another group of participants who met the same eligibility criteria described in Experiment 1 and who had not taken part in the previous experiment were recruited from Prolific. Participants were assigned randomly to one of four conditions. Four participants (one in each condition) were excluded on the grounds of not meeting the 60% pre-surprise trial accuracy quota. The final sample was composed of 78 participants. The average age of the participants was 24.8 years (*SD* = 3.12, range 19–30); 29 self-identified as female, 48 as male, and one preferred not to specify their gender.

Materials, Design, and Procedure

The materials, design and procedure in Experiment 2 were identical to Experiment 1 except for the following changes. In Experiment 2, as shown in Figure

1, the surprise memory test was on verbal information. More exactly, participants were asked to click on which of the 4 words presented at test was the same as the word that they just saw on that trial (RED, BLUE, YELLOW, PURPLE)**.** The final question of the experiment was also adapted to reflect that procedural change: "Were you anticipating the surprise memory test where we inquire about the word you recently viewed?".

Results and Discussion

After exclusions based on poor pre-surprise trial accuracy, the mean average score in the pre-surprise trials across conditions was 46.231 (SD= 2.608) out of a total of 48 trials, meaning that the error rate was 3.685%. The mean average response time for these pre-surprise trials was 746.046 ms (SD= 1784.414 ms). Understandably, given the need to read and process new instructions, the surprise trial response time average of 5554.962 ms (SD= 4282.337 ms) was higher. In the control trials following the surprise trial, wherein participants likely knew that they would need to recall the identity of the color word, their error rate was 8.654%.

Again, the key comparison for these data is between the incongruent surprise trial error rates and chance. These were calculated by dividing the number of participants who made errors by the total number of participants who took part in each surprise trial type.

Inferential Analysis

To compare these results to chance, a Chi Squared Goodness of Fit test was conducted, which found that the Incongruent Word-First results did differ significantly from chance ($c^2(1) = 29.491$, $p < .001$), but the Square-First results did not significantly differ from chance ($c^2(1) = 1.067$, $p > .05$). When the word information which was probed for recognition was presented first, participants appeared to remember it. However, when this information was presented second, participants performed no better than they would if they were to guess. Again, we ran a second Chi Squared analysis on these data comparing participants' surprise trial performance to their performance in the first control trial. This analysis revealed that the Incongruent Word-First results did not differ significantly from the first control trial performance ($c^2(1) = 0$, $p > .05$), but the Incongruent Square-First results did

significantly differ from the first control trial ($c^2(1) = 67.222$, $p < .001$). See Table 3 below.

In this version of the experiment, we predicted that error rates, and therefore evidence of source amnesia, would be lower than in Experiment 1, due to the comparatively reduced capacity to induce Stroop interference that color information has compared to word information. This prediction is not supported by the results here, with a Chi Squared Goodness of Fit analysis suggesting that the error rates did not significantly differ across the two studies for the tested-item-first ($c^2(1) = 2.235$, $p > .05$) nor the tested-item-second ($c^2(1) = 0.875$, $p > .05$) condition. These results support the idea that this phenomenon is domain general: there is seemingly no difference in the extent of source amnesia when participants are tested on their ability to recall color patches or color words.

The existence of source amnesia that occurs so rapidly poses problems for most models of working memory, but the current results of the phenomenon occurring equally in a second domain lend stronger support to the domain-general models such as the Embedded Processes (Cowan, 1999), Resource-Depletion (Popov & Reder, 2020) and Interference models (Oberauer & Lin, 2017; 2023). Meanwhile, models which suggest that visual and verbal information are stored or maintained differently to each other may find this result more challenging.

	Error Rates	
	Congruent	Incongruent
Word-First	10.520%	21.053%
Square-First	10%	65%

Table 3 shows a comparison of the error data for both congruent and incongruent surprise trials when participants were asked to recall the word that they saw (Experiment 2).

Regarding misattributions, the number of errors in which the incorrect answer given matched the untested information type for that trial was divided by the total number of errors. Since misattributions were only possible in Incongruent surprise trials, these are the only trials for which data is shown. From these results, we can

conclude that misattributions appear to be roughly as prevalent in word recall as there are in color recall, especially when errors are common. See Table 4 below.

Trial Type	Errors	Misattributions
Word-First	21.053%	10.526%
Square-First	65%	50%

Table 4 shows a comparison of the misattribution data from word-first and square-first incongruent trials when participants were asked to recall the word that they saw (Experiment 2).

The misattributions seen in this paradigm may look on the surface to be comparable to the well-documented phenomenon of Stroop interference. First, the stimuli are color words and color squares which are very commonly-used Stroop paradigm stimuli; and second, when participants are asked to recall the color square, we sometimes see a bias towards instead recalling the content of the written word, which mirrors the Stroop effect of failure to inhibit word meaning when responding to visual color information. On the basis that participants struggle much more to inhibit interfering word stimuli during color naming than they do interfering color stimuli during word reading (Stroop, 1935), we hypothesised that misattribution errors might be less common in this paradigm when participants were asked to recall word information than when they were asked to recall color information. The results of Experiment 2 refute this idea, with the rates of errors and witnessed primacy effect being stable across both information types, leading us to conclude that it is unlikely that source amnesia occurs as a result of the same interference documented in Stroop effect research. It seems not to matter therefore which stimulus is causing interference toward the other. Instead, this finding supports Chen et al.'s contention that in this phenomenon, presentation order predicts which feature is dominant in memory: it is the first-encoded feature, regardless of its form. It is possible that these observed error rates will persist in any stimulus type which might be tested, though of course, further study would be required to say this with certainty.

Experiment 3

Following results from their Experiments 1 and 2 which could equally suggest failure to encode stimulus format as well as they suggest forgetting of stimulus format, Chen et al. (2018) conducted a third experiment. In the surprise trial of this

experiment, they showed participants a written word probe, the meaning of which aligned with the colored square which was presented first on that trial and asked them directly to choose whether the color represented by that word was presented in word or colored square format (thus the correct answer was always the “colored square” option). They found that participants were very poor at this explicit version of the task, performing very close to the fifty-fifty level expected by chance despite being always probed on the first-presented item, to which they responded accurately in the previous experiment. From this finding, Chen et al. thus concluded that in this paradigm: 1) The format in which the stimulus is presented is never encoded when it is not known to be needed; and 2) Participants are merely biased towards choosing the response which matches the semantic representation of the first-presented item. They suggest that this primacy bias is what leads them to do well in Experiment 2 when the item probed was presented first, and badly when the item probed was presented second.

If Chen et al.’s explanation is correct, and in this paradigm, participants are indeed entirely failing to encode an item’s source when it is not required for the task (though see Wyble et al., 2019 for a discussion on when this is not the case), this would be problematic for the Interference model (Oberauer & Lin, 2017; 2023), which emphasizes that context is the necessary cue which allows items to be recalled. We argue that it is safe to assume that participants can indeed recall the two items presented to them in the surprise trial, given the high prevalence of correct or misattribution answers observed. However, there is a possible alternative which we can see which might allow both Chen et al. and Oberauer and Lin’s suggestions to co-exist in harmony. It is possible that stimulus format is never encoded, but that a different type of context cue *is* encoded. When that context cue cannot be used to answer the surprise question, the primacy bias comes into effect. Commonly suggested types of ‘context’ are an item’s location and an item’s position in serial presentation order. Since all stimuli in this experiment are presented in the same location at the centre of the screen, it is unlikely that location context cues can be effectively used to distinguish them. On the other hand, the stimuli all necessarily have different positions in the serial order. This therefore could be the context cue by which participants are able to access item information in-line with the Interference model.

To test the suggestion that the context by which participants can recall item information in this paradigm is their serial order or position information, another variation of the previously used paradigm was created, with the pre-surprise trials remaining the same, but some key alterations made to the surprise and control trials. During the surprise and control trials, participants were asked which item was presented first and which item was presented second. According to Oberauer and Lin's model, if serial order is the context cue by which items in the source amnesia paradigm are encoded and retrieved, participants will respond correctly or will extrapolate semantically (choose the response option which aligns with the semantic color representation of the correct response, but in the other format) more frequently than they will misattribute (choose a response in either format which depicts the semantic color which they saw in the not-probed position) or be entirely wrong (guessing) because they have access to correct serial order information. A finding of chance-level performance in this task would be compromising for fundamental assumptions of the model, whereas evidence that participants succeeded in this task would provide very positive support making the Interference model the best contender among the working memory models considered here to explain the source amnesia phenomenon.

Method

Our Experiment 3 was identical to our Experiments 1 and 2 except for the surprise memory test. In this experiment's surprise memory test, we asked participants to recall which item was presented first and also which was presented second (order randomized) on that trial. Participants did this by clicking with their mouse on what they believed to be the correct color word or color square item (a total of eight response options instead of four as had been presented in previous experiments). This manipulation allowed us to investigate whether participants had access to a different kind of 'source' information than has been tested previously in this paradigm.

Participants

In Experiment 3, another group of participants who met the same eligibility criteria described in Experiment 1 and 2, and who had not taken part in the previous two experiments were recruited from Prolific. Participants were assigned randomly to

one of two conditions (their surprise trial was either square-first or word-first). For each condition, the presentation of the test order was counterbalanced across participants to control for order effects, but these were collapsed to form two groups of 80 participants each. Consequently, the sample was larger than in previous experiments. One participant (from the square-first condition) was excluded on the grounds of not meeting the 60% pre-surprise trial accuracy quota. The final sample was composed of 159 participants. The average age of the participants was 26 years (*SD* = 3.72, range 19–30); 92 self-identified as female, 61 as male, five as a different gender and one preferred not to specify their gender.

Materials, Design, and Procedure

The materials, design and procedure in Experiment 3 were identical to Experiments 1 and 2 except for the following changes. In Experiment 3, as shown in Figure 2, the surprise memory test asked participants to recall the first and second-presented items (order counterbalanced). More exactly, participants were asked to click on which of the 4 words and 4 colored squares presented at test were the same as the first and second items that they just saw on that trial (RED, BLUE, YELLOW, PURPLE)**.** The final question of the experiment was also adapted to reflect that procedural change: “Were you anticipating the surprise memory test where we inquire about the which one was presented first or second?”.

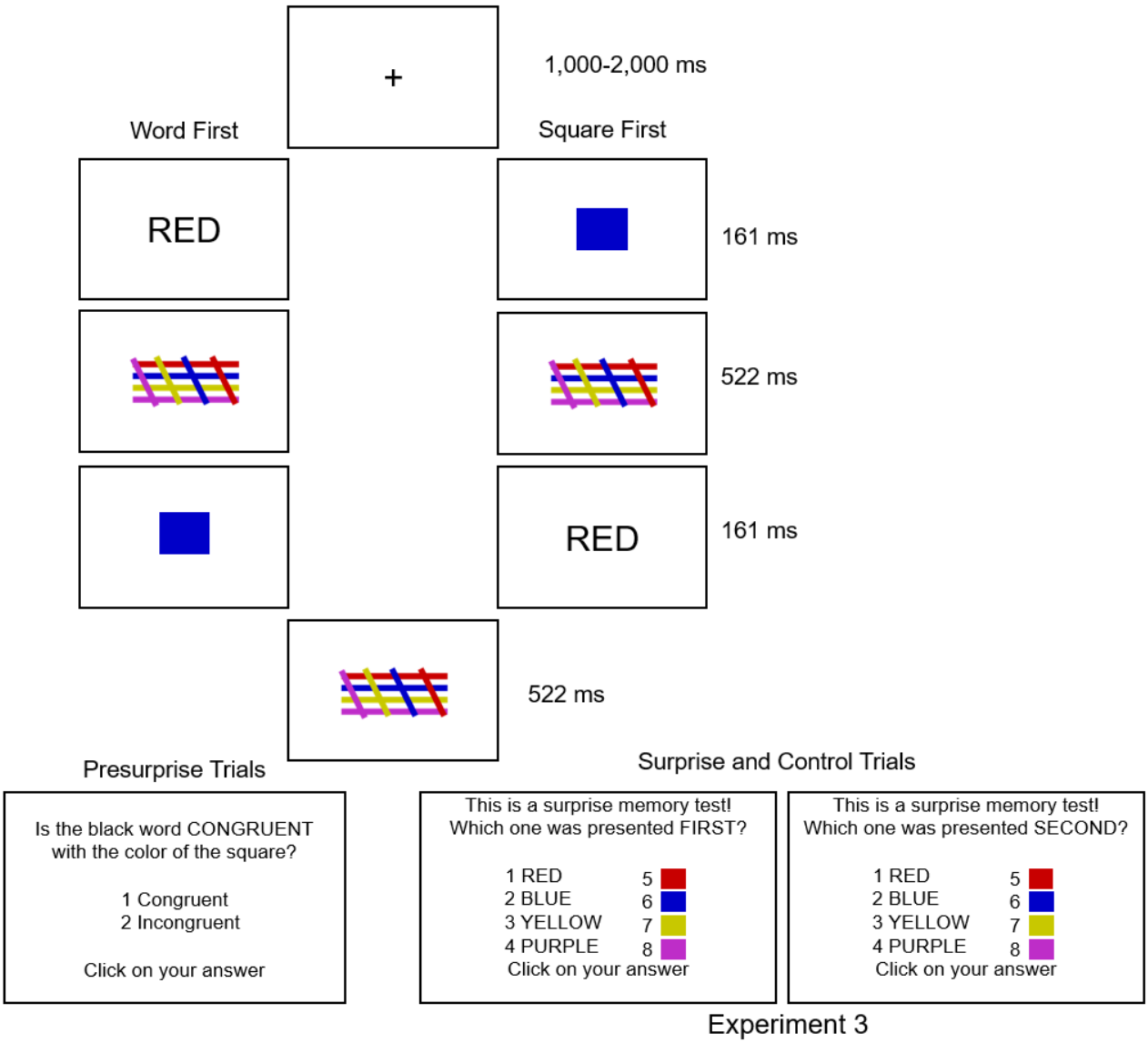

Figure 2 shows an illustration of the procedure in Experiment 3.

