THE PUPILLOMETRIC PRODUCTION EFFECT:

MEASURING ATTENTIONAL ENGAGEMENT DURING A PRODUCTION TASK

by © Hannah V. Willoughby

A thesis submitted to the School of Graduate Studies

In partial fulfillment of the requirements for the degree of

Masters of Science in Experimental Psychology

Department of Psychology

Faculty of Science

Memorial University of Newfoundland

October 2020

St. John's Newfoundland and Labrador

Abstract

Reading to be remembered material aloud improves memory at test compared to material read silently; a finding termed “the production effect”. Aloud words are theorised to be better remembered because they stand out and benefit from distinctive processing during which unique features of the words (i.e., visual, phonological, semantic) are attached to the encoding trace. Recent accounts challenge this notion suggesting that an attentional or effortful component contributes to the effect. The following applies pupillometry to measure cognitive effort during a production task; the pupils dilate when one is engaged in an effortful task (e.g., solving a difficult math problem) compared to a less effortful task (e.g., solving an easy math problem). Across two experiments, I separated components of the production task (e.g., presenting the stimulus separately from speaking) to provide insight into the underlying active cognitive processes. In Experiment 1, a pre-cue was implemented, participants viewed the instructional cue (e.g., read aloud) prior to viewing and producing the stimulus. In Experiment 2, participants viewed the word and instruction simultaneously and waited until a later “Go” cue to speak. Both experiments demonstrated evidence showcasing the importance of distinctiveness and also the influence of other cognitive processes within the production paradigm.

Keywords: memory, recognition, production effect, pupillometry

Co-authorship Statement

This project was completed as a follow-up to a study previously completed as a directed studies project by Jenny Tiller, cited in-text as Tiller, Hourihan and Fawcett (2018). As such, Experiment 1 was implemented and pilot data collected by the authors of the previous study. The changes in Experiment 2, changes to programming and all data collection (apart from help by volunteers acknowledged) were completed by the author.

General Summary

Many people try to repeat or read things aloud to memorize them. Indeed, saying things aloud does improve memory for those items–this phenomenon is known as the production effect. Researchers proposed that aloud words are remembered better than silent ones because aloud words are unique compared to the silent ones. This notion has been challenged by suggestions that speaking aloud might increase effort or attention to aloud words. I assessed this possibility by measuring pupil size as an index of cognitive effort. I modified the production paradigm by separating the steps involved in the production task in two ways: (1) In Experiment 1, I added a pre-cue where participants viewed the instructional cue prior to speaking aloud; and (2) In Experiment 2, the cue to speak aloud was delayed. My results suggest an additional process facilitates the production effect which will be detailed in the following text.

Acknowledgements

I would like to thank both of my wonderful supervisors for their continued guidance and support. Notably I would like to thank Jon for encouraging me to challenge myself and to Kathleen for helping me move forward at times that I was feeling overwhelmed. I would also like to extend my gratitude to Aime Hirwa and Andy Moldwin for their aid in data collection. Furthermore, I would like to express my appreciation to my partner Tyler for his continued support. Being confined together through an unprecedented snowstorm and global pandemic while writing my thesis surely revealed the strength of our relationship. Lastly, it would be remiss of me not to mention my parents who have always supported me and my two cats, Paul and Jade, who put a smile on my face every day.

Table of Contents

List of Figures

Chapter 1: Literature Review

When people deliberately try to remember information, whether it is a birthday or information from class that will appear on a test, they use a variety of study techniques. For example, self-report measures from university students suggest that within the university environment, they will often reread passages from their study notes and that university instructors reinforce these habits (Morehead, Rhodes, & DeLozier, 2016). But is this an effective strategy? Indeed, research suggests that reading material when accompanied by some form of production (e.g., reading aloud, writing, etc.) does improve memory relative to material that is not produced (Ozubko, Hourihan, & MacLeod, 2012). At first glance this phenomenon – titled the production effect – seems straightforward. However, there are complex cognitive processes underlying this seemingly simple effect. The dominant explanation of the production effect is distinctiveness; that is, the produced materials are remembered better because they have extra features (e.g., motoric, auditory) that the silently read words do not (Hunt, 2006; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Although there is a great deal of support for this explanation, there is also reason to believe that attention and/or effort contribute to the enhanced memory performance. In the following, I will outline research on the production effect, describe the evidence for or against a distinctiveness explanation, and discuss how attention or effort may play a role. Then, I will describe the advantages of using pupillometry (an unobtrusive psychophysiological technique) as a proxy to measure attention or effort. The ultimate purpose of this thesis is to closely examine the various cognitive processes used in this memory paradigm to inform the use of the production effect as a device to improve memory performance.

The finding that reading words aloud resulted in a memory advantage over words that were read silently was first recorded by Hopkins and Edwards (1972). The authors observed that recognition for words said aloud was better when the aloud and silent words were intermixed, but not when compared across two separate pure-lists made entirely of either condition. In the following decades several other researchers observed this effect, sometimes inadvertently when production was used as an experimental manipulation to study other cognitive processes, such as attention and modality effects in long-term memory (Conway, & Gathercole, 1987; Gathercole & Conway, 1988; MacDonald & MacLeod, 1998). The effect itself did not gain traction nor its name until MacLeod, et al. (2010) later elaborated on previous findings by completing a series of experiments designed to identify the breadth and boundaries of the effect. The structure of the studies completed by MacLeod, et al. (2010) is considered typical for the paradigm, whereby a list of words was presented one at a time simultaneously with a cue (e.g., a symbol or colour) indicating whether the cued word was to be read aloud or silently. The study list was then followed by one of two memory tests: (1) a recognition test, where participants were presented with a list of words, half which were previously studied half which were not; (2) a recall test where participants must remember as many words as possible without a cue. Since their initial examination, the production effect has been replicated by many authors and shown to improve memory in diverse domains such as visual imagery (Fawcett, Quinlan, & Taylor, 2012; Hourihan & Churchill, 2019), longer texts (Ozubko, Hourihan et al., 2012), and across age groups (Lin & MacLeod, 2012). Indeed, the effect has been applied to practical domains such as within educational contexts, and the memory benefit for items studied *aloud* has been demonstrated to sustain itself when the interval between study and test was extended by a week (Ozubko, Hourihan et al. 2012). Considering the results of these experiments it was proposed that the production effect was the outcome of the aloud words standing out against the backdrop of the silently read ones, resulting in the former items being more distinctive within memory (Hunt, 2006; MacLeod et al., 2010). The following sections will describe evidence for – and against – this perspective as a prelude to the present studies.

1.1 Distinctiveness in Production

Distinctiveness is a mental process whereby unique items stand out against common ones and therefore, are processed differently; this differential processing benefits later recollection of those unique items (Hunt, 2006). Importantly, memory improvement is not the result of the exceptional features of those items, but rather of a specific type of cognitive processing used, which facilitates later retrieval of those items. According to Hunt (2006), to initiate distinctive processing there must be some form of difference or salience within the items that are processed distinctively: they must stand out in some way. Within the framework of the production effect the predominant theory is that the aloud items stand out against the silent items because they are encoded with additional information specific to those words, such as the unique phonemes, auditory, visual and semantic traits that are encoded when a word is said aloud (MacLeod et al., 2010; Ozubko & MacLeod, 2010). The distinctive processing which occurs in response to the aloud items, incorporates those word's distinctive features into memory, thereby improving later performance.

The notion that distinctiveness was a central feature of the production effect was evidenced by three key findings in MacLeod et al.'s (2010) study. First, producing the same response for all produced words (e.g., saying "yes" for all the aloud items rather than reading the specific word aloud) did not result in a memory benefit for those items, indicating that the verbal production itself had to be unique to encode the distinctive information of the words. Second, the production effect was found even for nonwords, showing that production did not necessitate the study items be meaningful as long as they were unique. Third, the production effect occurred when words were already strengthened by virtue of generation (a process where participants must come up with the word themselves from a definitional cue). This finding demonstrated that production was additive to words that had already been strengthened, suggesting that production did not just improve memory by increasing encoding strength. Rather, it was thought that the additional distinctive information (e.g., motor and auditory features) was incorporated into the memory trace during encoding that increased memory for the item. In summary, these results supported the notion that the distinctiveness account was consistent with the memory benefits observed as a result of production.

In addition to reading words aloud activating distinctive processing, participants might also use the "distinctiveness heuristic", whereby distinctive information encoded during study is used strategically, to inform their decisions at test (Dodson & Schacter, 2001). During encoding, participants may begin to incorporate some of the distinctive elements of the study phase into their memory; specifically, they may encode an element associated with the motoric articulation itself. Later, at test participants may think back to their subjective experience at encoding and use this "production trace" (Fawcett, 2013) to identify whether they remember saying the word aloud at study. If successful in retrieving this trace, then it can be heuristically inferred that the word appeared on the study list (Dodson & Schacter, 2001). Supporting this notion, producing words is ineffective at improving memory on implicit memory tasks (e.g., speeded reading), which do not require consciously assessing whether a word had been studied (Hourihan & MacLeod, 2008; MacDonald & MacLeod, 1998; MacLeod et al. 2010). These observations lend support to the suggestion that participants are able to access the production trace at test and indicate that production is an effective mnemonic device only if participants are actually able to access it.

Further evidence for the use of a distinctiveness heuristic during production includes studies using source or list discrimination tasks. For example, source discrimination tasks – where at test participants identify whether a given word at study was read aloud, silently or if it is a new word – evidenced that participants had a higher likelihood of correctly identifying the mode of presentation for words that had been read aloud (Conway & Gathercole, 1987; Ozubko, Gopie, & Macleod, 2012). This held true even when silent words were strengthened by repetition (Ozubko, Major, & MacLeod, 2014). These findings suggest that participants are successful at accessing distinctive information attached to the encoding trace such as motor, auditory, or semantic information, at test and use it to discriminate between test items. Further evidence comes from list-discrimination tasks where participants are shown two separate lists at study and then at test must indicate which list a given word was from. When a mixed-list of items is followed by a pure-aloud list, accuracy for source discrimination is lower than when the mixed-list is followed by a pure-silent list. Supposedly, reading the second list aloud renders the aloud items on the first list no longer useful as a diagnostic tool at test, since these aloud items cannot be parsed from those read aloud on the second list (Ozubko & MacLeod, 2010). Participants' success at these types of tasks implies that they are encoding distinctive information at study and that they are able to access this information and perhaps use it strategically to inform their decisions at test. Evidence that participants are strategically utilizing distinctive information supports the distinctiveness theory and more specifically the distinctiveness heuristic.

An extension of the evidence for distinctiveness in production came from Forrin, MacLeod and Ozubko (2012), who demonstrated that the effect persisted across different modes of production (reading aloud, mouthing, writing and whispering). Moreover, they observed that different modes of production resulted in differing magnitudes of the effect. For example, reading words aloud resulted in better memory performance than when words were whispered, written, or mouthed. These results replicated a previous finding by Conway and Gathercole (1987) who also observed an intermediate production effect for whispering. Because words that are read aloud utilize two forms of additional processing—motor articulation and auditory feedback—it is implied that they benefit from additional distinctive information that is encoded within the two forms of sensory feedback (Conway & Gathercole, 1987; Forrin et al., 2012). In a similar vein, Quinlan and Taylor (2013) lent further support for a distinctiveness account by obtaining a larger production effect for words that were sung compared to words read aloud; it was surmised that singing the word results in more distinct encoding elements than speaking. The production effect appears to be additive and the magnitude of the effect may vary depending on the number of distinct processes which occur during encoding (Forrin et al., 2012; Ozubko & MacLeod, 2010; Quinlan & Taylor, 2019). For example, when one is singing they are engaged in motor articulation, auditory processing and changing qualities of their voice such as their tone and pitch; when one is talking they are still engaged in motor and auditory function but there is less uniqueness in their vocal qualities (Hassall, Quinlan, Turk, Taylor, & Krigolson, 2016; Quinlan & Taylor, 2013). Additionally, Jamieson and Spear (2014) demonstrated evidence for the offline production effect – better memory for words that were imagined to be typed than silently read ones. This effect was also smaller compared to the magnitude of the effect when words were actually typed out. The results from these studies demonstrated that distinctive processing can be additive and the amount of distinct processing that occurs can influence the effectiveness of production.

1.2 Theoretical and Empirical Challenges

Although substantial evidence points towards distinctiveness as a fundamental process in yielding the production effect, not all experimental observations are compatible with a distinctiveness framework. At first the observation that the production effect was only observed in within-subject designs (experiments where participants view a mixed-list of both words to be produced or read silently) was thought to be a cornerstone of the effect. This type of design aligns with a distinctiveness account – words said aloud would stand out compared to the words that were read silently causing the memory benefit (Hopkins & Edwards, 1972; Macdonald & MacLeod, 1998; MacLeod et al., 2010). In a between-subjects design participants are only exposed to one condition and therefore this design does not fit into a distinctiveness framework (Bodner & Taikh, 2012; Bodner et al. 2016; although see Jamieson, Mewhort, & Hockley, 2016). Indeed, attempts to resolve the effect in between-subjects designs have historically failed to produce any difference (Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010). Nonetheless, this claim was recently challenged by a series of meta-analyses showing consistent evidence for a between-subjects production effect (e.g., Bodner, et al., 2014; Fawcett, 2013). Notably, it was revealed that using a between-subjects design had a moderate effect size and showed a reliable trend of hits for aloud words being greater than those for silent words.

Following the initial evidence for a between-subjects production effect, the finding has been replicated successfully by multiple researchers (Bodner, Jamieson, Cormack, McDonald, & Bernstein, 2016; Bodner et al., 2014; Forrin, Groot, & MacLeod, 2016; Forrin, Ralph, Dhaliwal, Smilek, & MacLeod, 2019; Taikh & Bodner, 2016). Although, note that while the within-subject effect is robust across both recognition and recall designs the between-subjects production effect has only been observed in recognition studies (Fawcett, 2013; Forrin & MacLeod, 2016;

MacLeod & Bodner, 2017). This discovery opened a dialogue that the production effect might not be solely driven by distinctiveness and that a second mechanism may also contribute to the effect (Fawcett, 2013; Fawcett & Ozubko, 2016; Ozubko, Gopie, & MacLeod, 2012). For example, saying the word aloud might strengthen the memory trace of the word (Bodner & Taikh, 2012). This notion is closely related to the concept of level of processing, introduced by Craik and Lockhart (1972) who posited that a deeper processing of an item (e.g., deciding if the word “cow” is semantically related to the word “grass” as opposed to silently reading the word “cow”) would strengthen the encoding process and create a stronger memory trace of the word. Considering this, the strength account posits that because aloud words are more strongly encoded, they will be more likely retrieved at test. A second cognitive mechanism proposed to play a role in increasing memory performance in the production paradigm is attention (Bodner, Taikh, & Fawcett, 2014; Fawcett, 2013; MacDonald & MacLeod, 1998; Ozubko, Gopie et al., 2012). Perhaps saying the word aloud allocates more attentional resources to the word thereby strengthening the encoding trace or facilitating the effectiveness of distinctive processing. In the text that follows, studies that observed results that do not fit into the traditional framework of the distinctiveness account, or that supported a strength or attentional account, will be discussed. Finally, I will discuss how psychophysiological measures, such as pupillometry, have been applied to examine cognitive processes and discuss how similar techniques can be used to observe attentional and strength processes during production.

1.3 Evidence for Alternative Types of Processing

1.3.1 Evidence that production increases attention or encoding strength. Multiple researchers have suggested that the greater magnitude of the production effect in a within subject design is the result of a cost incurred to the silent items when presented in a mixed-list (Bodner et al., 2014; Hopkins & Edwards, 1972; Jones & Pyc, 2014; Jonker, Levene, & MacLeod, 2014; Lambert, Bodner, & Taikh, 2016). In free-recall studies, accuracy for silent words was demonstrated to be lower in a mixed-list design when compared against silent items in a pure-list design (Lambert et al., 2016; Jones & Pyc, 2014). Surprisingly, comparing aloud words in this manner revealed no performance differences between the two lists (but see Jonker et al., 2014). This evidence suggests that the production effect in within-subject designs, using a mixed-list is driven by a cost incurred to the silent items rather than a benefit to produced items within free-recall studies. Both Jonker et al. (2014) and Lambert et al. (2016) suggested that during studies using a within-subject free-recall procedure, participants engage in both item-specific and item-order processing for studied items. Item-specific processing is activated by unusual stimuli (e.g., aloud words) which engages encoding of information specific to the features of an individual item (i.e., visual, phonological, semantic features). In contrast item-order processing is the encoding of relational information, such as the order of word presentation (e.g., remembering that the word ‘dime’ followed the word ‘toast’). In within-subject recall production studies enhanced item-specific processing for aloud words interferes with item-order processing; since increased item-specific processing for aloud words reduces the amount of item-order processing that occurs overall. Silent words are particularly affected by a reduction to item-order processing since they do not benefit as much as aloud words from item-specific processing but only when aloud and silent items are presented together. Additionally, some of these same studies determined that the cost to silent items is actually greater than the memory boost for aloud words (Jones & Pyc, 2014). The notion that the production effect may be caused in part by a cost to silent items rather than a boost to aloud items poses a challenge to the distinctiveness account of production which predicts better memory for aloud items due to advantageous distinctive processing; this account does not predict that distinctive processing would lessen encoding for silent words (cf. Macleod et al., 2010).

Bodner et al. (2014) again demonstrated a cost to silent items in a within-subject design, but this time using a recognition paradigm rather than free-recall, confirming that the cost to silent items discovered in studies using free-recall at test extended to those using recognition as well. They determined that in a mixed-list design, the measure of sensitivity (*d′*) for silent items was lower than the same measurement for silent items in a pure-list design. This suggests that the presence of aloud words decreases sensitivity for silent items in a mixed-list design meaning that participants are less able to identify silent words during recognition in a mixed-list compared to a pure-list design. The same authors obtained a significant between-subjects production effect by comparing *d′* across the two pure-lists finding that sensitivity was greater for aloud words than silent words. Finally, they observed that sensitivity for aloud words in the mixed-list improved when compared against the pure-silent list, confirming that reading words aloud on the mixed-list did improve memory for aloud words. However, the same authors did not observe improved sensitivity for aloud items in the mixed-list when compared to the pure-aloud list, providing concrete evidence that there is a memory benefit from production, even after factoring in costs. Although these results also revealed that production might not be particularly useful when used in a mixed-list context – if the goal is to remember all list items – since improved memory for the aloud items may be at the detriment to memory for silent items.

In a subsequent study Forrin et al. (2016) followed up on the previous study by Bodner et al. (2014). They argued that comparing *d′* across different experimental designs was problematic since it is impossible to separate false-alarms (FA) in a mixed-list design (i.e., having both a mixed-aloud and mixed-silent FA rate), an issue previously discussed by Fawcett et al. (2012).

By not being able to separate the FA rates for the two mixed list conditions, there is no way of knowing if the participant thought the FA was a word they had read aloud or had read silently. Therefore, the measure used by Bodner et al. (2014) was biased since it assumed an equal probability for a FA being incorrectly identified as a word from either condition. Forrin et al. replicated the previous study using three different designs. First, they used a blocked design where the same participant viewed all three list designs (i.e., mixed-list, pure-silent, pure-aloud) replicating Bodner et al.'s (2014) previous findings using a similar d' measurement. Next, in their second experiment they altered the design by having participants make modality attributions at test (i.e., asked participants if they studied the word aloud, silently or if it was new). Using this method, they were able to obtain separate FA rates for the mixed-aloud and mixed-silent words. Finally, in their last experiment they followed the Experiment 2 protocol but separated the test following the mixed-list into two response options (e.g., aloud-new, silent-new) so that a cost-benefit analysis could be completed. By examining the dissociated FA rates in Experiments 2 and 3 the authors demonstrated that words that were read aloud in the mixed-list were better remembered than those read aloud in the pure-list, evidenced by a higher d'. However, they did not observe this advantage in Experiment 1 when the FA rates were not dissociated. Instead, this evidenced the notion that new items are not equally likely to be identified as aloud or silent; FA rates were more likely to be identified as silent items than aloud ones when separate FA rates were used but using separate FA rates did not impact the observed cost to silent items in the mixed-list. It should be mentioned that their approach fundamentally changed the design of the production experiment and while providing evidence towards differential FA rates, the results might be different in an unmodified production study.

The discrepancy in results between studies that have found a cost to silent items and no benefit to aloud items (e.g., Jones & Pyc, 2014; Lambert et al. 2016) and those that have shown improved sensitivity for aloud items (e.g., Bodner et al., 2014; Forrin et al., 2016) is likely determined by the type of test used to asses memory performance. The set of studies that showed no advantage for aloud words used a free-recall testing procedure, whereas those studies that showed increased sensitivity for aloud words used a recognition test. Recollection and recall depend on different processes to elucidate memory. In free-recall where there is no cue at test to guide memory processes, participants rely more on relational processing – a process where information about words in relation to other words that appeared on the list is encoded (Lambert et al., 2016; Jones & Pyc, 2014). At test relational information can benefit memory by acting as a cue for related words. For example, the previously studied word "cow" might cue the word "grass" as these two words are semantically related. Conversely, in recognition studies a target word is provided, therefore, relational information is less important. Instead memory depends more on item-specific processing, the encoding of specific information about the stimulus itself (e.g., visual, semantic, phonetic, etc.) so that it can be compared to the test word. This is evidenced by the sensitivity scores being used to guide judgement for aloud words in recognition studies. Unlike studies examining the production effect in recognition, there is presently no evidence supporting the effect between subjects in recall (Fawcett, 2013; Forrin & MacLeod, 2016; MacLeod & Bodner, 2017). It seems that production might only be beneficial for memory when used in a between-subjects design followed by a recognition test.