Results and Discussion

After exclusions based on poor pre-surprise trial accuracy, the mean average score in the pre-surprise trials across conditions was 46.396 (SD= 2.670) out of a total of 48 trials, meaning that the error rate was 3.342%. The mean average response time for these pre-surprise trials was 708.500 ms (SD= 1480.758 ms). Understandably, given the need to read and process new instructions, the surprise trial response time average of 7112.607 ms (SD= 6988.999 ms) was higher. In the control trials following the surprise trial, wherein participants likely knew that they would need to recall the positions of the stimuli, the percentage of participants selecting either the precisely correct or 'semantically correct' answers (answers

which had the same meaning as the precisely correct answer, but in the incorrect stimulus format) was 65.566% across all conditions and both positions. This demonstrates that participants could complete the task if they knew that they would be asked to do so.

	Percentage of response types							
	Testing Word				Testing Square			
	Correct	Semantically correct	Misattri-bution	Guess	Correct	Semantically correct	Misattri-bution	Guess
Word-First	12.500 %	13.750%	33.750 %	40%	17.500 %	6.250%	23.750 %	52.500 %
Square-First	7.595%	16.456%	32.911 %	43.038 %	13.924 %	16.456%	30.380 %	39.241 %

Table 5 shows a comparison of the proportions of responses for both word-first and square-first surprise trials when participants were asked to recall the first and second items that they saw (Experiment 3). "Correct" refers to responses which selected the same semantic meaning and stimulus format as was presented on that trial. "Semantically correct" refers to answers which had the same meaning as the precisely correct answer, but in the incorrect stimulus format (e.g., if correct response would be the blue square, the word BLUE was chosen instead). "Misattribution" refers to answers which corresponded to the non-probed item presented on that trial, regardless of stimulus format. "Guess" refers to answers which did not correspond with a stimulus presented on that trial, belying random guessing.

Inferential Analysis

Again, Chi-Squared analyses were conducted to compare the observed spreads of data for each condition to the spread which would be predicted by chance responding. The observed distribution of frequencies in the square-first group did not significantly differ from chance (for which the expected response proportions would be 12.5%, 12.5%, 25% and 50%, in-line with the order of response types in Table 5) regardless of whether they were tested on the identity of the first-presented item, the square, ($c^2(3) = 3.860$, $p > .05$), or the second-presented item, the word, ($c^2(3) = 5.250$, $p > .05$). It is the same situation in the word-first group: neither the results for the first-presented item, the word ($c^2(3) = 4.150$, $p > .05$), nor the second-presented item, the square ($c^2(3) = 4.250$, $p > .05$) differed significantly from chance. Additionally, we ran

Chi Squared analyses to compare surprise trial performance to performance on the first control trial. For the square-first group, these distributions differed for both the square ($c^2(3) = 71.130$, $p < .001$) and the word ($c^2(3) = 50.037$, $p < .001$). This pattern was the same for the word-first group for both the word ($c^2(3) = 120.864$, $p < .001$) and the square ($c^2(3) = 107.345$, $p < .001$).

Chance performance primarily indicates that participants did not know which stimulus was presented in which position during the surprise trial, refuting the hypothesis that serial order position information is being encoded in this paradigm. Even when expanding our definition of "correct" answers and taking semantic extrapolation responses into account, (where participants knew which color semantically was presented but selected the wrong format, e.g., *blue square* when the answer was "blue"), participants' performance is not indicative that they could use the correct serial order position cues to recall the items they saw. These results taken with the previous experiments reported here support Chen et al.'s (2018) notion that no feasible type of context or 'source' is encoded in this phenomenon. This is problematic for the Interference model (Oberauer & Lin, 2017; 2023) as discussed earlier, because without a linked context, the model predicts that items should not be accessible in working memory, but in some select instances (e.g., when the first item is probed by format in Experiments 1 and 2), the information is accessible.

A counter to this argument might be made in the form of the Interference model's Focus of Attention element, which is proposed to confuse the context-content links of items held within it at the same time (Oberauer & Lin, 2017). If the stimuli in this paradigm are thought to be held in the focus of attention simultaneously, their links would be confused regardless of which context type they consisted of, and they would not be expected to know which item was presented in which format (Experiments 1 and 2), nor in which position (Experiment 3). These findings therefore argue strongly for the inclusion of the focus of attention element in this model for maximum explanatory power. This is an important argument because the inclusion of this element of the model has previously been debated following mixed results from testing model fits (Oberauer & Lin, 2023). Alternatively, perhaps this finding warrants a clearer definition of what can and cannot be considered a 'context' in the model. For instance, could the stimuli in this paradigm be linked to the planned congruent/ incongruent response which participants intend to make about them?

It is additionally interesting that the results of this final experiment suggest a total loss of item information: participants select response options that correspond with one or the other of the presented stimuli just as often as they would if they were guessing. This was not predicted by either the Interference model (even with a focus of attention adjustment), nor the primacy bias suggestion made by Chen et al. (2018). If the stimuli are proposed to be held in the Interference model's focus of attention, they should be more or less guaranteed to be accessible on a semantic level. If participants are biased to report the first-presented item, regardless of source, why do they guess randomly in this instance? Further, this finding is in stark contrast to the previous results reported here wherein the prevalence of incorrect, non-misattribution responses (i.e., guesses) has consistently been in the realm of 10-15%, much lower than the 50% guess rate expected by chance in those previous experiments (where two of four possible response options were correct or misattributions).

It is possible that this inconsistent result is due to the introduction of extra response options. In this third experiment, participants chose between eight instead of four response options, which could feasibly be overwhelming and either delay responses to the point where time-based decay might have the chance to act (the average surprise trial response time was higher in this experiment than in the previous two by about 1,500 ms), or cause interference as extra items which must be processed before the response can be made. Or it could be simply that participants were asked explicitly about the order in which items were presented, and this cued their recall very poorly when they expected to judge congruency. Whatever the mechanism, clearly this change had a strong negative effect on participants' performance compared to previous experiments, to the point where they could no longer reliably recall which two semantic items they saw. An important takeaway from this study is that we may still not fully understand the impact of surprise questions on memory performance or the factors which mediate this effect.

General Discussion

To review, the three studies reported here had two major aims: first, to replicate and extend previous research to investigate whether the extent of source amnesia would differ depending on the type of information which was tested. This

subject is of high theoretical interest because replication of such a phenomenon in a second domain is very convincing of its importance for accommodation in memory models. Alternatively, if the finding had replicated in Experiment 1 when memory for color items was tested but not Experiment 2 when memory for verbal items was tested, this might have spoken to an essential difference between these stimulus types which would also need to be explained by models hoping to accommodate this phenomenon. The finding of a disparity between information types would also have mirrored the well-established Stroop interference disparity with the same information types (Stroop, 1935) and might have indicated similar underlying mechanisms in these two phenomena, opening avenues for better understanding of both. The second aim of this study was to test whether participants would be able to successfully identify which item was presented first and which was presented second, which would indicate that they were encoding the context of position in presentation order instead of stimulus format (colored square or written word). The implications of the findings of the final experiment are important for the Interference model (Oberauer & Lin, 2017; 2023), which emphasizes that associated context information is essential for the recall of an item.

In Experiment 1, our results closely replicated the findings of Chen et al. (2018): that source amnesia occurred to a greater extent when the probed information type was the one which was presented second in the trial and that the majority of errors were misattributions. Our novel finding from Experiment 2 is that source amnesia occurred to a very similar extent when participants were asked to recall the source of word information. Additionally, the proportions of errors which can be labelled as misattributions were very similar across the two experiments. We therefore conclude that regardless of whether color or word memory was tested, participants were likely to choose the option at test which was consistent with the meaning of the first-presented feature. Replicating a phenomenon such as this in a second domain bolsters its credibility and strengthens the argument for models of working memory to be amended to accommodate these findings. In addition, the chance-level results from our Experiment 3, which tested participants' memory for order information, lead us to conclude that no form of context which we can see is necessarily encoded alongside semantic representations of item memory when presented so rapidly.

Though it is interesting that these semantic representations appear to be very susceptible to loss, with participants guessing at random from the eight response options during Experiment 3, seemingly having lost even the previously preserved semantic item memory of what they had just seen. This particular finding leads us to wish for a better understanding of the factors influencing the impact that surprise questions can have on participants' memory performance. A study by O'Donnell and Wyble (2023) has already begun to address this and has concluded that while surprise questions do have an impact on participants' memory performance, this cannot account for the magnitude of information loss in source and attribute amnesia. Further, Muter (1980) compared trigram recall performance following a distraction task when participants were surprised with the recall prompt to when they were made aware from the beginning that this would occur on a small number of trials. They reported no notable difference in performance as a result of knowing that they would experience these "surprise"-esque trials, which implies that the element of surprise is not likely to cause the forgetting they witness in their method, and perhaps by extension, in this paradigm. In light of our novel finding, it seems that there is more to be uncovered on this subject and that it warrants more in-depth further study.

Addressing the Models

Following the experiments detailed here, we are more confident that participants do not encode the source nor any obvious context for the items they observe in this paradigm. For the task that participants intend to carry out (the pre-surprise task), they do not need to know which form the information they process was presented in: they only need to compare the semantic meanings of the stimuli they observe. We believe that the mind often conserves its resources where possible, and since source information is not needed in the pre-surprise task, it stands to reason that instead of being forgotten, it may purposefully not be encoded at all (an idea which has already been explored in Chen & Wyble, 2016). This is a problematic assumption for the Embedded Processes model, which posits that all attended information enters the focus of attention (and therefore activated long-term memory) at least briefly, and thus there should be some trace of source information accessible after so short a period. The Embedded Processes model could adjust to allow for rapid forgetting by introducing new boundary conditions on entry to the

focus of attention and/or allowing for de-activation of long-term memory under these circumstances. This is also potentially an issue for the Interference model, which, as discussed earlier, would argue that without associated context information, items should not be retrievable from memory. One would expect that a failure to encode source information would be problematic for the Multicomponent model because it would necessarily assign the verbal item to the verbal short-term store and the visual item to the visual short-term store, meaning that their source would be inherent depending on the store in which they are maintained. The same could be said for TBRS model here, given their suggestion of a verbal-only memory mechanism – if an item is being stored by that mechanism, it follows that it was presented in a verbal format. The Resource-Depletion theory suggests that unless context-item bindings are necessary, cognitive resource is not dedicated to forming them (Popov & Reder, 2020). This seems to be the case in this example, but this claim is discordant with the wealth of literature documenting the occurrence of incidental bindings (e.g., Treisman & Zhang, 2006; Campo et al., 2010; Morey, 2011; Logie et al., 2011; Santana & Galera, 2014; Elsley & Parmentier, 2015), so perhaps there is room to elaborate in this model which circumstances do and do not permit incidental binding when it is not explicitly called for.

Our results from Experiment 2 align with some of the findings by Xu et al. (2020), who used a similar methodology of visually presenting words. Interestingly, however, our findings diverge from theirs in their experiment wherein the words were presented auditorily, as they did not observe rapid forgetting. This suggests that memory for visually presented words is more susceptible to rapid forgetting compared to spoken words. At first glance, these results may seem difficult to reconcile with existing memory models. However, they align well with established phenomena such as the modality effect (Watkins & Watkins, 1977; 1980), the superior memory performance for recently presented items when information is presented auditorily rather than visually. Thus, our findings, along with those of Xu et al. (2020), may be reconciled with memory models that propose auditory presentations have distinctive characteristics that make them more resistant to forgetting or interference at least across periods this brief (e.g., Nairne, 1990; Saint-Aubin et al., 2021). Nevertheless, future research will be needed to directly evaluate this proposition.