Many researchers have questioned if the cost to silent items is due to disengagement with those items or participants viewing them as less important. This notion is referred to as the "lazy reading hypothesis" (Begg & Snider, 1987; MacLeod et al., 2010). However, there are mixed findings that both support and refute the hypothesis. Contradictory evidence suggests that participants are engaging with silent items to the same degree as aloud ones. For example, using a self-paced study design, where participants moved on to the next study word at their own pace, Lambert et al. (2016) found no difference between the average duration of study time for silent and aloud words. If participants regarded the silent words as being less important, it would be expected that they would spend less time studying those words. Additionally, participants who read the pure-silent list actually took longer than those who read the pure-aloud list. In a similar vein, studies that have strengthened encoding for silent items by adding a deeper processing manipulation (e.g., generation, imagery) contradict a lazy reading stance as the expected production effect remains intact even when silent items are strengthened (Forrin, Jonker, & MacLeod, 2014; MacLeod et al., 2010). Likewise, in a blocked list design where the first half of the list is read silently and the second half is read aloud (or vice versa), the cost to silent items is reduced (Bodner et al., 2016). It seems that the cost is incurred by differences in encoding for silent words that are presented in a mixed-list rather than differences in engagement. This observation refutes the notion that participants see silent words as being less important or are taking less time to encode them.

Nevertheless, there is still conflicting evidence suggesting that participants might not attend to aloud and silent words equally: On subjective reports of attentional allocation at test participants report more engagement with aloud items (Fawcett & Ozubko, 2016) and are less likely to mind-wander while reading aloud than reading silently (Varao Sousa, Carriere & Smilek, 2013). Similarly, Ozubko, Bamburoski, Carlin, and Fawcett (2020) included a third study condition of silent items that participants were told were “more important” than both aloud and other silent items. This procedure resulted in superior performance for important silent items over non-important silent items; indeed, performance was comparable to that observed in the aloud condition. Although there is no concrete evidence to support the lazy reading hypothesis per se, there is evidence that points towards different levels of attention and motivation between trials where participants speak aloud and those where participants read silently, which could potentially result in a cost to memory for the silent items.

The assumption that producing words might influence attentional processes was considered by MacDonald and Macleod (1998) who sought to examine the effect of differing attention on direct and indirect memory. Across a series of three experiments attention was manipulated by having participants read words aloud as a means of increasing attention to those items versus reducing attention by having participants "ignore" the other items. In the first experiment participants were asked to say "pass" for items that were in the ignore condition. Later, in the second experiment the same methodology was used except participants were no longer required to make an overt response to the unattended items. Finally, in the third experiment participants were shown word pairs, with one item being the "attended" item to be said aloud and the other to be ignored. Note that across all three experiments participants were given no explicit instructions as to whether they should read the unattended word nor were they instructed to ignore it; they were unaware that the word list would be followed by a test.

In each experiment the study list was followed by an implicit test and an explicit test. In the implicit test participants read the words aloud as quickly as possible; the implication was that words that had been "primed" (previously presented) would have a lower latency period before speaking than those that were not primed. Therefore, if participants were really not attending to the "ignored" words then there should be no difference in latency between silent words and new ones. The second, explicit memory test was a traditional recognition test format. The first two experiments revealed similar results on the implicit test; the unattended words did prime memory as evidenced by faster reaction times for both attended and unattended items compared to new words. However, there was no priming effect for the ignored words when they were presented as a word pair. The explicit memory test showed better memory for the words that were read aloud – demonstrating a production effect - across all three studies. Interestingly, production only influenced memory on the explicit memory test. MacLeod et al. (2010) posited that production had no effect on the implicit memory test because without having time to make a conscious decision at test, participants would not be able to utilize distinctive information encoded during the study phase to discriminate between items. However, Macdonald and MacLeod (1998) also attributed the results of the explicit memory test to the increased attention that was used to encode the words that were said aloud. This study lent support for the theory that speaking aloud increased participants' focus on the word resulting in a stronger encoding trace of that word.

Further work examining the potential impact of attentional processes on the production effect were examined when the task was completed with background noise. Mama, Fostick, and Icht (2018) had participants complete a typical production task under different noise conditions. These included a steady-state noise condition (i.e., energetic noise presented in normal speech range); a babble condition (i.e., nonsense language syllables); and a silent condition where no noise was presented. In all noise conditions noise was played through headphones so that auditory feedback from production could not be heard by the participant. The production effect was observed in the no noise and the steady-state noise conditions. However, babble noise eliminated the effect rendering similar performance for aloud and silent recall. The authors were not sure if the babble noise interrupted the effect because of its informational nature (i.e., sounding like real words) or because it drew attention away from the task. A second experiment was completed using fluctuating energetic sounds to see if it was the linguistic nature of the babble noise or the fact that the noise fluctuated that eliminated the production effect. They found that fluctuating noise also eliminated the production effect although average overall recall was similar to that of the no noise and steady-state noise conditions, which were all significantly higher than the babble noise condition. Furthermore, there was no difference between means across the four noise conditions for silent words. It seemed that it was not the informational quality of the background noise that interrupted the production effect but rather the fluctuation in the noise presented that reduced it. The authors suggested that noise impeded attentional processes during production and especially affected the aloud words which benefit from additional attention during production, drawing from Bodner et al.'s (2014) idea of "shallow processing" that might occur to silent words.

A later study by Mama and Icht (2018a), further examined the role of attention in the production effect by studying a sample of adults with attention-deficit/hyperactivity disorder (ADHD). This disorder is characterized by an impairment in cognitive functions, especially those associated with attention (Faraone et al., 2000; Fuermaier, et al., 2015). The authors tested a clinical sample of adult participants with diagnosed ADHD who were already prescribed and taking methylphenidate (MPH) - a medication used to treat the attentional deficits associated with the disorder. The researchers had them complete a typical free-recall production task with and without MPH and compared their results to a healthy control group. The authors predicted that if there was an attentional component to the production effect, and that vocalizing improves attention, then they should see a production effect across both groups (control, ADHD) but that the production effect would be larger in the ADHD group after taking MPH. Their results confirmed their hypothesis: all three groups showed a significant production effect, and the production effect was larger for participants after they had taken MPH. Interestingly, the difference in the production effect between the ADHD groups was driven by improved performance for aloud words in the MPH group and there was no difference in the recall of silent words. Furthermore, testing after administration of MPH revealed better memory for aloud words boosting recall scores to that comparable to the control group but the amount of silent words recalled was lower than the control group. These results were consistent with an attentional account of the production effect previously put forth by Mama et al. (2018), described above, whereby completing the experiment in a noisy environment that interfered with attention eliminated the memory benefit from production, but the attentional manipulation did not have an effect on their silent counterparts. This finding suggests that attention could be an integral process in the production effect. The fact that manipulating attention in both studies had an effect on aloud words but not silent words suggests that attention is not just a supporting background process, but rather that it plays a specific role in the improved memory outcome for aloud words.

In a similar vein, Forrin et al. (2019) theorized that silent items in the mixed-list design may be weakened by aloud words due to performance anticipation: Participants know that they will be saying an upcoming word aloud and their anticipation causes their attention to be diverted from the task at hand. In a series of four experiments they modified the typical production task by using different designs that would alert participants as to whether the upcoming words would be said aloud or silently. The experimenters theorized that if performance anticipation was interfering with encoding for silent items then the closer the silent word was to the next performance block the more it would be affected by performance anxiety. This notion is similar to the “next-in-line” effect whereby a participant has worse memory for a word they hear right before they know they will be speaking aloud compared to prior words they had heard earlier (Brenner, 1973). The hypothesis was confirmed across three experiments; although observing the expected production effect, the researchers noted a linear decline in the performance of silent words as their proximity to words that were to be said aloud increased. Providing further evidence that performance is affected by the social factor of performance anticipation, when the participant completed the same task without an experimenter in the room the cost to silent items was lessened. Based on their findings the authors proposed that performance anxiety affected silent items in two ways: (1) attention was diverted from the silent items due to a preoccupation with upcoming performances; (2) upcoming performance created anxiety which decreased encoding for silent items (Forrin et al., 2019). One issue with this study is that in a standard production paradigm the participant is not privy to what word will appear next. Forrin et al. (2019) suggest that in these designs, a 'blanket cost' is imposed – all silent items are affected by a performance anticipation as participants do not know when the aloud items will appear.

In addition to attention, other research suggests that encoding strength may also contribute to the effect. Mama and Icht (2018b) examined the notion that the production effect could be influenced through effortful processing using delayed production. In the delayed production condition participants would complete a typical production trial but rather then saying the word aloud when it was presented they held their verbal response until they were prompted by a later cue. The word did not remain onscreen during the delayed interval between viewing the stimuli and producing the word. This meant that participants would have to retrieve the word from their working memory - akin to the testing effect, whereby memory for items that are tested is superior to memory for items that are studied a second time (see Eisenkraemer, Jaeger, & Stein, 2013, for review). However, the distinctiveness theory of production would predict that there would be no difference between immediate vocal production and delayed vocal production since the same number of distinctive elements (e.g., auditory, semantic, motor features) would be present.

Mama and Icht (2018b) found that delaying vocalization by either 1 or 3 s improved recall scores compared to reading the word aloud immediately. The advantage of delayed production persisted even when study time for the aloud words was increased to three seconds. However, the authors noted that the increase in memory observed following the delayed trials could be caused by factors related to withholding the production response or extra processing of the word during the delayed time period. To examine if the benefits of delayed production were related to effortful retrieval and not these other factors, they added a delayed reading condition. In both the delayed production and the delayed reading condition participants first viewed the stimuli followed by a delay (no stimulus) of 3 s, then participants retrieved the word from memory (i.e., delayed production) or the word reappeared on the screen and the participant read it aloud (i.e., delayed reading). The memory advantage for delayed production persisted even when compared against delayed reading, evidencing that improved memory performance resulted from the increased effort that was used in retrieving the items and not increased study time (Mama & Icht, 2018b). This outcome supports the notion of a strength account of production, that rendering production more effortful increases the encoding strength of the word and results in a stronger memory trace (Bodner & Taikh, 2012).

In a similar procedure, Thoms, Fawcett, Hourihan and Willoughby (2019) used a delayed production procedure but modified it by including differing stimulus onset asynchronies (SOAs) for both aloud and silent words. The timing between the stimulus onset and the ‘Go’ signal to produce the word varied by 500 or 1500 ms. Additionally, they included a third, less frequent, catch condition where participants thought they would have to produce the word, but the Go signal never appeared. The authors supported the delayed production effect reported by Mama and Icht (2018), observing a trend towards a higher accuracy for both aloud and silent words after the 1000 ms delay compared to the 500 ms delay. Importantly, unlike the prior study there was no immediate production condition; therefore, no comparison could be made between memory performance for delayed items and immediately produced ones. Nevertheless, the observed trend of greater recognition scores over the increasing intervals of the SOAs suggests that the comparison would be similar to that recorded by Mama and Icht (2018b) where the delayed production effect improved memory relative to immediate production.