An alternative reason as to why context may not be encoded in this paradigm could be that it is a result of the stimulus presentation rate. Popov, So and Reder (2022) found that the binding of some items (low-frequency words) to contexts (locations) was worse at very fast presentation rates (500 ms compared to 750 ms and 1000 ms). In the current experiments, stimuli were presented for even less time, perhaps suggesting that in some cases, it may be a natural consequence that item-context bindings are not made if presentation times are too brief. Further support for this may come from the Attentional Blink phenomenon frequently observed in experiments of the Rapid Serial Visual Presentation (RSVP) paradigm, which consistently show that at very fast list presentation times (e.g., 107 ms per item), a second target for detection and later recall is often missed if presented between approximately 200-500 ms after the successfully detected first target (Broadbent & Broadbent, 1987; Nieuwenstein & Potter, 2006; Potter et al., 2010). These findings could be taken to indicate that during a specific time window following encoding of the first target, the second target is not successfully bound to the context (which is what gives it its target status among the distractors, e.g., the color of the letter item or being marked by some punctuation indicator). In the RSVP task, the presence of multiple non-target distractors may mean that the item information for the second target is confused with distractors before recall can occur at the end of the list, but in this source amnesia paradigm where there are no distractors (only a brief mask), both items are remembered, and it seems that only the source is forgotten.

An alternative explanation to the failure to encode argument is that once information is removed from our focus, it may be specifically inhibited or suppressed to aid in task switching or conserve cognitive resources. This idea is discussed by Lewis-Peacock et al. (2018). In the current paradigm, if the second-presented item is removed from focus and specifically suppressed in favor of generating and holding a response plan to the pre-surprise trial incongruency judgment task (which is what participants would expect to do in the surprise trial before they see the new instructions), this might explain why memory for that second-presented item is poorly accessible. One could argue that this suppression would equally apply to the first-presented item and that it would be even harder to access given that it was presented earlier, but this might be counter-acted by some level of short-term consolidation (Jolicœur & Dell'Acqua, 1998) which was carried out to hold the first-

presented item during the very short mask between first and second items. Our confirmation in Experiment 2 that the preservation of the first-presented item occurs for verbal as well as visual features underscores the need to think further about potential boundary conditions on proposed maintenance processes in working memory. For example, complete removal might be more likely for information that has not yet been encoded to a particular degree, or perhaps has not figured into any plan.

Addressing Primacy Bias

Chen et al. put forward the idea of a primacy bias, which is not unsupported by the working memory literature: the primacy effect in memory (Oberauer et al., 2018) is a well-replicated effect which is often targeted for explanation by models. However, here Chen et al. would argue specifically that it is not the source of first-presented items is remembered, but instead that participants are biased to report the semantic representation of the first-presented item more often than that of the second-presented item. This is an incomplete explanation however, as it stands to reason that they should not only be blindly biased to report the first-presented item when the first-presented item was probed: they should 'guess' the first-presented item to the same extent whether they are in the word-first or the color-first condition. This is not what is seen in their Experiment 2 however: only in the square-first condition is the first-presented item most likely to be chosen. Additionally, in the Experiment 3 reported here, no such primacy bias was witnessed when the paradigm was altered very minorly to ask participants about serial order positions instead of stimulus format. An explanation is needed which accounts for this asymmetry of response better than an omnipresent bias towards the first-presented item. Perhaps in the source-probing version of the paradigm, some proportion of participants actually know the answer and there is a bias towards the first-presented item only in the case that a participant is unsure.

A particular strength of Popov and Reder's (2020) Resource-Depletion model is that it tidily explains the primacy effect in serial recall memory with its resource depletion mechanism (although see Popov, 2023 for discussion of a phenomenon within the primacy effect literature which does pose a problem for the model as it stands). The model states that the amount of resource dedicated to encoding each subsequent item declines as less resource is available for the task, and that the less

resource that is dedicated to encoding an item, the less easily it is retrieved. This seems to provide a good account for the primacy bias here: with such a short delay between presentation of the first and the second item, there would assumedly be very little opportunity (if any) for resource recovery, and thus we would expect the first item to be better recalled than the second. In addition, it is unclear what possible explanation this model could suggest for the knock-out effect which occurred in our Experiment 3 when participants were asked for serial order information instead of source information. Why would participants not be inclined again to rely on the primacy bias which they had used so consistently in the first two experiments? Surely with such emphasis in this model on the superiority of the first-encoded item, we would expect our participants to do very well when asked for the identity of that item.

We conclude that at very short presentation times, participants do not automatically encode any form of context when they do not require it for the task at hand. The performance data from our control trials and those reported in published literature in this and other realms of extremely rapid forgetting (Chen et al., 2018; Chen & Wyble, 2015; 2016) suggest that participants *can* maintain this context information when they believe that they need to do so. This therefore implies that there is some cost associated with encoding context information during such brief stimulus-presentation time periods. Working memory is ultimately for action in service of some goal. Perhaps, besides attention-based assumptions about what is encoded, models should focus on the fate of information prioritized for responding, emphasizing why that seems to differ from more incidental details.

Declarations

Funding

While working on this manuscript DG was supported by Experimental Psychology Society small grant and NC by NIH Grant R01-HD21338.

Conflicts of interest/Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

The studies reported were approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

Informed consent was obtained from all individual participants included in the study at the same time as taking consent to participate.

Availability of data and materials

All materials, program, data and the analysis scripts for this study are available at the OSF page (https://osf.io/mkwb2/). None of the experiments reported here were pre-registered.

Code availability

The program code for this study is available at the OSF page (https://osf.io/mkwb2/).

References

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific american*, *225*(2), 82-91. http://www.jstor.org/stable/24922803

Baddeley, A. D., Hitch, G. J., & Bower, G. A. (1974). Working memory. *Recent advances in learning and motivation, 8*, 47-89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., Hitch, G. J., & Allen, R. (2021). A multicomponent model of working memory. *Working memory: State of the science*, 10-43. https://doi.org/10.1093/oso/9780198842286.003.0002

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*(6), 575-589. https://doi.org/10.1016/S0022-5371(75)80045-4

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., & Camos, V. (2021). The time-based resource-sharing model of working memory. *Working memory: State of the science*, 85-115. https://doi.org/10.1093/oso/9780198842286.003.0004

Broadbent, D. E., & Broadbent, M. H. (1987). From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & psychophysics*, *42*(2), 105-113. https://doi.org/10.3758/BF03210498

Campo, P., Poch, C., Parmentier, F. B., Moratti, S., Elsley, J. V., Castellanos, N. P., ... & Maestú, F. (2010). Oscillatory activity in prefrontal and posterior regions during implicit letter-location binding. *Neuroimage*, *49*(3), 2807-2815. https://doi.org/10.1016/j.neuroimage.2009.10.024

Chen, H., Carlson, R. A., & Wyble, B. (2018). Is Source Information Automatically Available in Working Memory? *Psychological Science*, *29*(4), 645–655. http://doi.org/10.1177/0956797617742158

Chen, H., & Wyble, B. (2015). Amnesia for object attributes: Failure to report attended information that had just reached conscious awareness. *Psychological science, 26(2),* 203-210. https://doi.org/10.1177/095679761456064

Chen, H., & Wyble, B. (2016). Attribute amnesia reflects a lack of memory consolidation for attended information. *Journal of Experimental Psychology: Human Perception and Performance, 42(2),* 225. https://doi.org/10.1037/xhp0000133

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104*, 163-191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake, & P. Shah (Eds.), Models of working memory: Mechanisms of active maintenance and executive control (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24,* 1158-1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2021). An embedded-processes approach to working memory. *Working Memory: The state of the science*, *44*. https://doi.org/10.1093/oso/9780198842286.003.0003

Elsley, J. V., & Parmentier, F. B. R. (2015). Rapid Communication: The Asymmetry and Temporal Dynamics of Incidental Letter–Location Bindings in Working Memory. *Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://doi.org/10.1080/17470218.2014.982137

Jolicœur, P., & Dell'Acqua, R. (1998). The Demonstration of Short-Term Consolidation. *Cognitive Psychology, 36*(2), 138–202. https://doi.org/10.1006/cogp.1998.0684

Lewis-Peacock, J. A., Kessler, Y., & Oberauer K. (2018). The removal of information from working memory. *Ann N Y Acad Sci., 1424(1)*, 33-44. https://doi.org/10.1111/nyas.13714

Logie, R. H., Brockmole, J. R. & Jaswal, S. (2011). Feature binding in visual short-term memory is unaffected by task-irrelevant changes of location, shape, and color. *Mem Cogn, 39*, 24–36. https://doi.org/10.3758/s13421-010-0001-z

Morey, C. C. (2011). Maintaining binding in working memory: Comparing the effects of intentional goals and incidental affordances, *Consciousness and Cognition*, *20*(3), 920-927. https://doi.org/10.1016/j.concog.2010.12.013

Muter, P. (1980). Very rapid forgetting. *Memory & Cognition, 8*(2), 174-179. https://doi.org/10.3758/BF03213420

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*(3), 251–269. https://doi.org/10.3758/BF03213879

Nieuwenstein, M. R., & Potter, M. C. (2006). Temporal limits of selection and memory encoding: A comparison of whole versus partial report in rapid serial visual presentation. *Psychological science*, *17*(6), 471-475. https://doi.org/10.1111/j.1467-9280.2006.01730.x

Oberauer, K. (2021). Towards a theory of working memory. *Working memory: The state of the science*, 116-149. https://doi.org/10.1093/oso/9780198842286.003.0005

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological review*, *124*(1), 21. https://doi.org/10.1037/rev0000044

Oberauer, K., & Lin, H.-Y. (2023). An interference model for visual and verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001303

O'Donnell, R. E., & Wyble, B. (2023). Slipping through the cracks: The peril of unexpected interruption on the contents of working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49(6),* 990–1003. https://doi.org/10.1037/xlm0001214

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, *58*(3), 193. https://doi.org/10.1037/h0049234

Popov, V. (2023). Cognitive resources can be intentionally released when processed information becomes irrelevant: Insights from the primacy effect in working memory. https://doi.org/10.31234/osf.io/gct58

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127(1),* 1–46. https://doi.org/10.1037/rev0000161

Popov, V., So, M., & Reder, L. M. (2022). Memory resources recover gradually over time: The effects of word frequency, presentation rate, and list composition on binding errors and mnemonic precision in source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(9), 1263. https://doi.org/10.1037/xlm0001072

Potter, M. C., Wyble, B., Pandav, R., & Olejarczyk, J. (2010). Picture detection in rapid serial visual presentation: Features or identity?. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1486. https://doi.org/10.1037/a0018730

Saint-Aubin, J., Yearsley, J., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language, 118*, 104219. https://doi.org/10.1016/j.jml.2021.104219

Santana, J. J. R. A. d., & Galera, C. (2014). Visual-spatial and verbal-spatial binding in working memory. *Psychology & Neuroscience, 7*(3), 399–406. https://doi.org/10.3922/j.psns.2014.048

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are there multiple visual short-term memory stores? *PLoS ONE, 3(2),* e1699. https://doi.org/10.1371/journal.pone.0001699

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*, 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolKit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*, 24-31. https://doi.org/10.1177/0098628316677643

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18(6),* 643-662. https://doi.org/10.1037/h0054651

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory and Cognition*, *34*(8), 1704-1719. https://doi.org/10.3758/BF03195932

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, *6*(1), 13. https://doi.org/10.5334/joc.259

Virzi, R. A. & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13(4),* 304-319. https://doi.org/10.3758/BF03202499

Watkins, O. C., & Watkins, M. J. (1977). Serial recall and the modality effect: Effects of word frequency. *Journal of Experimental Psychology: Human Learning and Memory, 3*(6), 712–718. https://doi.org/10.1037/0278-7393.3.6.712

Watkins, O. C., & Watkins, M. J. (1980). The modality effect and echoic persistence. *Journal of Experimental Psychology: General, 109*(3), 251–278. https://doi.org/10.1037/0096-3445.109.3.251

Wyble, B., Hess, M., O'Donnell, R. E., Chen, H., & Eitam, B. (2019). Learning how to exploit sources of information. *Memory & Cognition, 47*, 696-705. https://doi.org/10.3758/s13421-018-0881-x

Xu, M., Fu, Y., Yu, J., Zhu, P., Shen, M., & Chen, H. (2020). Source information is inherently linked to working memory representation for auditory but not for visual stimuli. *Cognition*, *197*, 104160. https://doi.org/10.1016/j.cognition.2019.104160

5. A New Approach to Measuring Verbal-Spatial Binding Asymmetry

Introduction

To understand the complex and multi-faceted world around us, it is important that our minds are capable of knitting together the myriad of individual features which we perceive into cohesive units: objects and living things. To begin to develop a fundamental understanding of how this is done for multipartite stimuli in our busy environment, it is useful to break down the problem into smaller parts and study feature integration in isolation, for instance, '*how do colours and shapes/shapes and locations/letters and colours become bound to one another?*'. This is the purpose of studies that investigate 'binding'. Over the years, a large body of research has been dedicated to understanding the integration of a wide variety of different features. Regardless of the type of features involved, there are two overarching categories of binding: 'intentional' and 'incidental'. Intentional binding can be classified as instances wherein participants know that they will be tested on the conjunction of the bound features, whereas incidental binding is considered to occur accidentally when participants believe that they are only required to commit one of the features to memory, but happen to remember both. Morey (2011) found a substantial difference between participants' performance when binding was possible (but not required) compared to when participants were instructed to bind in an otherwise identical task. A noteworthy finding was that intentional binding instructions elicited fewer false alarms for recognition of bound items than incidental binding. Morey concluded from her data that the intention to bind conferred considerable cognitive benefits which did not occur during the incidental binding version of the task. Similarly, Lekeu et al. (2002) found a difference in the errors that participants made in their comparison between incidental and intentional binding. While there was seemingly no benefit to intentional binding in their corrected recognition score (number of hits minus number of false alarms), an analysis of the errors that participants made revealed that false alarms for bindings were less common in their intentional binding condition than in their incidental binding condition.

The process behind binding of verbal items to their locations is of particular interest due to its integral role in reading. The retention of both the locations of letters within a word and locations of words within a sentence are essential for complete understanding of phonographic written communication. Treisman and Zhang (2006) found early evidence to attest to the important role of location information in feature integration involving verbal stimuli. Their data from a delayed match to sample experiment suggested that when letter-colour bindings were maintained from memorandum to probe, recognition of a multipartite display was hampered by the stimuli being presented in new locations compared to old locations. Additionally, when letters were presented in a new colour (a different binding), being presented at test in their old locations hampered recognition performance compared to presentation in new locations. This finding of differential recognition benefits and false alarms depending on the interplay of location with the integrity of feature bindings suggests that location information is important in the process of feature integration involving verbal components. Some models of memory also promote the important and unique role of location bindings. For instance, Oberauer and Lin's (2017; 2023) interference model of working memory proposes that without clearly encoded 'context' information (typical examples are serial order position or location in a display) to use as retrieval cues, item information cannot easily be recalled due to interference from other items with overlapping context cues. Therefore, according to this model, accurate location binding is an absolutely essential process in our understanding of the world, which may explain why studies have indicated its prioritisation and uniqueness when compared to other feature bindings.

In the specific realm of research on incidental verbal-spatial binding, a few studies have reported an interesting difference when testing memory for the verbal compared to the spatial information. This asymmetry consists of more extensive binding when participants thought they were being tested on only verbal information than when they thought they were being tested on only spatial information. Campo et al. (2010) and Elsley and Parmentier (2015) ran experiments using very similar memory displays consisting of letters placed within rectangular frames which form a circle centred on the middle of the computer screen. During each trial, four consonant letters (chosen at random from the eight letters in the stimulus pool)

would appear, one per frame, within four of the eight frames which make up the circular display. See Figure 1 below for a demonstration.

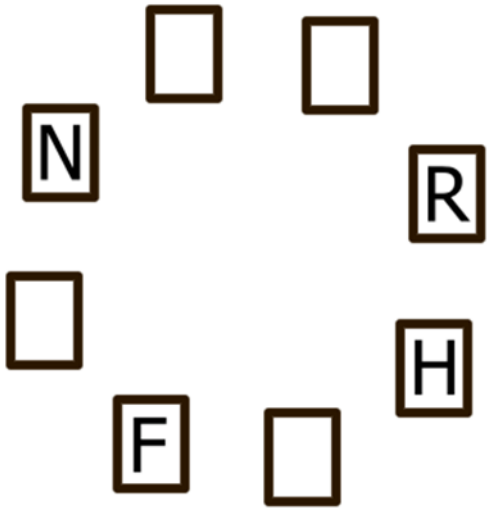

Figure 1 shows a rough example of a verbal-spatial memory array made up of eight frames positioned around an invisible circle. Four of the frames are inhabited by consonant letters.

Participants could either be asked to commit to memory the identities of the letters which appeared or the locations in which they appeared (the frames), which were the verbal and spatial tasks, respectively. To measure participants' memory, each trial ended with a recognition probe consisting of one of the possible letters in one of the possible frames. Participants had to indicate whether they had seen *either* the letter or the location in that trial, depending on the task type. The probes requiring a 'yes' response could be *intact*, meaning that the letter presented is in the same location as it was in the display, or *recombined*, meaning that the letter and location were both used in the display, but not paired together. Examples of these are both demonstrated in Figure 2 below. The difference in accuracy (or response times) between responses to intact and recombined probes is considered to be a measure of incidental binding, with more accurate (and faster) responses being expected to intact than to recombined probes. Campo et al. (2010) and Elsley and Parmentier (2015) saw in their data evidence of significantly more incidental binding when participants were tasked with committing letters to memory than when they were tasked with committing locations to memory.

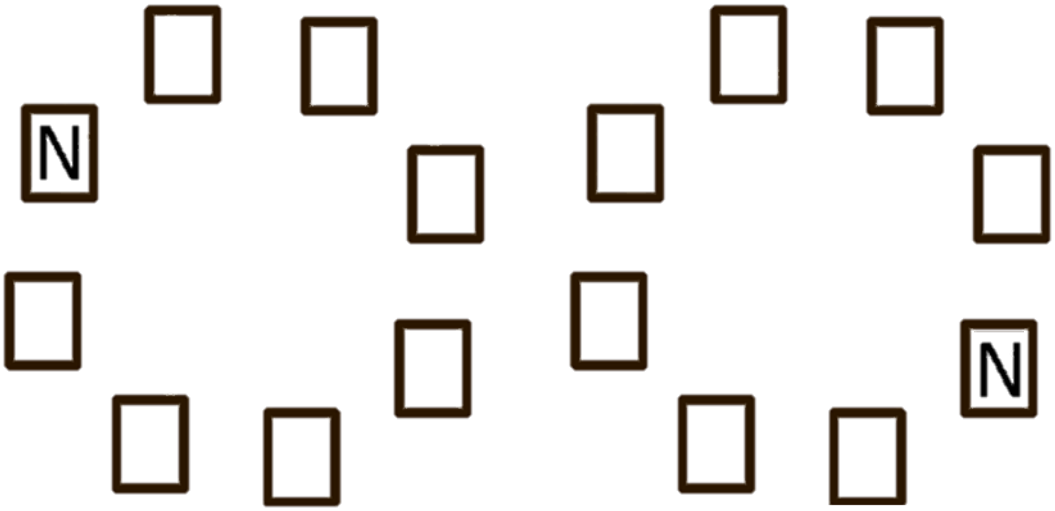

Figure 2 shows a rough example of an intact probe (left image), where a letter from the memory array appears in the same position, and a recombined probe (right image), where a letter from the memory array appears in a position where a different letter appeared.

Despite studies reporting this binding asymmetry (Campo et al, 2010; Elsley & Parmentier, 2015), a more recent attempt to investigate the factors affecting this phenomenon have failed to replicate the asymmetry (Delooze et al., 2022). In that study's first experiment, the stimuli varied marginally from those used before in that the consonants were a different set of 10 letters presented in circular frames. The procedure differed as well, with the task of remembering the letter or the location not separated into blocks as had been done previously, but instead the task varied from trial to trial and was either pre-cued or post-cued (this variable was blocked). Bayesian analysis methods suggested that there was moderately (in the accuracy measure, *BF=5.74*) and extremely strong (in the reaction time measure, *BF=352.38*) evidence against a pattern of incidental binding wherein focusing on letters elicited more binding than focusing on locations. In the study's second experiment, the experimenters attempted to replicate the previously used method more closely by using the same consonant letter set and removing all retro-cue trials to leave a letter focus block, a location focus block and a mixed focus block, the first two of which they reasoned should be close replications of Elsley and Parmentier's (2015) method. Analysis revealed that this data still favoured a conclusion of no asymmetry (*BF=200.29* for the reaction time measure). However, if an asymmetry were to be accepted, the graphs suggested stronger binding in the location task rather than the letter task, opposite to the way in which the asymmetry had been documented

before. These failures to replicate the proposed asymmetric binding pattern suggest that the phenomenon may be very weak, or possibly susceptible to interference from factors we do not currently understand.

A noteworthy issue with this recognition-centred design for measuring binding is that it is necessary to collect a lot of data that cannot be used to assess memory for binding. To measure binding, only the responses to intact and recombined probes are useful, both of which require 'yes' responses from participants. If we were to run an experiment wherein only these trial types are included, participants would very quickly notice that they are always required to give a 'yes' response, and would begin to respond without actually considering the probe, rendering the data useless. Therefore, a significant proportion of trials requiring various types of 'no' response must also be included to balance participants' experience, but the data from these trials is not hypothesised to reveal anything about binding and gets overlooked. This attrition of data wastes time for both researchers and participants and limits how much useful data can be collected before participants start to suffer from the effects of fatigue, or possibly other transitions that might affect how they perform the task. It would be useful to deploy an experimental design capable of assessing binding but which does not rely solely on recognition responses, and which can provide richer data about what is remembered about incidental letter or location features.

A further motivation to run an experiment on this topic using a non-recognition measure of binding is that there is evidence from previous studies on memory to suggest that encoding is different when participants intend to give a recognition response than when they intend to recall information (e.g., Carey & Lockhart, 1973; Tversky, 1973; Hall et al., 1976, Uittenhove et al., 2019). Those findings suggest that in the existing paradigm to measure incidental verbal-spatial binding, participants' intention to make a recognition response may involve weaker encoding processes and lead to different results than if they were to intend to recall the to-be-remembered information. Thus, to use this method but instead implement a recall-style memory response would potentially also allow us to add commentary to a wider-reaching debate on memory, beyond this narrow question of whether verbal-spatial binding occurs to varying degrees depending on the feature participants are focusing on. Therefore, we set out first to investigate whether incidental binding can

be detected using a recall response and also whether there is any evidence to suggest an asymmetric pattern using this new kind of response.

In response to these ideas, we created an experiment similar to those used by Campo et al. (2010) and Elsley and Parmentier (2015), wherein for each trial, participants would see an array of four letters in four locations around a circle and would be required to commit to memory either the identities of the letters which appeared or the locations in which they appeared. Where our experiment differed is that after a short delay, participants were presented with a response screen containing all possible letters and all possible locations. With a computer mouse, they were required to select and drag four letters into four of the locations. If their current task was to remember the letters, they should choose the letters which they remember from the array and place them into any of the eight possible locations. If their current task was to remember the locations, they should choose any of the eight possible letters and place them into the locations that they remember from the array. With this freedom in responding, would participants gravitate toward placing letters in their original locations, even though this was not required?

Methods

Participants

Forty-one undergraduates from the School of Psychology at Cardiff University participated in the study in exchange for course credit or payment of approximately £10. They were recruited opportunistically through the school's Experiment Management System. Demographic information was not collected from participants directly, as it was not deemed relevant to the experimental aims at the time. This limits the conclusions which could be drawn from this data regarding generalizability. The student population from which participants were taken was mostly female, mostly aged 18-24 years, and all were fluent speakers of English, as required for students taking an undergraduate course in the medium of English. No participants met our exclusion criterion of accuracy performance lower than the 66% threshold, so all 41 were included in analyses.

Materials & Apparatus

The letters which could appear within the study were the consonants N, J, R, D, T, Q, F and H. These appeared in their lowercase forms within the memory display (an example of this can be seen in Figure 3 below) and in their uppercase forms during the response phase. The font size for stimuli throughout the experiment was 30pt. and all text, including instructions, appeared in white font to maximise contrast to the black background. The locations in which these letters could appear were eight rectangular boxes around the circumference of a circle (with a radius of 200pt.), which remained in identical positions throughout all phases of all trials, 45° apart. The experimental program was built and run in OpenSesame, (Mathôt et al., 2012). Participants responded using a standard wired computer mouse positioned on a mouse pad on the desk in front of them. The computer monitors on which they viewed the experiment were iiyama ProLite XUB2294HSU 21.5-inch monitors with a maximum resolution of 1920x1080 pixels.