Importantly, the key finding by Thoms et al. (2019) was that a significant production effect was still observed during the catch trials even though the word was never produced. This was akin to the offline production effect previously observed by Jamieson and Spear (2014) where a production effect was obtained for words that were imagined to be typed. However, an important difference between the two studies is that Thoms et al. (2019) did not instruct participants to imagine producing the word. Both the finding that accuracy for aloud words increased over a time delay and the observation of a production effect in the absence of a productive act, challenge the distinctiveness account. First, as stated previously by Mama and Icht (2018b), delaying the onset of production does not manipulate the amount or type of distinctive information that is encoded, therefore, it is likely that either processing strength or attention is increased by delaying production. Second, the authors found evidence for the production effect in the absence of distinctive processing elicited by the act of production, finding a memory advantage for aloud words even on stop trials. Although the authors could not rule out the possibility that participants imagined producing the word, there is no reason to expect they would be motivated to do so in the absence of an instruction. Alternatively, these trials might have been distinctive to participants themselves as they may have caught participants off guard. A second explanation is that the participants may have engaged in more attentional or effortful processing of the word when they thought that they would have to produce it, whereas in a silent trial they may have not stayed engaged with the stimulus after it had left the screen. Nonetheless, this study, along with the others mentioned above, suggest that either attention or processing strength play a role in the production effect (Forrin et al., 2019; Mama et al., 2018; Mama & Icht, 2018a, 2018b). While the above findings do not refute the distinctiveness account, they do highlight outcomes of the production effect that do not fit into a traditional distinctiveness account framework, perhaps suggesting an additional component.

1.3.2 Neurological evidence for the production effect. Given the competing theoretical frameworks within this literature, there has been a growing interest in using psychophysiological and neuroimaging approaches to better characterize the mechanisms involved in the production effect. In the first such study, Hassall et al. (2016) measured brain activity using electroencephalography (EEG) during a standard production paradigm. Specifically, the researchers were interested in the P300 – a positive deflection over the frontal and parietal lobes comprised of two subcomponents: (1) the P300a which is localized to the frontal cortex and is associated with stimulus driven processing; and (2) the P300b which is localized to the parietal lobes and reflects distinctive processing (although see Polich, 2007). Of particular note, the P300 is largely characterized by its nature to be activated by unique stimuli and has previously been used to identify the impact of distinctive items on memory in paradigms such as the primacy and von Restorff effects (Fabiani & Donchin, 1995; Kamp, Forester, Murphy, Brumback, & Donchin, 2012).

Since the P300 had previously been associated with distinctive encoding, Hassall et al. (2016) inferred that the signal should be observed during a production task and that the amplitude of the signal should be larger for produced words compared to silent ones. Furthermore, a mode of production that incorporates more distinct elements into the encoding trace (e.g., singing) should result in a larger P300 signal than words read aloud. Consequently, in addition to aloud and silent conditions the researchers added a third “sing” condition. Participants completed the task while their neural activity was measured with EEG; the word and instruction were presented before enacting production because speaking can interfere with neural recording. The event-related brain potentials (ERPs) included in the analysis were obtained from 300-500 ms post instruction – well before production was enacted. The researchers predicted that amplitude would increase depending on the amount of distinctive processing which occurred and therefore posited that the mode of production which would elicit the highest amplitude would be singing aloud, followed by speaking aloud and lowest for reading silently. Furthermore, they predicted that memory performance would follow a similar trend and that a larger peak deflection would be associated with better memory for more distinctive words.

The authors did not find an increased behavioural production effect for singing (i.e., larger difference in memory performance between words sung aloud versus read silently than for words read aloud versus words read silently). However, recognition was higher both for items that were sung or spoken compared those read silently, demonstrating the production effect for both conditions. These behavioural findings were accompanied by the observation of P300 waveforms that peaked at 400 ms and were localized to the parietal lobe consistent with a P300b signal (Polich, 2007). Analysis of the P300 revealed that the peak amplitude increased for both spoken and sung words relative to silent words. Contrary to the predicted outcome, the peak amplitude for the sung words was not greater than the words said aloud. Although, the results of the study did not show strong evidence for the distinctiveness account (there was no evidence for more distinctive processing for sung items over spoken items), it also did not refute distinctiveness since produced items still benefited and showed a stronger P300 signal compared to those in the silent condition (Hassall et al., 2016). However, given that the change in signal was observed prior to the word being produced it is possible that this was reflecting a separate process, such as increased encoding strength or attention. The P300 is also sensitive to attentional and memory load processes and the amplitude changes may reflect a response to both attentional load and mental capacity (Kok, 2001; Yu, Prasad, Thakor, & Al-Nashash, 2015). Therefore, the difference in amplitude observed between the production conditions and the silent conditions may have been influenced by either one of these processes. As Hassall et al. (2016) did not demonstrate a clear relationship between distinctive processing and the P300 it cannot be directly inferred that purely distinctive processing is occurring, and a strength or attentional account cannot be ruled out.

A second endeavour to investigate the underpinnings of the production effect by observing the neural mechanisms behind the effect was completed by Fawcett et al. (2020). Their investigation was exploratory and aimed to determine if neural activity during the production effect was consistent with a distinctiveness account. However, instead of EEG, functional magnetic resonance imaging (fMRI) was used to measure blood flow in the brain. By detecting changes in blood flow in the brain fMRI can be used to localize cognitive activity to specific areas of the brain – a key difference from EEG. The authors reasoned that if distinctive processing during the production effect was due to the encoding of more distinct elements (i.e., motoric, semantic, etc.) then areas responsible for this output (i.e., motor and auditory cortices) should be more active during encoding for words that were produced. In addition to reading the words aloud and silently the authors added a third “check condition” where participants were instructed to read the word silently to themselves but say “check” out loud. The purpose of this third condition was to act as a motor control to enable the comparison of any neural responses that would occur due to the effort of speaking alone.

Their behavioural results revealed the expected production effect with more aloud words being labelled ‘old’ than check or silent items; the latter two conditions did not differ. Finally, more responses were made to all study conditions compared to foil words. Examination of neural activation during the study phase revealed more substantial activation in the motoric, auditory and semantic brain areas during the aloud and control conditions compared to the silent condition. Although similar areas of the brain were activated in both the aloud and control conditions, difference comparisons revealed that activation in the motor and auditory cortices were more substantial in the aloud condition. Additionally, there was greater activation of the brain areas involved in semantic processing (i.e., the inferior frontal gyrus and the premotor cortex), suggesting more semantic processing was occurring during aloud trials. Considering the similarities in activation between the control and aloud conditions did not result in behavioural differences, it was suggested that these similarities were the result of cognitive processes unrelated to the production effect. Lastly, a correlational analysis was completed comparing the magnitude of activation during the aloud condition from the baseline measurement, to the magnitude of the production effect (i.e., the difference in memory performance between words in the vocal conditions and the silent words). The relationship between activation in the aloud condition was related to memory performance, meaning that participants who showed more activation were also more likely to remember the words they said aloud. The authors concluded that the activation of the auditory and motor cortices during encoding was consistent with the prediction of a distinctiveness account: greater activation during the aloud condition was related to better memory performance. Furthermore, the upregulation in the motor, auditory and semantic brain areas suggests that participants were encoding more distinctive information during aloud trials compared to control or silent ones. Observations by both Hassall et al. (2016) and Fawcett et al. (2020) provided a different perspective of the cognitive mechanisms behind the production effect using neuroimaging techniques. Next, I will examine another technique that can be used to measure cognitive processes during a production task.

1.4 Using Pupillometry to Measure Cognitive Effort

In addition to neuroimaging techniques, there has also been growing interest in the use of unobtrusive psychophysiological measures, such as pupillometry. This approach involves measuring pupil size during cognitive tasks and allows researchers to quantify mental effort. Pupil dilation is measured while participants are engaged in cognitively demanding tasks. In particular, past research has found that while engaging in a cognitive task, pupil diameter will vary as a function of the task demands: As the task becomes more difficult (e.g., completing a difficult math problem) the pupils dilate, and when the difficulty of the task lessens pupils retract (e.g., completing a simple math problem; Hess & Polt, 1964). Similar relations between pupil size and cognitive effort have been demonstrated in a number of cognitive domains, including recognition memory (Geller, Still & Morris, 2016), working memory (Kahneman, & Beatty, 1966), and attention (Unsworth & Robinson, 2018). Given its functionality in quantifying cognitive effort, this method could perhaps shed light on the cognitive processes underlying the production effect. Throughout the remainder of the document change in pupil size will be discussed as a measure of cognitive effort only, and not as a response to another factor, for example, the response to a change in luminance.

Applying pupillometry to the production effect may shed light on the processes involved, providing a quantitative measure of attention and effortful cognitive processing. For example, it would be expected that an increase in pupil dilation would occur to a greater degree when participants say the word aloud but not when they read it silently since it is predicted that reading aloud should be more cognitively demanding than reading silently. As an alternative, if participants were relatively less engaged during silent trials, a rapid drop-off in the pupillary signal may be expected to occur during those trials, perhaps reflecting mind wandering which is characterized by a decrease in pupil diameter (Grandchamp, Braboszcz, & Delorme, 2014; Smallwood, et al., 2011; Smilek, Carriere, & Cheyne, 2010; although see Franklin, Broadway, Mrazek, Smallwood, & Schooler, 2013).

Accordingly, in a recent study, Tiller, Hourihan, and Fawcett (2018) measured variation in pupil diameter during a production task to determine if aloud trials were more cognitively demanding than silent trials. They reasoned that if there was a strength component to the production effect, then this should be reflected in differences in cognitive effort at encoding. Using a similar procedure to that of Fawcett et al. (2020) they further controlled for any cognitive effects that might have been associated with the effort that it took to speak aloud during production by including a third ‘check’ condition, where participants always gave the same verbal response (“check”) to control for cognitive effort that might occur as a result of speaking alone. They predicted that pupil diameter would increase to a larger degree on the aloud trials than silent trials with check trials falling in the middle. Overall, their predictions were supported, with data demonstrating a pupillometric production effect defined by larger pupils on the aloud trials compared to the silent trials. A particularly compelling finding was that pupil dilation on aloud trials was sustained over a longer period of time than during the check trials, while dropping off quickly in the silent trials. Finally, the magnitude of the aforementioned pupillometric production effect, was itself predictive of the magnitude of the subsequent behavioural production effect (i.e., more aloud words remembered than silent at test). That is, participants who showed more effortful or attentive encoding during the study phase – as indicated by pupil dilation – were more likely to show a larger production effect.

Considering these findings, it was suggested that encoding was more effortful for aloud items as evidenced by an increase in pupil diameter during study. Additionally, it was suspected that pupil size was influenced by an attentional component since participants' pupil size retracted during the silent conditions over the course of the trial, whilst dilation was sustained on aloud and check trials. This could be evidence that participants attentionally disengaged from silent items at a faster rate but sustained attentional efforts on their primary task – either processing the aloud item or saying "check", respectively – for a longer duration during aloud and check conditions. This notion is consistent with studies on mind-wandering which show that pupils retract when participants are not attentionally engaged during task performance (Unsworth & Robinson, 2018).

However, in considering this early study there is limited ability to draw firm conclusions pertaining to the observed pupillometric effects because any putative differences in attentional allocation necessarily overlap temporally with effects driven by production itself; this is indicated by the finding that comparative differences in pupil dilation on the check and aloud trials was only marginally significant (Tiller et al., 2018). A stronger test would somehow separate the production and other cognitive processes in a manner such that they might be investigated individually during the production task. In the following experiments I aim to dissect the cognitive processes that are specific to interpreting the production cue and the stimulus separate from the act of producing.