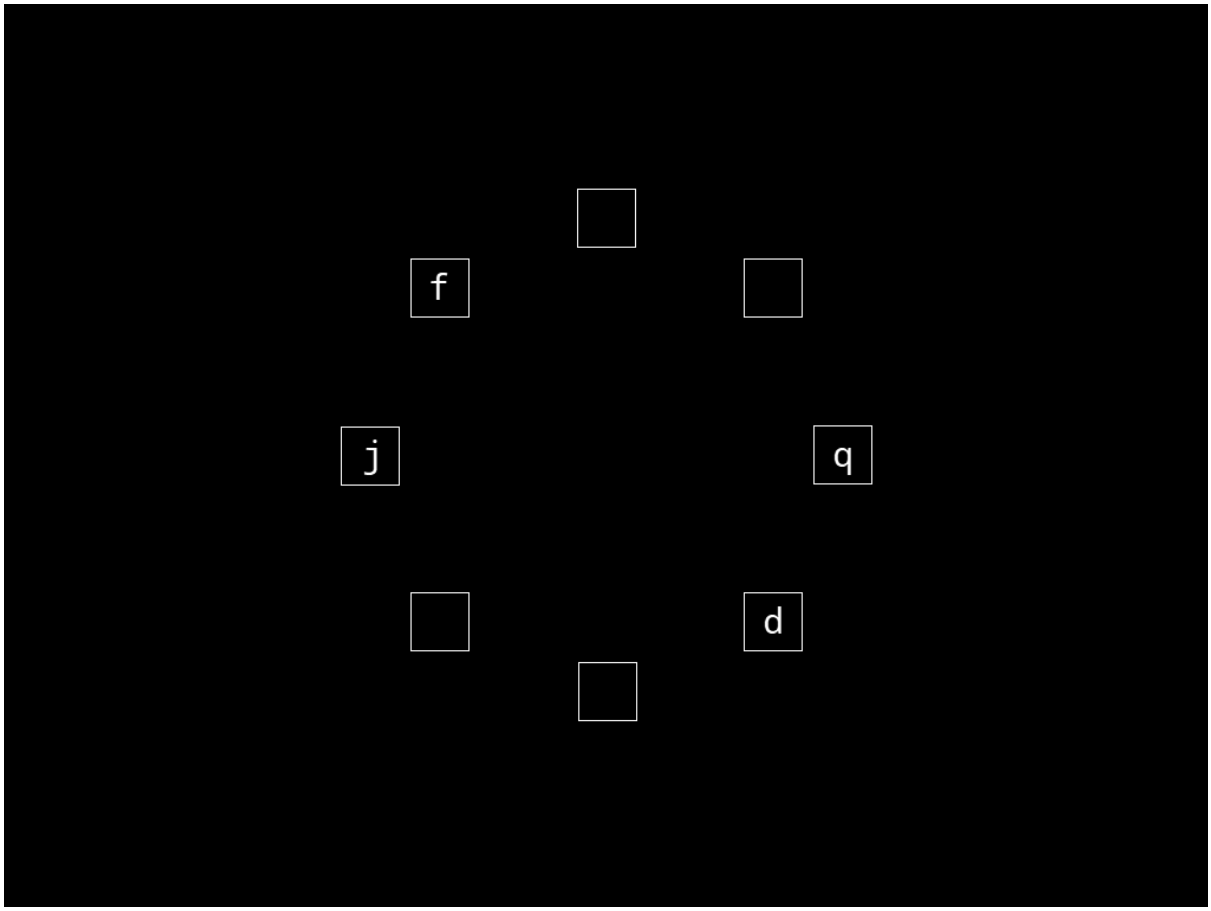

Figure 3 shows the memory display used in the current experiment, demonstrating the locations in which letters could appear.

Design

The experiment was a 2 (task focus: letter or location) x2 (retention interval: 400ms or 4900ms) x2 (counterbalance: letter-block-first or location-block-first) design, with the factors of task focus and retention interval manipulated within-subjects and the factor of counterbalance manipulated between-subjects. Task focus was blocked so that each half of the experiment consisted of a different task, but retention interval was not blocked, so that trials of shorter length were interspersed among trials of longer length. The dependent variables of interest were accuracy (how many items participants answered correctly with regard to the feature on which they focused) and binding (how many items participants answered correctly with regard to both the feature on which they focused and also with regard to the feature which they were instructed to ignore).

Procedure

Participants were welcomed to the lab and seated in one of two experiment booths. The experimenter briefly explained the procedure to participants, with their key goal to clearly communicate and emphasise that participants did not at any point need to remember the bindings of the displayed items, only the one feature which they were tasked with remembering. Participants were given the opportunity to ask any questions they had. They were then directed to read about the details of the study procedure and consent to take part within the experimental program by clicking a button marked “I Consent”. Written instructions were provided to participants in addition to, but reiterating, the verbal instructions given by the experimenter. All instructional and consent form text can be found in the supplementary files of this project’s OSF page (https://osf.io/48uxn/). The experiment consisted of a total of 80 trials, which were divided evenly in half between the two tasks, and further subdivided evenly between the two retention interval lengths. Participants were given the opportunity to take a self-timed break at the half-way point during the instruction screen which detailed the nature of the second task.

The beginning of each trial regardless of task type consisted of a screen, blank except for a centrally presented small white fixation dot which lasted for 500ms. In both tasks, participants were shown a memory display which endured for 2000ms wherein four of the eight boxes were populated with lowercase consonants

(from the sample described in *Materials & Apparatus*). In the letter focus block, participants were required to commit the identities of the letters to memory, and in the location focus block, participants were required to commit the locations of the filled boxes to memory. Following this display period, the boxes disappeared from view and fixation appeared again in the centre of the screen for the duration of the retention period (either 400ms or 4900ms). Finally, the eight empty boxes reappeared, in addition to a randomly ordered horizontal array of the eight consonants in the sample, situated just below the boxes. This marked the beginning of the response phase, and these items would not disappear from the screen until participants responded. To respond, participants used their mouse to drag four of the letters at the bottom of the screen into four of the empty boxes. Once a letter was placed in a box (by participants releasing the mouse button while holding a letter over the box – there was no tracking to guide their answers if they did not release within the confines of a box), that letter was not replaced in the array at the bottom, and no other letters could be placed in the inhabited box. At the time that four of the boxes had been populated, the screen advanced to a feedback phase, which communicated to participants whether they answered the trial correctly or incorrectly using a small circle in the centre of the screen, now coloured green or red, respectively.

The cursor was always forcibly returned to the same central location at the beginning of each new response (after each time after they drop a letter into a box). We were also able to see the order in which participants addressed each letter in their response, e.g., which box was filled first, and which letter was selected to inhabit it. This study was approved by the Cardiff University School of Psychology Research Ethics Committee, and conducted in keeping with the principles of the Declaration of Helsinki.

Results

Analysis Plan

The analyses reported here are exploratory, and not pre-registered. First, we will report descriptive statistics for the participants' accuracy. If any participant has an overall score of less than 66% correct, their entire data set will be removed from the analysis. A 2-way repeated measure ANOVA will then be conducted on the accuracy

data to determine whether there are any systematic differences in accuracy as a result of the task type (letter or location focus) or retention interval (400ms or 4900ms) or an interaction between the two. Next, we will move on to binding. We will explain how we will measure binding and report descriptive statistics for that measure. Then, we will analyse order effects on the binding data using a 3-way mixed measures ANOVA with the two factors already outlined above plus the counterbalance group (letter task first or location task first) as a between-subjects factor. Following up on this idea of potential change in behaviour over time, we will investigate whether or not participants demonstrated fatigue or practise effects. This will be done by correlating the extent of binding with trial number within each block. Finally, we will test using Chi-Squared Goodness-of-Fit analyses our suspicion that the method permits participants to exhibit a 'lazy bias', wherein they favour placing task-relevant letters into the bottom few locations (which require less physical effort) while carrying out the letter focus task, rather than responding randomly in the task-irrelevant feature.

Accuracy

Only two participants scored lower than 66% accuracy on any given trial type, but these did not result in an overall participant score of less than 66% in either case, so they were not excluded from analysis. The sample's average performance is shown in Table 1 below. Performance on the task was significantly better (though the effect size is small and both averages are near to ceiling) in the letter block than the location block ($F(1,40)=9.035$, $p=.005$, $\eta_p^2=.184$). Retention interval did not significantly affect participants' accuracy, nor was there a significant interaction between the two factors.

Trial type (focus_retention interval)	Mean trial accuracy (/20)	Std. Deviation
letter_400ms	19.122	1.345
letter_4900ms	19.244	1.135
location_400ms	18.756	1.410
location_4900ms	18.293	2.136

Table 1 shows the descriptive statistics to three decimal places for accuracy in the specified task.

Binding

Having established that participants were capable of the task and that there were no large fluctuations in accuracy which might hinder further interpretation of results, next we address the extent to which binding occurred. To define binding in this analysis, an “instance” of binding occurs when a participant drags a letter into the same box which it inhabited during the memory display. Therefore, any one trial can detect as many as four instances of binding. With 20 trials in each task focus by retention interval cell, this means that a perfect binding score would be 80. If a participant were to bind on every item on every trial across all four cells, they would have an overall binding score of 320. Table 2 below shows the mean number of items which were bound for each task focus by retention interval cell across both counterbalance groups. It demonstrates a slightly greater tendency to bind for the letter focus trials overall than the location focus trials, but that the variability was very high for both. Importantly, it suggests that more incidental binding is occurring than one might have expected based on chance alone. If participants disregarded the incidental feature and chose randomly where to drop their recalled letter or which letter to place in their recalled location, one would expect to see intact binding about 16% of the time, or on approximately 13 out of 80 items.

Trial type (focus_retention interval)	Mean number of bindings (/80)	Std. Deviation
letter_400ms	47.512	28.806
letter_4900ms	46.561	29.177
location_400ms	38.537	29.093
location_4900ms	37.122	30.132

Table 2 shows the descriptive statistics to three decimal places for the number of items which were bound.

Order effects

A 3-way ANOVA was run on the counts of bound responses data, with task focus and retention interval as within-subjects factors and counterbalance group as a between-subjects factor. This revealed a significant main effect of task focus ($F(1,39)=5.116$, $p=.029$, $\eta_p^2=.116$) and two significant interactions: task focus by counterbalance ($F(1,39)=20.195$, $p<.001$, $\eta_p^2=.341$), which is shown in Figure 4 below, and retention interval by counterbalance ($F(1,39)=4.711$, $p=.036$, $\eta_p^2=.108$). This latter interaction reflects a negligible effect of retention interval on the extent of

binding in the letter-first group (a very small difference of 0.5 items bound), and a tendency to bind more in the shorter trials in the location-first group (a difference of 5.095 items bound). No other main effects or interactions were significant.

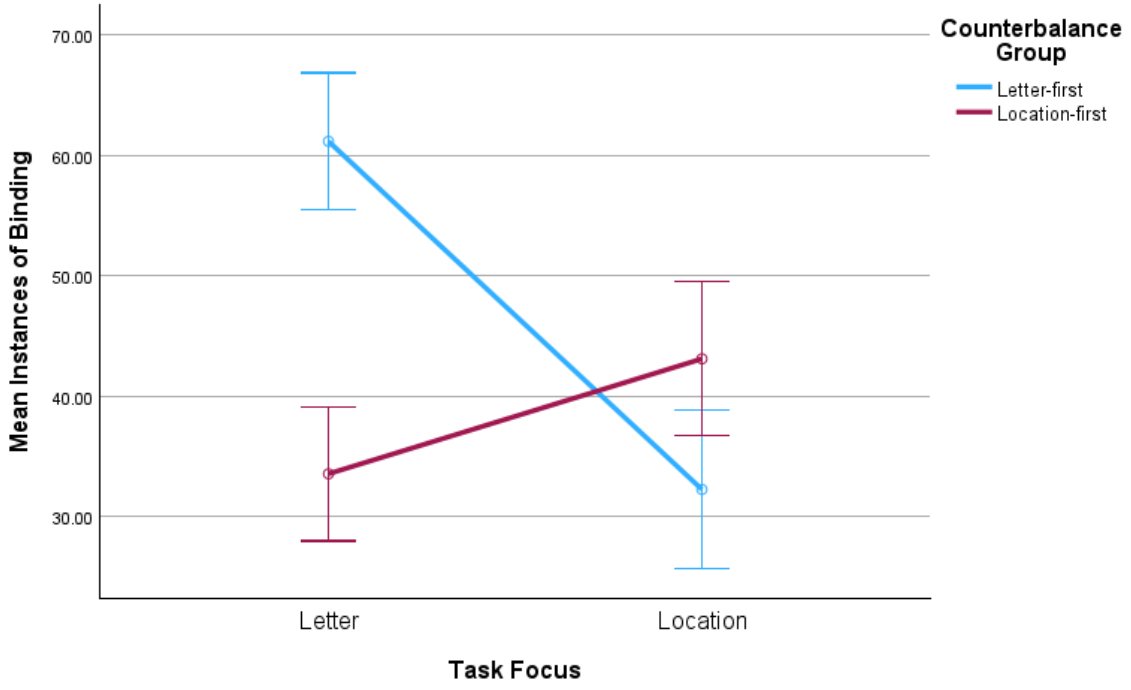

Figure 4 shows the average instances of binding as a function of the within-subjects factor of task focus and the between-subjects factor of counterbalance group. Error bars represent +/- 1 standard error.

To better understand these between-subjects differences, further analyses were run on the groups separately. In the letter-first participants, a 2-way ANOVA revealed that the effect of task focus was significant ($F(1,19)=19.064$, $p<.001$, $\eta_p^2=.501$), with instances of binding occurring approximately twice as often in the letter block compared to the location block. Neither the effect of retention interval nor the interaction between retention interval and task focus were significant. In the location-first participants, the effect of retention interval was significant ($F(1,20)=10.588$, $p<.005$, $\eta_p^2=.346$) according to a 2-way ANOVA, with binding occurring more often in the short retention interval than in the long retention interval. Neither task focus nor the task focus by retention interval interaction were significant. These results suggest that order effects exist in the letter-first counterbalance group but not in the location-first group. Something about the letter task encouraged participants to bind much more during the task, but only when it was experienced first.

Practise or fatigue effects

A Kolmogorov-Smirnov test for normality revealed that the data were sufficiently close to a normal distribution to satisfy a parametric correlational analysis method, so Pearson's r tests were conducted. In the letter-first group, there were no significant correlations between trial number and average instances of binding for either the letter or the location focus block (Figure 5 below). In the location-first group, there was a significant positive correlation between trial number and average instances of binding ($r(38)=.277$, $p=.042$, one-tailed) in the location focus block, but not in the letter focus block. This positive correlation shown in Figure 6 below reveals that participants who took part in the location-focus block as the first half of the experiment became more likely to bind as more trials passed. Possibly, this reflects the task becoming easier over time and thus more cognitive resource being available to devote to encoding unnecessary information. However, this is not supported by the accuracy data for the intentionally-remembered feature: a Spearman's rho analysis conducted on the arcsine square root transformed accuracy data from the same trials suggests that there is evidence against a correlation between trial number and accuracy ($r_s(38)=-.119$, $p>.05$). To summarise, despite evidence for order effects indicating in some cases, namely the letter-first counterbalance group, that participants were more likely to bind in the first task than in the second task they experienced, there is no evidence to suggest that this decline in binding is due to fatigue or increased capacity for the task as a result of practise. Contrarily, the only trials which showed evidence of change over time within-block were the location task trials from the location-first counterbalance group, which did not show evidence of order effects in the inferential analyses detailed above.

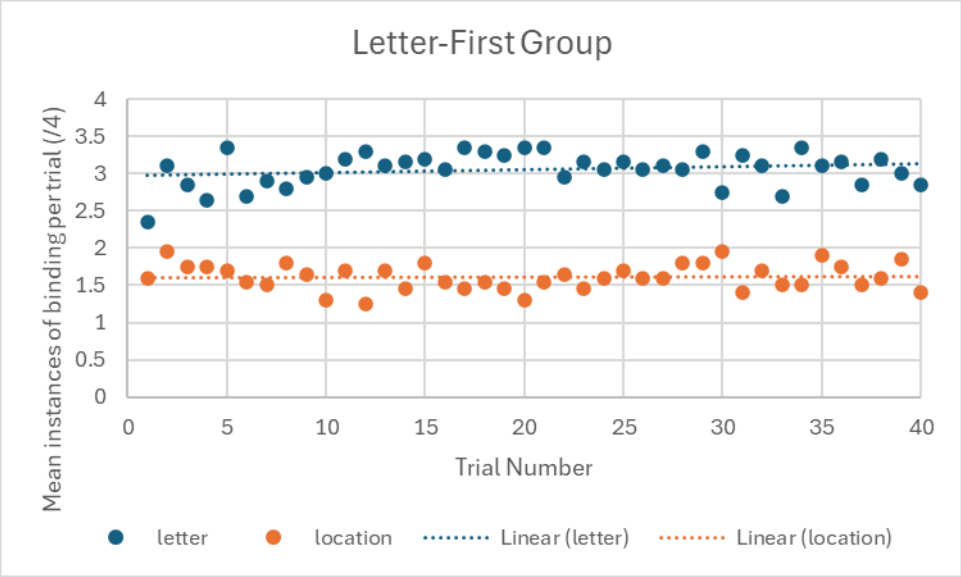

Figure 5 shows the average instances of binding per trial (maximum possible value of 4) as a function of trial number within a block and task focus for the letter-first counterbalance group.

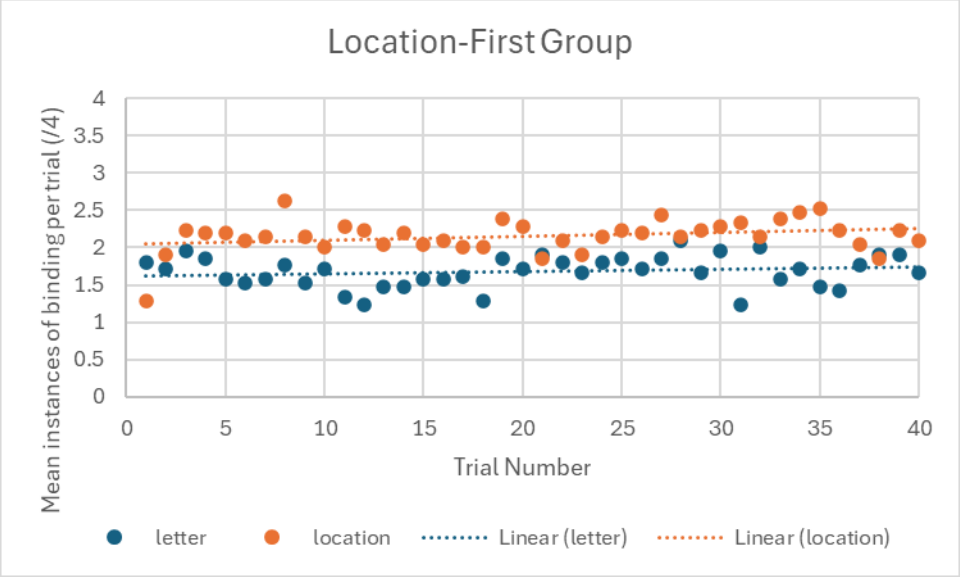

Figure 6 shows the average instances of binding per trial (maximum possible value of 4) as a function of trial number within a block and task focus for the location-first counterbalance group.

Lazy bias

We wished to assess whether participants were likely to demonstrate a lazy bias, i.e., a preference to place letters in the lower-positioned boxes over the higher-positioned boxes when the task did not require a location-based response. Our logic was that placement in the lowest three boxes (plus one or the other of the boxes in the middle) of the response display required less physical effort than positioning the

letters into random boxes, and therefore participants who are not binding would perhaps demonstrate this behaviour. We collated data containing which boxes were selected (had a letter placed into them for the response) on each trial in the letter focus task and used it to run Chi-squared goodness-of-fit inferential analyses. These analyses revealed that for the letter-first counterbalance group, distributions of responses did not significantly differ from chance for either the short ($X^2(2, N=20) = 2.138$, $p > .05$, $\varphi c = .231$) or long ($X^2(2, N=20) = 2.086$, $p > .05$, $\varphi c = .228$) retention interval. However, for the location-first counterbalance group, distributions of responses did significantly differ from chance for both the short ($X^2(2, N=21) = 21.332$, $p < .001$, $\varphi c = .713$) and long ($X^2(2, N=21) = 22.817$, $p < .001$, $\varphi c = .737$) retention intervals. This data, illustrated in Figure 7 below, suggests that participants in the location-first counterbalance group were likely to demonstrate a 'lazy bias', characterised by a significantly increased tendency to put letters into the lower positioned boxes. Whereas participants in the letter-first counterbalance group had a considerably more modest tendency to place letters into the lower positioned boxes, they responded at chance level for their distribution of responses. Since letters were distributed at chance level into each box in the memory display, this chance-level pattern of responding aligns with the evidence that participants in the letter-first group were binding a lot in this task, whereas participants in the location-first counterbalance group were binding less and typically exhibiting a lazy bias.

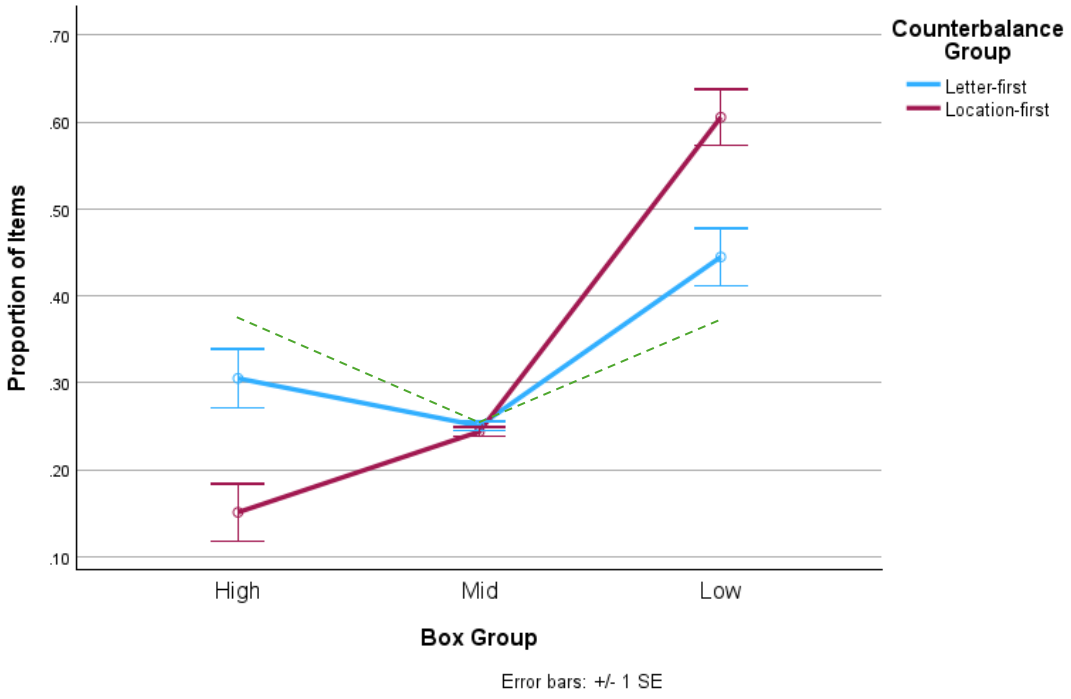

Figure 7 shows the proportion of items placed in each group of boxes (high, mid and low) as a function of counterbalance group. Error bars represent +/- 1 standard error. Chance responding (which reflects the number of boxes within that box group) is illustrated in the green dashed lines.

Discussion

The primary aim of the current experiment was to determine whether a recall version of the incidental binding paradigm would be able to detect any evidence of binding, and if so, would it help with testing whether binding is more likely when focusing on a particular kind of feature? Our motivation was that the current paradigm for measuring incidental binding relies on data only from trials which require a 'yes' recognition response, which unfortunately cannot be all or a large majority of trials, else participants may notice the trend, which would cause a bias in responding. The necessity to include 'no' response trials in a substantial proportion means that a lot of trials in the recognition probe paradigm are wasted. We wanted to design a method of responding that would maximise the usefulness of all trials, provide rich data, and assess using a very simple version of the experiment whether it was capable of capturing any evidence of binding, which we believe has been achieved given the data presented here.

The major influence on the data was the interaction between counterbalance group and task focus, reflecting that binding was most likely to occur in the first block

when that first block was the letter focus task. Meanwhile, participants in the location-first counterbalance group did not bind significantly more or less in one or the other task. These data serve to somewhat support the trend of binding asymmetry reported in the literature discussed earlier (Campo et al., 2010; Elsley & Parmentier, 2015), as they replicate the finding that encoding a letter can sometimes confer with it a benefit to the task of recalling the location in which that letter was presented. However, the support that these data provide is tempered by the moderating factor of task order (the between-subjects independent variable of counterbalance group): it is important to note that when participants experienced the location focus task first, they were not nearly as likely to bind in the letter task. In fact, the data indicate a small but non-significant trend for more binding to occur in the first task, the location focus trials, than in the letter focus task, suggesting that it is the dual effects of the nature of the letter task and its position as the first of the two tasks which cause so much binding to occur, rather than a strong asymmetry which endures despite circumstances outside of task focus. Elsley and Parmentier (2015) and Campo et al. (2010) specify that their task orders were counterbalanced, but do not report any assessment of whether their counterbalance groups differed. In the analysis from Delooze et al. (2022), we also did not address such a possibility. It is possible that the same interactions would emerge in those three data sets if the appropriate analyses were to be conducted, or this may be unique to this recall-style response.

The nature of the measurement of binding in this task is arguably more explicit than the previously used measure of taking the difference in response times to recognise a recombined compared to an intact probe. Our intuition is that some participants may have taken cues from the response method that they should bind, despite the fact that they were reassured that maintaining the bindings was unnecessary. It is possible therefore, that our method is more susceptible to accidentally measuring intentional binding than the recognition response method used by Campo et al. (2010), Elsley and Parmentier (2015) and Delooze et al. (2022). A caveat, however, is that we have no reason to expect that this influence would occur to a greater extent for one task compared to the other, so the fact that letter-first participants behaved so differently to location-first participants is possible evidence to the contrary, that participants on the whole may have acted exactly as

instructed. If future studies were to adopt this method of measuring incidental binding, we suggest that they also implement some sort of question at the end of their methods to probe whether their participants intentionally maintained the task-irrelevant information.