1.5 Current Experiments

The current studies addressed the problem of not being able to separate the motoric act of production from the other cognitive processes. Both studies were designed to replicate Tiller et al. (2018) with some procedural changes used to tease apart the different processes that might contribute to the production effect. For example, by separating motoric production from the initial encoding of the stimulus I can address whether processing effort differs across conditions in response to distinctive stimuli and if there are preparatory differences in initial encoding if a participant knows that they will be later producing the word. Additionally, by separating vocal responses from the stimulus presentation I can elucidate attentional differences that might occur in response to words that are produced versus those that are not. For example, if the word is encoded prior to production, will participants still engage with the word they have read silently during the time given to produce the word or will they disengage with the task (e.g., mind-wander or prepare for the upcoming trial)? Following procedures from past studies applying psychophysiological measures to the production effect, I added a third control condition where participants read the word silently but said “check” aloud, to control for any effortful processes that could occur in response to speaking alone (Fawcett et al., 2020; Tiller et al., 2018).

To explore these ideas, first, in Experiment 1 a pre-cuing procedure was implemented to highlight preparatory processes that might occur when participants know they will be producing something. Subsequently, in Experiment 2 I then implemented a delayed production procedure in which participants will be presented with an instructional cue simultaneously with the word but then asked to delay production until a later ‘Go’ cue denoting that they should carry out the preceding instruction (Hassall et al., 2016; Mama et al., 2018). Using delayed production I can examine the hypothesis that participants will disengage with silent items during the time allotted for production after the stimulus is removed from the screen.

Together, I predict that that in both experiments a behavioural and pupillometric production effect will be obtained – that is, better recognition for aloud than silent items at test and greater pupil dilation for aloud items than silent items during encoding. I expect that recognition performance for check items will be equivalent to silent items since past research has indicated no resulting production effect when a non-unique word is said aloud (MacLeod et al., 2010). Furthermore, I anticipate that pupil size in the check condition will be greater than that of the silent condition due to the effort of speaking, but smaller than dilation in the aloud condition due to less effortful cognitive processing being used when saying the word. In Experiment 1, I expect that the pre-cueing procedure will result in differences in pupil size concurrent or slightly preceding stimulus onset, indicative of preparatory cognitive processes. In Experiment 2, I predict that the post-cueing procedure will reveal differences in attentional disengagement with pupil size rapidly decreasing for the silent items compared to the aloud and check items, especially during the cue to produce. Finally, I expect to replicate Tiller et al.’s (2018) observed relationship between the magnitude of the pupillometric production effect and the behavioural production effect across both studies. Considering the two experiments in totality should give further insight into the roles of cognitive effort and the possibility of an attentional role in the production effect.

Chapter 2: Experiment 1

1.1 Overview

The primary objective of Experiment 1 was to elaborate on the observations made by Tiller et al. (2018) by using a pre-cueing procedure in which the participants would view the instruction before the stimulus appeared on the screen. In Experiment 1 participants viewed the instructional cue (e.g., if the word was to be read aloud) before the word presentation and verbal production. Therefore, participants would have time to prepare for their upcoming response. Importantly, this procedure will allow me to infer if differences in cognitive effort occur in the period surrounding the cue onset when participants are not giving a motoric response. It is possible that participants will engage with the trials differently depending on the action they will be taking; for example, participants might show more attentional engagement on trials where they know they will be speaking, even before they know what they are meant to say (and therefore, prior to the possibility of any form of distinctive processing).

With the intention of investigating the preparatory (i.e., pre-cue) and motoric responses separately, I grouped my pupillary analysis into two separate time windows: (1) an early window spanning the pre-cue period to assess preparatory processes; (2) a late window including the time the stimulus was onscreen during which the participants would be speaking, to investigate processes related to the stimulus and motoric production. I anticipate that pupil size between the vocal conditions (i.e., aloud, check) and silent condition will diverge during this pre-cue period, which could be indicative of preparatory cognitive processes. For example, participants might exert greater task specific attentional efforts if they know that they will be speaking in the second half of the trial. If differences in pupil size prior to word presentation were observed this outcome would be difficult to reconcile within a distinctiveness framework as these differences could not be a response to distinctive elements of the word before the word is on screen. Furthermore, I expect to replicate the early behavioural and pupillometric findings by Tiller et al. (2018). Specifically, I expect to observe a greater increase in pupil size during aloud trials than both silent and check trials, with dilation during check trials being intermediate. Finally, I expect that I will find the behavioural production effect, but that memory for words presented during the check condition will not be remembered better than silent items, despite a difference in pupil size during study.

2.2 Methods

2.2.1 Participants. Sixty-six students at Memorial University participated in exchange for 1 credit point towards a psychology class or were paid $10.00 for an hour of their time. Those who received points were recruited through the university psychology pool, paid participants were recruited through poster advertisements. Participants were asked to refrain from wearing makeup as it might interfere with the accuracy of the pupillometric equipment. Data from 10 participants were excluded from the study as over 50% of their data was removed during preprocessing (see Section 2.2 for an outline of the preprocessing procedure and exclusion criteria). Of these, data from (a) two participants were excluded due to interference with the eye tracker (i.e., persistent glare from eyeglasses or a black rim interfering with the eye tracker); (b) two participants were removed because they blinked frequently during the baseline collection period; and, (c) one participant was removed because their head moved out of the chin rest during the study, invalidating their calibration. The remaining exclusions were caused by lost data from frequent blinking that occurred throughout the trial.[1] The final sample consisted of 56 participants.

2.1.2 Materials. *Stimuli and Apparatus.* The current study was created using the Experiment Builder (SR Research, 2020) software package developed for the Eyelink 1000 Plus (SR Research, 2010). The experiment was loaded on the display computer, a MacMini Computer running OSX 10.12 and was displayed on a 22" 1020x768-resolution Benq. All stimuli were presented in black (RGB: 0, 0,0) size 24 Courier font on a grey (RGB: 128, 128, 128) background. Pupil and gaze data were monitored by a separate, networked computer running specialized recording software. During the experiment, output from the eye tracker was displayed to the experimenter on a V176L Acer screen, where they monitored data quality throughout the task.

Pupillary data from each participant's right eye was recorded at a rate of 500 Hz using an Eyelink 1000 Plus eye-tracker (SR Research, 2010) deployed across from the desk-mounted configuration. The eye tracking hardware was placed below the monitor and a chin rest was installed on the desk to reduce unnecessary head movement during the study. The chin rest was positioned so the participant's eyes were 105 cm away from the display screen and their forehead was positioned against the forehead rest 50 cm above the desk. The forehead rest was maintained at a constant height meaning every participant's gaze was similarly positioned in relation to the camera. The distance from the camera attached to the eye tracker to the chin rest was also 50 cm. The participant sat in a standard computer chair onto which wheel covers were placed to further

[1] Exclusion criteria were determined apriori with the intent of maximizing data quality; however, exploratory analyses revealed that inclusion of all participants did not impact my primary conclusions.

prevent unwanted movement during the study. The participant responded using a keyboard placed on the desk directly in front of the head rest.

Stimuli consisted of 200 words randomly selected from the MRC Psycholinguistic Database (Coltheart, 1981). These words were subdivided into five randomized lists of 40 words. Lists were matched for word length (M = 5.42, SD = 1.28) and the Kucera-Francis written word frequency (M = 63.76, SD = 81.02). To control for item-level variability, I counterbalanced the five lists across all conditions (aloud, silent, check and two foil lists). Counterbalancing and randomization were implemented using a custom *Python script* and run using *PsychoPy3* (Peirce et al., 2019).

Production instructions (aloud, silent, check) were represented by a box (720 pixels x 540 pixels) in centre screen. The perimeter of the box was made up of either ‘x’ for silent trials, ‘+’ for aloud trials, or ‘✓’for “check” trials. Each symbol was comprised of four identical straight lines, reconfigured to create the desired shape, ensuring that they were matched for luminance. A fourth box was made out of squares, also matched for luminance, and acted as a neutral interstimulus placeholder between trials to maintain equal luminance between screen changes.

Pretest Questionnaire. Prior to completing the experiment participants were asked to complete a brief demographic questionnaire. Data collected with the questionnaire included age, gender, handedness, language, hours of sleep the previous night and caffeine consumption prior to study. These questions were in multiple choice or Likert scale format.[2] These data were collected as part of a standard questionnaire used in this laboratory and were not analyzed.

[2] Due to university closure because of a global pandemic, I was unable to access the previously collected demographic information.

2.2.3 Procedure. The experiment consisted of four phases - a familiarization phase, a practice phase, a study phase and a test phase. Calibration of the eye tracking device was completed prior to each phase, excluding the familiarization phase, to ensure a precise measurement of the pupil. During calibration the participant was asked to fixate their gaze on a dot on the screen, which moved around to nine points approximating a square. The fixation points appeared one at a time, disappearing once the eye tracker recorded fixation of the pupil. After calibration, this procedure was repeated a second time during a validation procedure, and an accuracy measurement of the difference between the calibration fixation and the validation fixation was recorded. This process was repeated on a participant-by participant basis until the difference measurements were minimized as much as possible.

Initial Setup. Before initiating the study the experimenter assigned the participant to one of five counterbalance conditions to determine which of the three lists would comprise the study list presented during the study phase and which two would be later integrated as foil items in the list presented during the test phase. Following consent and the demographic questionnaire, the researcher introduced the participant to the eye tracking device, ensuring the participant was seated at a comfortable height for the head rest. Adjustments were first attempted by moving the height of the chin rest if there was a large height difference the height of the chair was also manipulated. Modifications were made so that the eye tracker was focused on the right eye. The researcher explained that the study would consist of four phases: (1) the familiarization phase where the participant would be told the instructions for the study; (2) the practice phase where they would be given a short list and practice reading the words while following the instructional cues; (3) the study phase where they would memorize a list of words; and (4) the test phase where they would be shown a second list of words and indicate if they had previously viewed them.

Familiarization Phase. In this phase, the participant was familiarized with the study procedure. Specifically, each instruction cue (box) was presented with explicit instructions written in the centre noting their meaning. For example, the ‘x’ instruction was presented with the instructions “Read the word silently”. The ‘+’ instruction was presented with the instructions “Read the word out loud”. Lastly, the ‘✓’ instruction was presented with the instructions “Say ‘Check’”. Each of the instruction cues was presented three times for 6000 ms, for a total of nine trials. Note that participant was verbally instructed to also read the word on the screen silently to themselves in the “check” condition in addition to making the verbal response; they were also reminded not to say ‘check’ aloud until the word appeared on the screen. The participant was told that the words would appear on a later test so they should try their best to memorize them but were not given any instructions on how.

Practice Phase. After the familiarization phase, the participant was asked to repeat the instructions they had read to ensure understanding. If they could not remember the instructions the researcher verbally repeated them a second time. The practice phase followed an identical format to the study phase listed below; calibration and validation were completed prior to the presentation of the practice list. The practice phase was shorter than the study phase and only contained 9 words each three-letters in length (e.g., ‘CAT, ‘CAR’, and ‘BAR’). This shorter list was presented to ensure the participant understood the instructions and was capable of performing the tasks. The participants were told that they did not have to remember these words for the later test.