That our data has, in one group of participants, replicated the asymmetric binding effect demonstrated in Campo et al. (2010) and Elsley and Parmentier's (2015) work suggests that the intention to recall or recognise information does not qualitatively impact incidental binding in that it does not change which items are bound and when. However, the manipulation may have a quantitative impact. In the current experiment, participants knew that they would be required to recall either the identity of the letters which they saw or the locations in which they saw them, and that they would do this by choosing the correct four letters or the correct four locations from an array of eight possible letters or locations. In those experiments which have detected this binding asymmetry previously, participants knew that they would have to look at an example of a letter in a location and respond yes or no to whether that letter or that location had been used in the memory array. The fact that the same asymmetric pattern of results is found regardless of memory task type (recognition or recall) suggests that this difference in memory-related intention does not qualitatively impact the outcome of incidental binding, at least when using these stimuli. However, we report a much greater effect size for the size of the difference in binding as a result of the task focus (η_p^2=.501) than that reported by Elsley and Parmentier (η_p^2=.190 - the equivalent result is not reported in Campo et al., 2010). This quantitative difference of a greater difference in the extent of binding might be attributed to the stronger encoding which is posited to occur when participants intend to recall rather than recognise memory items.

This study was an exploration into a new method of measuring incidental verbal-spatial binding. By implementing a drag and drop recall response in place of the probe recognition which has been used previously, we were able to collect evidence of verbal-spatial binding while not only preserving every trial, but acquiring multiple data points per trial. Our data also recreates the asymmetry witnessed in those recognition studies, of binding occurring to a much greater extent in the verbal task than in the spatial task, with the important caveat that this asymmetry is tempered by the order in which participants experience the tasks: only participants

who did the letter task first bound more in that task. This may go some way towards explaining why our earlier attempts to replicate this phenomenon have failed (Delooze et al., 2022). By corroborating the pattern sometimes found in recognition memory, but to a considerably greater extent using recall responses, this research has also contributed to the discussion around the differences between recall and recognition. It appears that incidental verbal-spatial binding can be included among the areas of cognition wherein the intention to carry out one or the other of these two memory tasks produces quantitatively different results, likely by strengthening the encoding quality of the to-be-remembered items.

References

Campo, P., Poch, C., Parmentier, F. B. R., Moratti, S., Elsley, J. V., Castellanos, N. P.,…Maestú, F. (2010). Oscillatory activity in prefrontal and posterior regions during implicit letter-location binding. *Neuroimage, 49*, 2807-2815.

Carey, S. T., & Lockhart, R. S. (1973). Encoding differences in recognition and recall. *Memory & Cognition, 1,* 297-300. https://doi.org/10.3758/BF03198112

Delooze, M. A., Langerock, N., Macy, R., Vergauwe, E., & Morey, C. C. (2022). Encode a letter and get its location for free? Assessing incidental binding of verbal and spatial features. *Brain Sciences*, *12*(6), 685. https://doi.org/10.3390/brainsci12060685

Elsley, J. V. & Parmentier, F. B. R. (2015). Rapid Communication: The asymmetry and temporal dynamics of incidental letter-location bindings in working memory. *Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://dx.doi.org/10.1080/17470218.2014.982137

Hall, J. W., Grossman, L. R, & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4* (5), 507-513. https://doi.org/10.3758/BF03213211

Lekeu, F., Marczewski, P., Van der Linden, M., Collette, F., Degueldre, C., Del Fiore, G., ... & Salmon, E. (2002). Effects of incidental and intentional feature binding on recognition: a behavioural and PET activation

study. *Neuropsychologia*, *40*(2), 131-144. https://doi.org/10.1016/S0028-3932(01)00088-4

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324. https://doi.org/10.3758/s13428-011-0168-7

Morey, C. C. (2011). Maintaining binding in working memory: Comparing the effects of intentional goals and incidental affordances. *Consciousness and Cognition*, *20*(3), 920-927. https://doi.org/10.1016/j.concog.2010.12.013

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological review*, *124*(1), 21. https://doi.org/10.1037/rev0000044

Oberauer, K., & Lin, H.-Y. (2023). An interference model for visual and verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001303

Treisman, A. & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition, 34*(8), 1704-1719. https://doi.org/10.3758/BF03195932

Tversky, B. (1973). Encoding Processes in Recognition and Recall. *Cognitive Psychology, 5,* 275-287. https://doi.org/10.1016/0010-0285(73)90037-6

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is Working Memory Storage Intrinsically Domain-Specific? *Journal of Experimental Psychology. General*, *148*(11), 2027–2057. https://doi.org/10.1037/xge0000566

6. Discussion

To address the key themes of this thesis again, we consider the new findings that the current work has provided. We then briefly assess the feasibility of applying a theory of feature binding to the broader sphere of cognitive effects examined here to answer the question of whether visually presented information is more dominant than other visual stimuli.

Interference

First, we consider briefly what our findings mean for interference as a whole. The experiments detailed in Chapter 2 demonstrate that both forward and reverse Stroop effects can be observed in the same method in the same participants without relying on pre-exposures, stimulus obscuring or articulatory suppression. Adding to the already established finding in colour-word Stroop tasks, this work has demonstrated that in verbal-spatial Stroop tasks, the interfering information need not be currently present to elicit interference; it suffices that it is held in working memory. This is evidenced by the three experiments reported in Chapter 3, all of which demonstrated an interference effect exerted by the to-be-remembered memory item on the judgment task. Further, these experiments very often showed a reciprocal interference effect of the judged item exerted on the memory task. Partially contrary to the conclusions drawn previously, the current work has suggested that the interference which items held in working memory can exert on judgments of items seen in real time may not be a perfect likeness to the simultaneous Stroop effect, specifically if the strength of the encoding of the memory item is low. This is highlighted by the translational model data pattern (interference is stronger when a translation is required from stimulus to response domain) which manifests much less distinctly and completely in Chapter 3's Experiment 1 than in its Experiments 2 and 3. The major difference between these experiments is that Experiment 1 utilised what we believe to be weaker recognition memory, whereas the necessity to recall in Experiments 2 and 3 elicited assumedly stronger encoding, and this is where the Stroop-like interference which we observed was more comparable to that witnessed in simultaneous Stroop tasks. In simultaneous Stroop tasks, where interfering items need not be committed to memory, this effect may be driven instead by object

salience or the strength of representation of items which are visually processed but not encoded into memory per se.

Next, we consider what this work has contributed specifically to the literature on the translational model of Stroop interference. The experiments reported here have provided strong evidence for the notion that multiple responses can be fluently linked, without the need for translation, to one parallel processing system. This is demonstrated by the success of our letter key response in eliciting Stroop-like interference when the to-be judged stimulus is spatial (and thus requires translation) and not when it is verbal (and thus does not require translation), in both the simultaneous Stroop version of the task (online and in-lab) and in the working memory-Stroop hybrid version of the task. Experiment 3 of Chapter 3 using spoken responses confirms in the same paradigm that this type of response is also mapped to the system. In addition, our results suggest that responses which are physically similar can be mapped onto different systems on account of the difference in their meaning. This is demonstrated by the mirroring of the pattern of interference when judging locations compared to judging letters, even when using two responses which are very similar physically: pressing letter keys compared to pressing arrow keys. The differences between the similar responses are the meanings attached to the keys (verbal associations - letters, and spatial associations - arrows).

From a group of experiments which separated the presentation of the interfering stimulus from the to-be-judged stimulus and required that both receive a response of some kind, we found evidence to suggest that not only the process of translation, but also the occupation of both stimuli within the same processing system additionally contributes towards the slowing of responses which characterises Stroop-like interference (this is true at least for the task of judging letters). We found that when both the memory and judgment items required a translation for the required responses to be output, interference observed in the judgment response time data was significantly less than when only one response required a translation. Thus, we concluded that some of the interference we see with one translation is due to the time taken to translate, but some of that interference is also linked with response or decision competition which arises from both items occupying the same processing system. That the extent of interference is less when

the items are in different systems (both responses require translation) demonstrates the size of this extra layer of interference.

Moving into looser theoretical suggestions, an alternative approach to explaining some of these findings is by supposing a selective role of automatic planning of the complementary response to the to-be-ignored item. This hypothesis is less concerned with interference as a whole and more with eliciting specifically the translational model pattern of Stroop interference (high interference when intended response requires a translation and no or low interference when intended response does not require a translation). We have drawn this speculation from the finding in Chapter 3's Experiment 1 that when participants planned a recognition response to the remembered location stimuli, it knocked out the pattern of interference seen in the judgment of letters task. This did not occur in the other direction, with the expected pattern of results being observed when participants judged locations. As discussed above, this could be related to the strength of encoding – if letters are naturally more strongly encoded into working memory than locations, this could account for the difference. Alternatively, if one were to assume that when a stimulus enters its unique processing system it triggers automatic planning of the associated response, the competition which Virzi and Egeth (1985) suggested occurs at the decision stage may actually occur at the response stage when both planned responses clash. It would follow that in Experiment 1 of Chapter 3, we do not see interference manifesting as we expect because the recognition response which participants are planning has overwritten that automatically planned response which would usually provide competition and cause the specific pattern of Stroop-like interference which is typical of simultaneous Stroop tasks.

Since the pattern of results which was expected was only missing when locations were the target of the memory task, we would suggest that the activation of verbal response planning is strong enough to endure the overwriting process caused by planning a recognition response, but the same cannot be said for spatial response planning. It is unclear to us at this moment why that would be the case, but the same criticism can be applied to the alternative theory - why would there be differential strengths of encoding for recognition tasks across domains? Some limited support for this early-stage theoretical notion of differing strengths of automatic response planning is that it does resonate with the suggestion put forward by

Uleman and Reeves (1971) about relative habit strength as an individual difference measure linked to Stroop interference. They reported that the greater a participant's tendency to enact one Stroop-like response over another influences the strength of Stroop-like interference which occurs when these tasks are pitted against each other (in their experiment, this was scanning to find either words or colours).

One might suggest in support of the differing strengths of encoding theory that the strength of the representations of letters could be boosted by a verbal rehearsal mechanism within working memory which cannot be used to maintain locations as effectively as letters and words - a spatial stimulus might be internally verbalised for rehearsal, but we would expect that this would require a translation. Arguably, a verbal rehearsal mechanism might actually be part of or even synonymous with the process of verbal response planning. We sometimes imagine verbal rehearsal as vocalisations carried out within the mind – there might be some role that such a mental vocalisation plays in creating real vocalisations. In fact, there is evidence to suggest that verbal rehearsal could be a good target for attempting to untangle these different theories, given that Chmiel (1984) successfully used both articulatory suppression, which is thought to prevent use of the verbal rehearsal mechanism, and translation to elicit a strong reverse Stroop effect in their colour-word card sort task. This is similar to the multiple layers of interference observed in Chapter 3's Experiment 2. In fact, if participants *are* inclined to transform spatial information into verbal codes for maintenance within the verbal rehearsal mechanism, this might go some way towards explaining the unanticipated interference which we observed in our judgment response time data when no translation was required for the response. If participants were holding verbal versions of the locations in mind, perhaps this was exerting a small amount of interference onto their judgments of letters in the same domain. A simple way to test this could be to utilise articulatory suppression in a similar study to prevent verbalisation of the spatial memoranda. If participants are tending to verbalise the locations, this would likely remove that interference.

Working memory

Equally interestingly to their implications for the translational model, the experiments detailed in Chapter 3 also reflect on the capabilities of working memory. They provide yet more evidence to support the claims that items and action plans

held in working memory have the power to influence the processing of items presented in real time. Thus, our work contributes to expanding on an interesting realm of existing research (e.g., Kiyonaga & Egner, 2014, but also Fagioli et al., 2007; Heuer & Schubö, 2017; Teng & Kravitz, 2019; Trentin et al., 2023).

The experiments detailed in Chapters 3 and 5 also contribute more evidence to the existing commentary on the differences between recognition memory and recall memory (e.g., Carey & Lockhart, 1973; Tversky, 1973; Hall et al., 1976, Uittenhove et al., 2019). In both cases, our results seem to be congruent with the notion put forward by Postman et al., (1948) that recall of information requires a stronger memory trace than recognition. Therefore, it follows that when participants plan to recall information rather than just recognise it, they encode it more strongly. The experiments reported here also provide novel findings related to this subject in that they suggest that this stronger encoding in turn seems to impact the encoded items' ability to interfere with other processed items and also bind to related features, e.g., location.

We think that the results from the experiments detailed in Chapter 4 can be taken to suggest that under the influence of sufficient task entrenchment, action plans can replace some types of information in working memory, likely to conserve cognitive resources. The consistent results of this paradigm across domains demonstrate that the source information (which format they were presented in: coloured square or word) of both the items seen on the surprise trial was lost, even though these pertain to items which were essential to informing the decision about which action to plan (congruent or incongruent response). Research from Henderson et al. (2022) suggests that the method of storage of items in working memory can vary depending on whether the response which will need to be made is known. Information can be stored as sensory information or as action plans, as suggested by the locations of the BOLD signals they recorded in response to different versions of a very simple task: one wherein a response cannot be planned immediately and one wherein a response can be planned immediately, respectively. It would be interesting to measure activity like this during a source or attribute amnesia experiment, to assess whether the above suggestion holds any weight: with sufficient task entrenchment, might participants in this paradigm store information as an action plan rather than as a sensory signal?