Study Phase. Immediately preceding the study phase, the researcher calibrated the eye tracker, as described earlier. As depicted in Figure 2.1, each study phase trial began with two blank screens for 200 ms interposed by a 1000 ms “blink” screen – made up of three exclamation marks presented at center. During this time the participant was instructed to blink if needed so that they could keep their eye open for the remainder of the trial; otherwise they were instructed to avoid blinking outside this period, but that they could blink during the trial if needed. Next, the instruction cue for that trial (e.g., ‘+’, ‘x’, ‘✓’) was presented with a fixation dot centered in the middle for 1500 ms. As described above, a pre-cuing procedure (e.g., Fawcett, et al., 2020) was used in this study meaning that the cue was displayed preceding the target stimulus for 1500 ms after which the fixation dot was replaced by the target word for 3500 ms with the cue remaining onscreen. Once the word appeared on the screen the participant was asked to implement the appropriate instruction (e.g., reading the word aloud if it was an aloud trial). The trials automatically progressed without input from the participant. To control for changes in luminance that would affect pupil dilation, a placeholder that matched the pixilation of the cue screen (See Figure 2.1) was displayed for 500 ms before the pre-cue and 1000 ms after stimulus presentation. As described below in Section 2.2, the period preceding cue onset served as the baseline period for analysis purposes.

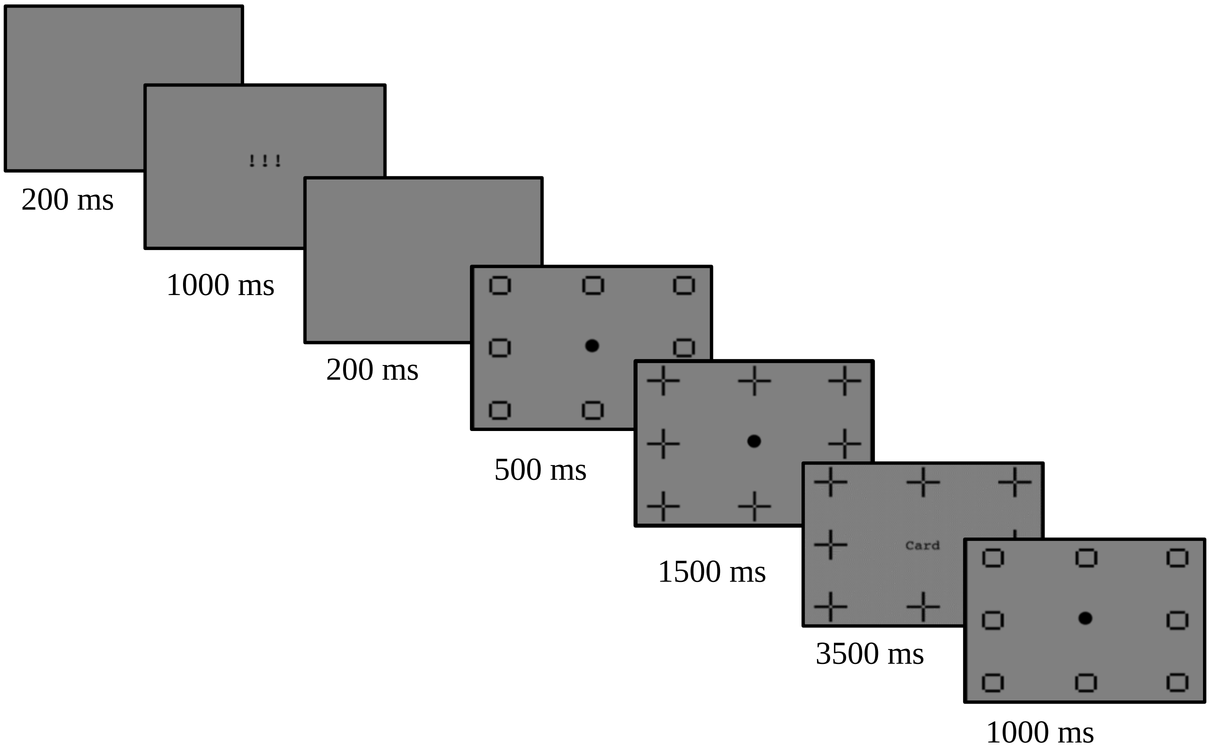

Figure 2.1: Outline of the screens presented to participants during a typical trial in Experiment 1. This experiment used a pre-cue followed by joint cue and item presentation. The “+” symbols displayed in the present figure indicate that the participant should read the word aloud. The three exclamation marks denote that the participant should blink at that point. Duration is indicated below each event.

Test Phase. The study phase was followed by a brief break in which the experimenter gave the participant instructions for the test phase and calibration and validation was again completed with the eye tracker. Participants were presented with all 200 words – the previous 120 words presented at study intermixed with the two 40 word foil lists. Mirroring the study phase each trial began with two blank 200 ms screens surrounding a 1000 ms blink screen during which participants were again instructed to blink if needed. After this a fixation dot was displayed in the center of the screen for 2000 ms and then replaced by the test-word which appeared for a period of 3000 ms. After, the target word disappeared, and the participant was given time to make their response using the numerical keys (1-6) at the top of the keyboard. The participant was asked indicate whether they had seen the word before (i.e., an old word) or if it was a word they had not seen before (i.e., a new word) by rating the word on the following scale: 1 – very sure new, 2 – mostly sure new, 3 – unsure new, 4 – unsure old, 5 – mostly sure old, and 6 – very sure old. The scale was displayed each trial following the removal of the target word and remained until a self-paced response was made, which initiated the beginning of the next trial.

2.3 Pupil Processing

Prior to analysis, the pupillary signal was pre-processed to maximize the signal-to-noise ratio. This section outlines each pre-processing step taken in its approximate order.

2.3.1 De-blinking and artifact removal. First, blinks in the data were identified and removed in a process known as de-blinking. Blinks are characterized by an occlusion of the eye which occurs when the eyelid covers the pupil. An algorithm developed by Hershman, Henik, and Cohen (2018) was used to identify blinks by measuring the pattern of rapid fluctuations represented by measurement error recorded by the eye tracker that is distinct from the physiological pupil fluctuations, which occur at a much lower frequency. During a blink there are no rapid fluctuations in the data, due to the absence of a pupillary signal. Thus, the algorithm picked up on the absence of the signal and identified these instances as blinks. Next, the onset and offset of the blink are identified by pinpointing the last signal that occurred prior to the loss of signal caused by the blink. The pupil signal rapidly changes during the time immediately before and after the blink due to the partial covering of the pupil by the eyelid. The algorithm next identified these periods of change. The data were first smoothed to increase the signal to noise ratio between rapid fluctuations of measurement error (i.e., noise) and the steady physiological pupillometric signal of interest. Then the differences in pupil size were compared between each consecutive sample using a moving average until a monotonic pattern emerged (i.e., the stable pupil signal reflecting cognitive activity). Once the start and endpoints of the blink were pinpointed, the data points between were removed only where these fluctuations began and ended. Hence, only those data points that were affected by the eyelid signal were removed.

Afterwards, I located and removed artifacts (missing trials or changes in the pupil signal that were unrelated to cognitive effort) that were not picked up by the blinking algorithm. A small number of trials contained no pupillary signal; there were a few circumstances where a pupil signal may be absent over the course of a trial. First, the participant may have engaged in a prolonged blink and kept their eyes closed for the duration of a trial. Although infrequent and discouraged, this may have occurred if the participant was fatigued by prolonged fixation on the centre of the screen. Second, in a few cases the eye tracker may have temporarily lost the pupil. This can be caused by interference from eyeglasses. For example, the eye tracker differentiates the pupil from the iris by picking up on an infrared reflection on the cornea that is reflected back from an infrared illuminator that is attached to the eye tracker, eyeglasses that have a black rim can reflect the illumination and therefore be picked up by the eye tracker (Gagl, Hawelka, &

Hutzler, 2011; SR Research, 2010). Additionally, in some cases light from the infrared illuminator may reflect off the lens of the eyeglasses; this reflection is also sometimes picked up by the eye tracker and can occlude the pupillary signal. In exceptional cases the pupil signal may have been lost if the participant moved their head. Movement was discouraged during the study and the participant was asked minimize movement after calibration; as an additional preventative measure they were also requested to keep their head in a chinrest during the study. Despite these protocols, occasionally a participant might have moved involuntarily, such as by coughing or yawning, which may have disrupted the pupil signal. Missing trials were easily identified due to the absence of data and subsequently removed from analysis.

Finally, I addressed artifacts in the data that were caused by the eyes being focused too far away from the center of the screen. This problem occurs when a participant looked away from the centre of the screen the shape of the pupil changed from a circle to an ellipse and additionally may have become partially occluded by the eyelid (Gagl, et al., 2011). In either scenario the measure of pupil diameter was contaminated since the entire diameter of the pupil was not visible and its shape was distorted. The present study was designed so that most task-related stimuli appear centre screen (or near centre screen), eliminating a need to focus the gaze away from the center. Therefore, I controlled for wandering gaze by restricting the area of screen that is of interest and removing pupil data collected when the gaze was focused outside this area. I defined the cut-off regions as being a focused gaze more than a quarter of the width or height of the screen away from the outside of the instruction box. Gaze which drifted outside of the cut-off region was identified and removed (i.e., treated as missing) using a custom function in *R 3.5.1* (R Core Team, 2018).

2.3.2 Downsampling. After data were processed and noise caused by blinks and artifacts was removed, it was then downsampled from the original sampling rate of 500Hz to a rate of 50Hz. This process was completed to reduce autocorrelation of the residuals and minimize the load of data on the processing software. A cut-off of 50Hz was selected based on van Rij, Hendriks, van Rijn, Baayen and Wood (2019) who determined it as an ideal minimal cut-off. Pupil change in response to cognitive activity is quite slow and sampling rates as low as 30Hz are sufficient to capture it (Winn, Indt, Koelewijn, & Kuchinsky, 2018). After downsampling the sampling rate was reduced to samples of 20 ms meaning that the trial was split into even 20 ms time periods; all subsequent processing and analysis occurred within these 20 ms time samples.

2.3.3 Baseline calculations. To measure the change in pupil size which occurred in response to cognitive effort, a baseline measurement – when the participant was alert but not exerting cognitive effort – was first obtained. The baseline measurement was collected during the interstimulus placeholder screen presented between the blink and pre-cue screen as seen in Figure 2.1. The screen appeared for a duration of 500 ms; I collected the baseline measurement for the last 200 ms of the presentation to avoid overlap with any blinking that may have been residually occurring from the previous screen. The baseline for each trial was calculated by averaging the recorded pupil size during that 200 ms time window.

Next, I calculated the change in pupil size for each trial from its respective baseline estimate. This was done by subtracting the average of the baseline period for a given trial from each sample within the remainder of that trial. I chose to use a subtractive baseline over a standardized baseline as there can be issues that arise when using standardization. For example, the same level of pupillary response can seem greater in magnitude due to an unrealistically small baseline (for example, if measurement error is present in the baseline sample), when standardizing the baseline, pupil size is divided by the baseline measurement, which inflates incorrect measurements to a greater degree compared to when subtraction is used (Mathôt, Fabius, Van Heusden & Van der Stigchel, 2018; van Rij et al., 2019).

2.3.4 Data cleaning. Finally, trials and/or participants were excluded if too much data were lost during the preceding steps. Individual trials were excluded if more than 50% of the baseline period or 75% of the remaining trial was missing respectively. Participants were excluded if more than 50% of their trials were missing. As noted earlier, 10 participants were excluded in this manner, resulting in 56 participants included in the final analysis.

2.4 Results

All statistical analyses were completed using *R 3.5.1* (R Core Team, 2018). Results calculated using analysis of variance (ANOVA) were completed using the *ez* package (*v*4.4-0; Lawrence, 2016). *Post-hoc* comparisons were carried out using Fisher's least significant difference (FLSD); these values are presented visually in the relevant figures, such that any two conditions for which the error bars do not overlap are considered statistically significant, so long as the relevant main effect or interaction was also found to be significant (Carmer & Swanson, 1973; Ramsey, 2002). Although my primary interest was the impact of production instruction on pupil dilation, I first analyzed the behavioural data.

2.4.1 Behavioural analysis. Confidence ratings (1-6) from the recognition test were converted into binary (yes/no) responses such that, 1, 2 or 3 were coded as 0 ("no") and 4, 5, or 6 were coded as 1 ("yes")[3]. The proportion of "yes" responses were then averaged for each test-phase condition (aloud, check, silent, foil) and compared using a repeated measures ANOVA. As

[3] Confidence ratings were collected for later analysis outside the scope of the thesis and were therefore collapsed into a binary (yes/no) responses for the present analysis.

depicted in Figure 2.2, my analysis revealed a significant main effect of condition, $F(3, 165) = 302.87$, $MSE < .01$, $p < .001$, $\eta_{G}^{2} = 0.74$. Accordingly, participants were more likely to respond 'yes' to items that had been read aloud followed by silently read items and check items, which themselves differed. Participants were least likely to respond "yes" to foil items (reflecting a false alarm), which were (falsely) recognized at a rate significantly lower than any of the remaining conditions.

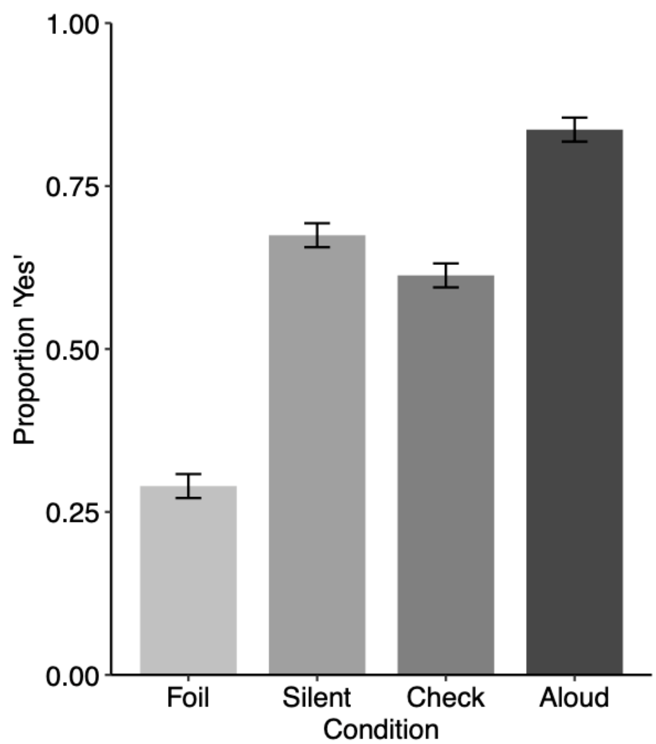

Figure 2.2 Proportion of 'yes' responses as a function of condition (foil, silent, check, aloud) in Experiment 1. Error bars represent ½ Fisher's Least Significant Differences (FLSD).

2.4.2 Pupillary analysis. Having established the presence of a behavioural production effect I next assessed the results from the pupil data (see Section 2.3 for pre-processing). Change in pupil diameter was aggregated to create a grand average by condition (aloud, check, silent) and across participants to produce the waveforms displayed in Figure 2.3. Note that a value of zero for time comparisons refers to the onset of the pre-cue.

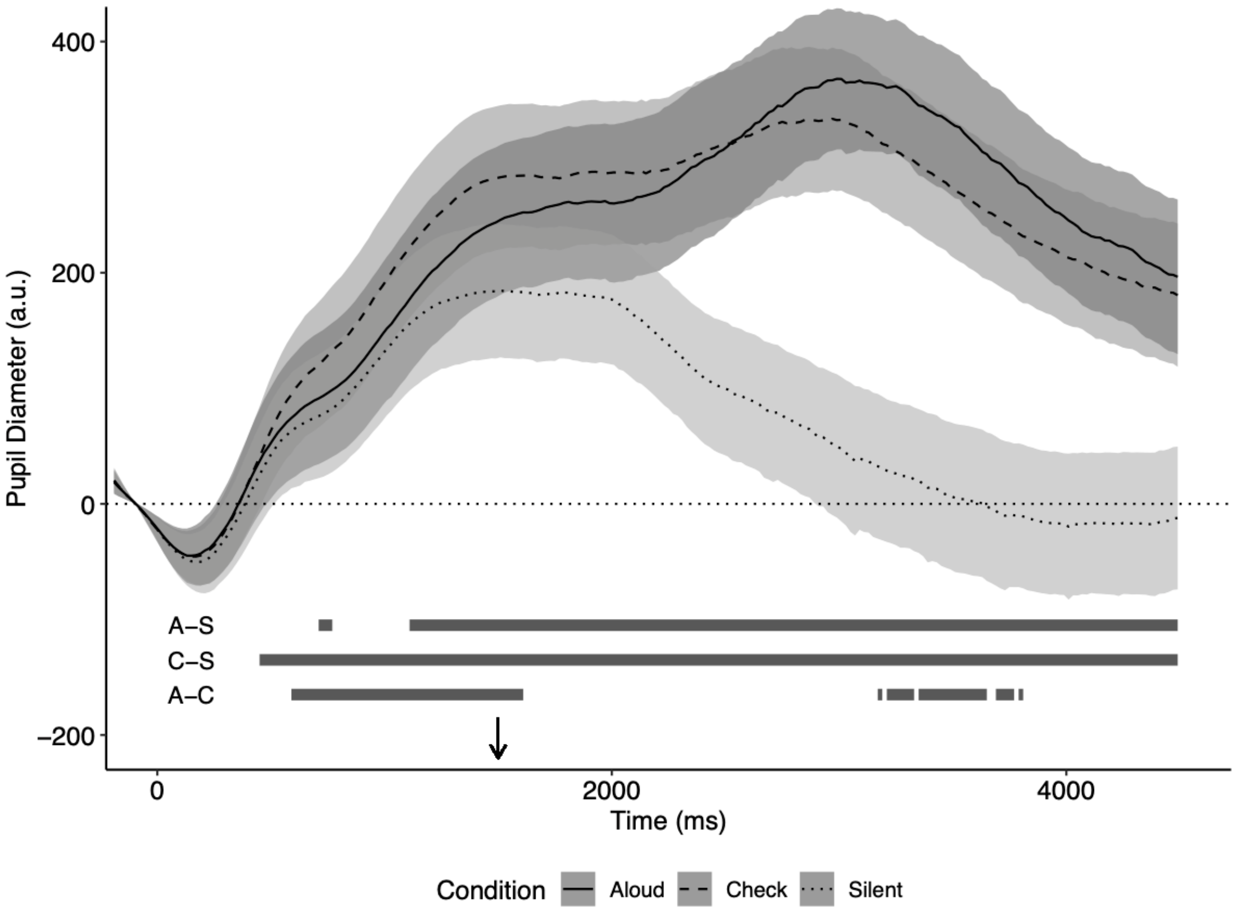

Figure 2.3. Mean change in pupil diameter in arbitrary units, as a function of condition (aloud, check, silent). On the x-axis, zero represents the onset of the pre-cue. The dotted horizontal line represents average pupil size during the baseline period. Error bars show 95% confidence intervals. The bars running along the bottom of the graph reflect periods for which the appropriate conditions (e.g., A-C = Aloud – Check) differ significantly ($p < .05$). Word onset is represented by an arrow.

Prior to conducting my planned, windowed analyses, I first applied an exploratory mass univariate analysis to the time-series (Groppe, Urbach, & Kutas, 2011); to this end, student's t-tests were computed comparing change in pupil diameter between conditions for all possible comparisons (a total of 225 tests per time series). Although the analysis was exploratory, and therefore required no multiplicity correction, the *p*-values were nonetheless corrected for False Discovery Rate (FDR), ensuring that no more than 5% of the differences deemed significant would be spurious; this approach is common in fMRI studies (see Benjamini, & Hochberg, 1995; Genovese, Lazar, & Nichols, 2002). Those regions where comparisons produced significant results are summarized by the lines at the bottom of Figure 2.3.

All three comparisons revealed significant differences prior to – and in the period surrounding – word presentation at 1500 ms. For comparisons against the silent condition, these differences emerged early in the trial (around 500 ms for the check trials and 1500 ms for the aloud trials), with pupil size plateauing and then diminishing steadily in the silent condition soon thereafter, whilst in both the check and aloud conditions, pupil size continued to rise. The check condition demonstrated a particular increase during the pre-cue period. Speculatively, unlike during the aloud trials, during the check trials participants would know what they would be saying and were likely preparing their response, accounting for the additional effort during this period. However, following presentation of the stimulus the increase in pupil size in the aloud condition became significantly greater than that in the check condition. Interestingly, in all three comparisons I observed differences in pupil dilation prior to word onset suggesting that production-related changes in pupil size occurred in response to the instruction alone even before viewing the stimulus.

Following the exploratory analysis of the time series, I then conducted a planned analysis comparing pupil dilation for each condition across two pre-specified time windows. The first “early” time-window aggregated the change in pupil diameter measured between 1000 and 2000 ms. This window was chosen to assess my hypothesis that differences in pupil size would emerge in response to cue onset, even preceding the word itself. Although this time window includes a small period (500 ms) following stimulus presentation, researchers who have systematically measured the pupillary response in similar cognitive tasks suggest that any cognitive response is delayed by 500-1300 ms following stimulus presentation (Hoeks & Levelt, 1993; Verney, Granholm, & Marshall, 2004). Thus, it is expected that any effects emerging in the pupillary signal within this window reflect processing related to the cue, rather than the word (and indeed, processing of the cue is likely to have been delayed so as to be represented within that time-frame).[4] The second, “late” time window included pupil samples collected from 2000-4000 ms, to observe differential processing in response to stimulus information alongside the cue and to capture changes in pupil size related to the productive act itself.