One of our major reflections from the work reported here is the importance of understanding the methods we use. While surprise trial paradigms have provided very interesting and also replicable findings about what participants do and do not know during a cognitive task, there is more to be learned about not only the impact of the interruption that surprise trials impose, but also the effects associated with the smaller nuances of how they are presented. Some work on this has already begun, for instance, O'Donnell & Wyble (2023) found that delivery of a cue to retain specific information just before the interruption somewhat reduced the forgetting which their participants experienced compared to when such a cue was not delivered. Meanwhile, our Experiment 3 in Chapter 4 wherein participants failed to recall even the item information (which was preserved in Experiments 1 and 2) suggests that inundating participants with too many response options and/or the specific nature of the question posed may also have an impact on what information is inaccessible by the end of the surprise trial. It would also be very valuable for the continued use of the paradigm going forward to know whether the experiment can still validly be run when implementing more than one surprise trial, as in the current version of the task, interpretation depends on performance on a single trial.

Feature binding

The exploratory analyses of the new drag and drop response method of testing memory for verbal-spatial feature bindings reported here suggest that previous studies reporting binding asymmetry may have failed to detect an important effect of task order. Our data indicate that binding occurs to a much greater extent when participants intend to commit letter identities to memory as the first of the two tasks they undertake. This is compared to when this task occurs second, or when they are tasked with committing locations to memory (at any stage in the experiment). Something about experiencing the letter task at the beginning of the experiment seems to facilitate participants to take on unnecessary additional information much more often than in other cases. Elsley and Parmentier (2015) concluded from their study that this difference in binding was likely due to a *strong asymmetry* wherein feature types (e.g., space, colour, shape, etc.) bind to one another selectively as a result of their relative positions within a feature hierarchy.

They do not suggest why a strong binding asymmetry would only apply within the first half of the experiment, so below we make some suggestions based on the data.

Acting *in addition* to a strong asymmetry, the short interruption of the break between blocks might be considered a candidate for reducing binding in the current study. This mechanism could be uniquely positioned to contribute to the explanation of the drop in binding following the letter task when it is first and also the suppression of binding in the letter task when it is second. If we assume that there is a diminished tendency to bind letters to locations than locations to letters anyway, as put forward by the strong asymmetry hypothesis, that theory can explain the overall difference between binding in the letter task compared to the location task. However, it has no explanatory power to account for why the letter task when it is experienced second shows much less binding than the letter task when it is experienced first. It also cannot explain the comparatively smaller reduction in binding in the location task as a result of task order. This is where the proposed effect of taking a break may be valuable. If it can be demonstrated that the simple act of interrupting the task exerts an additional reduction effect on the occurrence of binding, this may account for all of the results shown here not currently explained by asymmetry. This could be easily tested by comparing binding in two blocks of letter focus trials experienced back-to-back with only a short break in between. Would participants continue to bind to the same extent throughout both blocks? Alternatively, this proposed effect might be tied specifically to task switching, in which case, we would not observe a reduction in binding across two blocks of the letter task, but a more complex design implementing more than two blocks could feasibly be used to tease this apart (assuming that binding continues to drop with every instance of task switching and not just the first).

Instead of a strong asymmetry, this increased binding in the letter task could be a result of it being the easier task (though only marginally, given the extremely high and similar accuracy results across tasks), which affords participants the extra capacity to take on unnecessary location information. On the other hand, the more difficult and demanding task of remembering locations may not allow for the storage of additional unnecessary information. However, it still remains unclear why we would only see that when the letter task is first and not when it is second, so a second factor is still required. A straightforward suggestion for why a dependent variable may diminish over time is fatigue - perhaps participants are affected by

fatigue in the second half and are less likely to commit more information to memory than is required, even when the task is easy. However, if participants become fatigued as a result of taking part in the location task, we are unsure why this would not also manifest in a decline in binding over the course of the location block. Our correlational analysis results do not provide any evidence for a decline in binding over the course of any block of trials within the current experiment. In fact, they indicate a small but significant *increase* in binding over the course of the location block when it is experienced first – more likely to be evidence of a practise effect than one of fatigue. If, in spite of this finding, there actually is an effect of fatigue in play over the course of this experiment, it is possible that a block of 40 trials is too few in which to reliably detect the effects within the task. This could easily be addressed by substantially increasing the number of trials within each block. If participants continue until the end of a block of 80 location focus trials without giving evidence of a decline over time (which is the total length of the current experiment, so should elicit some fatigue-based decline if that is indeed the issue), one might conclude that fatigue is not a factor which influences binding within the scope of this experiment, and begin to look for other candidate effects.

In place of fatigue, this suggested effect on binding over time could easily be thought of as a result of diminished availability of cognitive resources, as outlined in Popov and Reder's (2020) Resource Depletion theory of working memory. This model suggests that we have a finite cognitive resource, a small amount of which is used for every cognitive operation which we enact (encoding, feature binding, etc.), and which slowly recovers over time. When we have less resource available, less of it is dedicated to encoding new items, which results in these newly encoded items being harder to recall. These findings are quite well explained by the model, because it may also account for why a more difficult task would have a stronger detrimental effect on performance than an easier task (by using up more resource faster). One limitation of the application of this model to the current data however is that we do not see how this mechanism of effect would explain the lack of a detectable decline in binding within a single block: whether the effect is fatigue or depletion of resources, we still think that we would expect to observe a downwards trend in binding.

Reflecting on a Domain Hierarchy

The domain hierarchy that Elsley and Parmentier (2015) refer to as the strong asymmetry hypothesis has intrigued us throughout the course of this thesis work. Elsley and Parmentier only apply it to the process of feature binding, suggesting that some feature types will naturally bind to some but not other feature types (i.e., locations bind to letters, but letters do not bind to locations) depending on their relative locations within said hierarchy. However, we intend to reflect on the work conducted here which investigates the interplay of verbal, colour and spatial information, and consider whether such a hierarchy may exist beyond just feature binding.

Experiments dating all the way back to Stroop's original paper (Stroop, 1935) have demonstrated that all other things being equal, the forward Stroop effect is typically larger than the reverse Stroop effect (see Chapter 3 for a fuller discussion). That is to say that ignoring incongruent colour words while judging text colour is harder than ignoring text colour when judging the meaning of colour words. However, the experiments detailed in Chapter 4 suggest that these same stimuli are not unequally salient in the Source Amnesia paradigm. This is demonstrated by the very similar error rates when participants were asked to remember coloured squares compared to when they were asked to remember colour words. It appears that colour words' stronger capacity for interference does not also afford them more memorability when participants do not know they will have to remember them.

On the other hand, we were beginning to think that spatial and verbal information were on a more equal playing field than colour and verbal information, given the ease with which bi-directional interference can be elicited using verbal-spatial Stroop stimuli as demonstrated in Chapter 2. However, we still detected some evidence of inequality between these stimulus types when an element of memory maintenance was introduced. In Experiment 1 of Chapter 3 where the working memory-Stroop hybrid tasks utilised weaker recognition-style memory, data from the location judgment trials conform to the patterns laid out by the translational model, but data from the letter judgment trials do not, seemingly because they require stronger encoding to do so (as in Experiments 2 and 3 of Chapter 3 which utilised recall-style memory). This asymmetry maybe suggests that verbal information (which

is held in mind in those location judgment trials) is naturally stronger in its capacity for interference than spatial information (which is held in mind during letter judgment trials), even at the low level of encoding strength associated with recognition memory. That is to say that verbal information held in memory interferes more strongly with judgments of locations than vice versa, demonstrating an inequality in the working memory-Stroop hybrid version of the task which is not seen in the simultaneous spatial Stroop task (Chapter 2). Further to this demonstration of stimulus inequality, the data in Chapter 5's exploratory memory binding study once again provide support for the notion of binding asymmetry, that we can encode a letter and get its location 'for free' very often, but much less frequently do we see the reverse. This is further evidence that there is an inequality between these stimulus types which may be dependent on participants' intention to remember them.

It seems from this dissection that in the results presented here, visually presented verbal information is only dominant over other types of visual stimuli when we intend to remember it (as is the case in Chapters 3 and 5). When we do not intend to remember it (as in Chapters 2 and 4), verbal information acts the same as other information types. This could be considered good support for unique verbal memory mechanisms or verbal-visual domain separation in working memory models. However, there is a caveat to this conclusion which we have already alluded to: it is well-established in Stroop research conducted by many researchers over the years that verbal information *is* dominant over colour stimuli even when no memory is required, because reverse Stroop effects tend to be smaller (all else being equal) than regular Stroop effects. That is to say that verbal information has a stronger capacity for interference on colour information than vice versa without other input, e.g., articulatory suppression, stimulus obscuring, etc. So, even when memory is not required, verbal information still holds some power over colour. In this particular case, we suggest that this inequality could be due to a *weakness* that colour information has, rather than any *strength* that verbal information has. This weakness is that colours cannot be expressed by humans without verbalisation or reference to an example of a colour in the environment, which might require verbal and spatial translations respectively, and causing an atypically unbalanced relationship regarding Stroop interference.

Conclusions

To conclude, the work described in this thesis has approached the topics of Stroop-like interference, working memory and feature integration, and within those realms, has attempted to shed light on the interplay of verbal and visual information using contemporary methods. We have frequently contrasted the effects on performance within these cognitive phenomena of the intention to recognise compared to the intention to recall information, and also the differences between recall responses with different domain associations. It appears that visually presented verbal information is more likely to assert dominance over other visual stimuli (manifesting as interference and incidental binding) when participants are intending to commit the stimuli to memory. When this is not the case, verbal information is typically treated equally to other stimulus types within our minds, however there are exceptions to this when other response-related elements are brought into play, such as the need to translate the stimulus code into a different domain for appropriate response output.

References

Carey, S. T., & Lockhart, R. S. (1973). Encoding differences in recognition and recall. *Memory & Cognition, 1,* 297-300. https://doi.org/10.3758/BF03198112

Chmiel, N. (1984). Phonological recoding for reading: The effect of concurrent articulation in a Stroop task. *British Journal of Psychology, 75,* 213-220. https://doi.org/10.1111/j.2044-8295.1984.tb01894.x

Elsley, J. V. & Parmentier, F. B. R. (2015). Rapid Communication: The asymmetry and temporal dynamics of incidental letter–location bindings in working memory. *The Quarterly Journal of Experimental Psychology, 68*(3), 433-441. https://doi.org/10.1080/17470218.2014.982137

Fagioli, S., Hommel, B., & Schubotz, R. I. (2007). Intentional control of attention: Action planning primes action-related stimulus dimensions. *Psychological research, 71*, 22-29. https://doi.org/10.1007/s00426-005-0033-3

Hall, J. W., Grossman, L. R, & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4* (5), 507-513. https://doi.org/10.3758/BF03213211

Henderson, M. M., Rademaker, R. L., & Serences, J. T. (2022). Flexible utilization of spatial- and motor- based codes for the storage of visuo-spatial information. *eLife, 11,* e75688. https://doi.org/10.7554/eLife.75688

Heuer, A. & Schubö, A. (2017). Selective weighting of action-related feature dimensions in visual working memory. *Psychonomic Bulletin & Review, 24,* 1129-1134. https://doi.org/10.3758/s13423-016-1209-0

Kiyonaga, A., & Egner, T. (2014). The working memory Stroop effect: When internal representations clash with external stimuli. *Psychological science, 25*(8), 1619-1629. https://doi.org/10.1177/0956797614536739

O'Donnell, R. E., & Wyble, B. (2023). Slipping through the cracks: The peril of unexpected interruption on the contents of working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(6), 990–1003. https://doi.org/10.1037/xlm0001214

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127(1),* 1–46. https://doi.org/10.1037/rev0000161

Postman, L., Jenkins, W. 0, & Postman, D. L. (1948). An Experimental Comparison of Active Recall and Recognition. *The American Journal of Psychology, 61*(4), 511-519. https://doi.org/10.2307/1418315

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Teng, C. & Kravitz, D.J. (2019). Visual working memory directly alters perception. *Nature Human Behaviour, 3,* 827–836. https://doi.org/10.1038/s41562-019-0640-4

Trentin, C., Slagter, H. A., & Olivers, C. N. (2023). Visual working memory representations bias attention more when they are the target of an action plan. *Cognition, 230*, 105274. https://doi.org/10.1016/j.cognition.2022.105274

Tversky, B. (1973). Encoding Processes in Recognition and Recall. *Cognitive Psychology, 5*, 275-287. https://doi.org/10.1016/0010-0285(73)90037-6

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is Working Memory Storage Intrinsically Domain-Specific? Journal of Experimental Psychology. *General, 148*(11), 2027–2057. https://doi.org/10.1037/xge0000566

Uleman, J. S. & Reeves, J. (1971). A reversal of the Stroop interference effect, through scanning. *Perception & Psychophysics, 9,* 293-295. https://doi.org/10.3758/BF03212651

Virzi, R. A. & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition, 13*(4), 304-319. https://doi.org/10.3758/BF03202499