Mean change in pupil diameter was analyzed as a function of condition (aloud, check, silent) and time window (early, late) using a repeated-measures factorial ANOVA. A main effect of condition was observed, $F(2, 110) = 64.81$, $MSE = 14700.93$, $p < .001$, $\eta_G^2 = 0.10$, although there was no main effect of time, $F(1, 55) = 0.02$, $MSE = 25326.81$, $p = .886$, $\eta_G^2 = <.01$. However, there was a significant condition x time interaction, $F(2, 110) = 66.333$, $MSE = 4587.69$, $p < .001$, $\eta_G^2 = 0.04$. As revealed by Figure 2.4, during the early window, increases in pupil size during the check condition was greater than the aloud followed by the silent condition, which also differed from one another. Conversely, during the late time window pupils were largest during the aloud condition, although this did not significantly differ from the check

[4] This analysis had the same outcome with the inclusion of the full time window.

condition. Both aloud and check conditions showed a larger increase in pupil size compared to the silent condition indicating that participants exerted greater mental effort during these trials compared to silent ones. Notably, pupil size increased in both aloud and check trials from the early to the late time window, but markedly decreased within this same period for the silent condition.

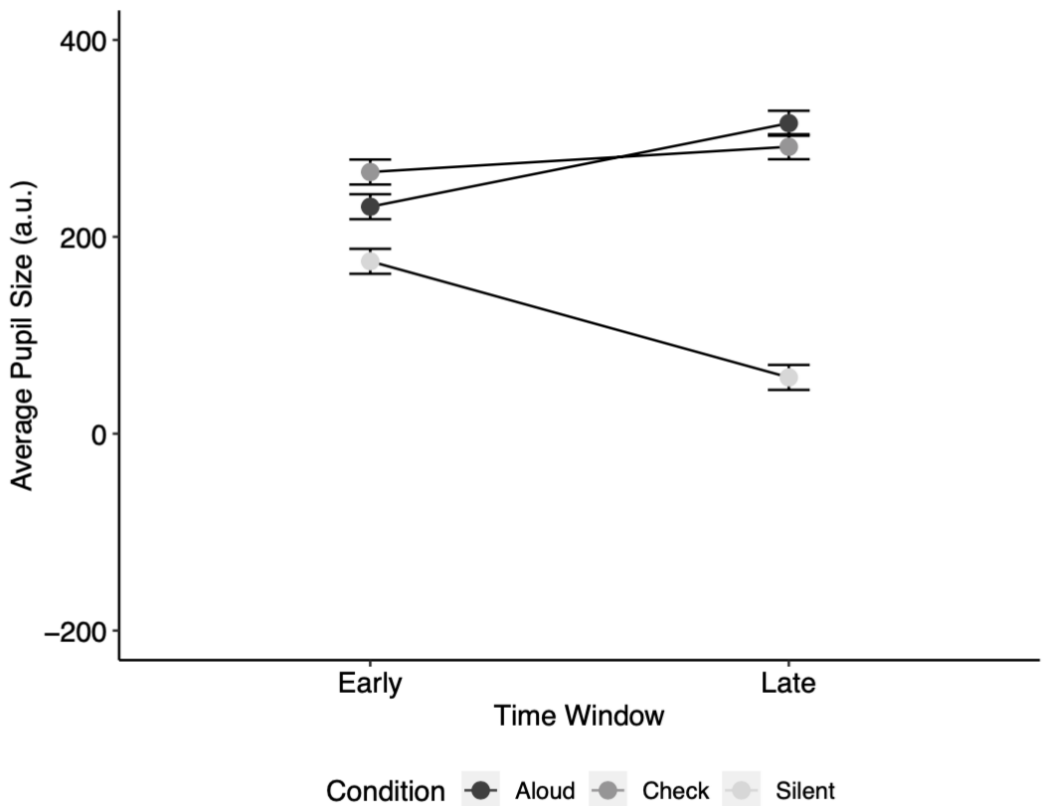

Figure 2.4: Average change in pupil size during early (1000-2000 ms) and late time-windows (2000-4000 ms). Error bars represent ½ FLSD.

Lastly, replicating analytical techniques used by Tiller et al. (2018), I conducted a correlational analysis to examine the relationship between the amount of mental effort exerted (i.e., change in pupil size) during each time window and the magnitude of the production effect; this was intended to determine if the increase in pupil size during study was predictive of later memory performance at test. To this end, I calculated the magnitude of the *pupillometric production effect* for each participant defined as the difference in pupil size between the aloud and silent conditions within each time window. Similarly, a measure of the behavioural production effect was calculated for each participant by subtracting the recognition performance for silent words from performance for aloud words; correlations were carried out to assess the relationship between these two variables for both the early and late time-windows. If cognitive processing (reflected in increased pupil diameter) was related to the resulting behavioural production effect then there should be a relationship between that and the pupillary production effect. Importantly, in Experiment 1 during the early time-window, I measured pupil activity in response to instruction alone, during which no distinctive information was presented; therefore, a relationship between pupil size and memory performance would be indicative of alternative (albeit possible preparatory) cognitive processes apart from distinctive processing.

Prior to analysis, inspection of these data revealed a small number of apparent multivariate outliers. To evaluate this possibility, I used the minimum covariance determinant (MCD; Fauconnier, & Haesbroeck, 2009; Hubert, & Debruyne, 2009). The MCD detects outliers by identifying extreme standard deviations of the outliers from the mean or center data points. Unlike other methods, the MCD adjusts for the effect that the outliers have on the standard deviation of the dataset which is why it is preferred for multivariate datasets and datasets with high covariance. I implemented this test using the *Routliers* package in *R* (Delacre & Klein, 2019). This analysis indicated four participant outliers; two of these were present in both time windows, one was present in the early time window only and the last was present in the late time window only. I completed the correlational analyses twice; the first time including the indicated the outliers and the second time after removing them. The outcome of both analyses is visually depicted in Figure 2.5.

The first analysis revealed a positive relationship during both the early, r = .30, t(54) = 2.30, p = .025, *95% CI* [0.04, 0.52] and late r = .51, t(54) = 4.40, p = <.001, *95% CI* [0.29, 0.68], time windows, replicating the observations of a positive relationship between these measures previously reported by Tiller et al. (2018). After excluding the previously identified outliers, a significant positive correlation was still observed in both the early, r = .30, t(51) = 2.26, p = .028, *95% CI* [0.03, 0.53] and late, r = .53, t(51) = 4.51 , p <.001, *95% CI* [0.31, 0.70] time windows. Again, the results were in line with previous research (Tiller et al., 2018); after excluding the outliers the strength of the positive relationship in both the early and late time windows was maintained. Of particular interest, the results confirm that there was a relation between cognitive processing during the pre-cue and the subsequent behavioural production effect, prior to the distinctive information provided by the stimulus. This suggests that there are preparatory cognitive processes that might facilitate the production effect. Similarly, an increase in pupil size in the late time-window reflecting processing that occurred during stimulus encoding and production was also related to better memory performance. Considered in whole the correlational analysis suggests that both preparatory processes and those that occur during production contribute to the memory outcome.

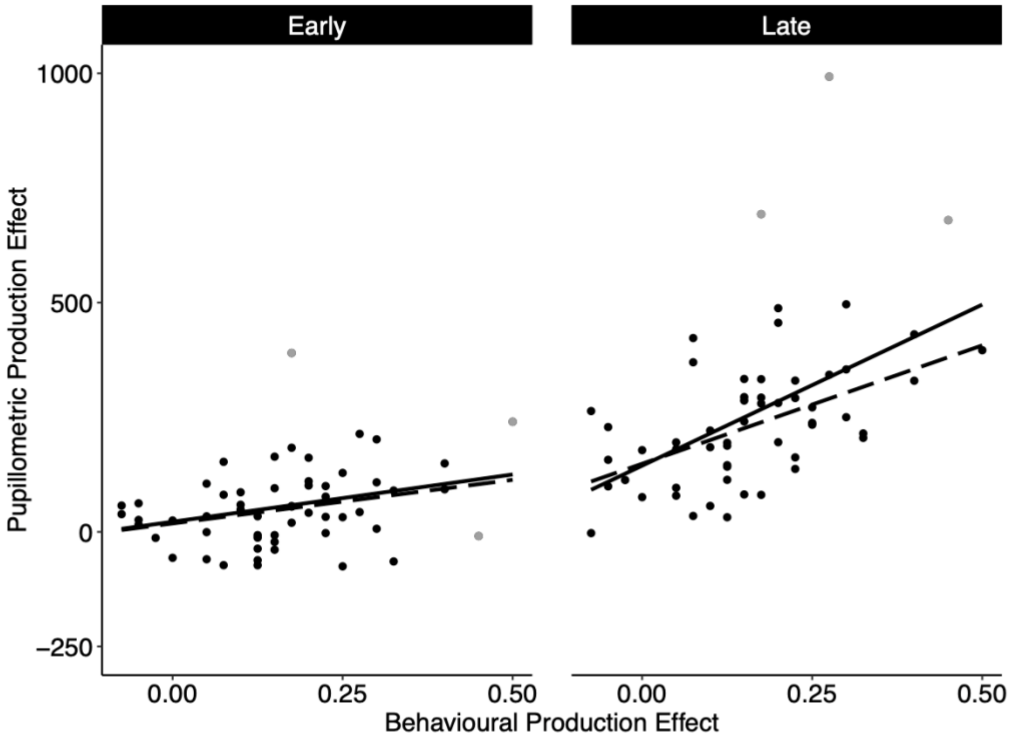

Figure 2.5: Relation of the pupillometric and behavioural production effect during early (1000-2000 ms) and late time-windows (2000-4000 ms). The solid regression lines demonstrate the relationship with all data included, the dotted regression lines represent the relationship after the removal of the outliers. Grey dots represent outliers.

An increase in pupil size was also evident in the check condition but unlike the aloud condition the behavioural data did not display an apparent increase in memory. A second correlational analysis was completed to determine if the change in pupil size during check trials would also be related to later memory performance. Perhaps, given that pupil size increased but memory did not, I would observe a negative correlation. To assess this possibility I repeated the correlational analysis for the check condition. A behavioural measurement of the *check effect* was obtained by subtracting each participant's silent recognition scores from their check scores. Next, the *pupillary check effect* was calculated by subtracting the magnitude of the pupil signal on silent trials from the check trials. Remaining consistent with the previous correlation the MCD was used once again to check for outliers. This analysis revealed a number of outlier participants, two of which were present in both time-windows; an additional three and four outliers were identified in the early and late time windows respectively.

The initial analysis, including all participants, revealed no significant correlation across both the early, $r = .00$, $t(54) = -0.02$, $p = .98$, *95% CI* [-0.27, 0.26] and late, $r = -.07$, $t(54) = 0.49$, $p = .63$, *95% CI* [-0.20, 0.32], time windows. The results of this analysis confirm that the change in pupil size during check trials was not related to later memory performance. Furthermore, I did not find evidence that pupil size was related to worse performance for check items. Removing the previously identified outliers did not change the outcome of the analysis; there was still no significant relationship between memory during check conditions and the pupillary signal in either the early, $r = -.12$, $t(49) = -0.83$, $p = .41$, *95% CI* [-0.38, 0.16], or the late, $r = .11$, $t(48) = -0.77$, $p = .44$, *95% CI* [-0.38, 0.17], time windows. The results of these analysis both are depicted in Figure 2.6.

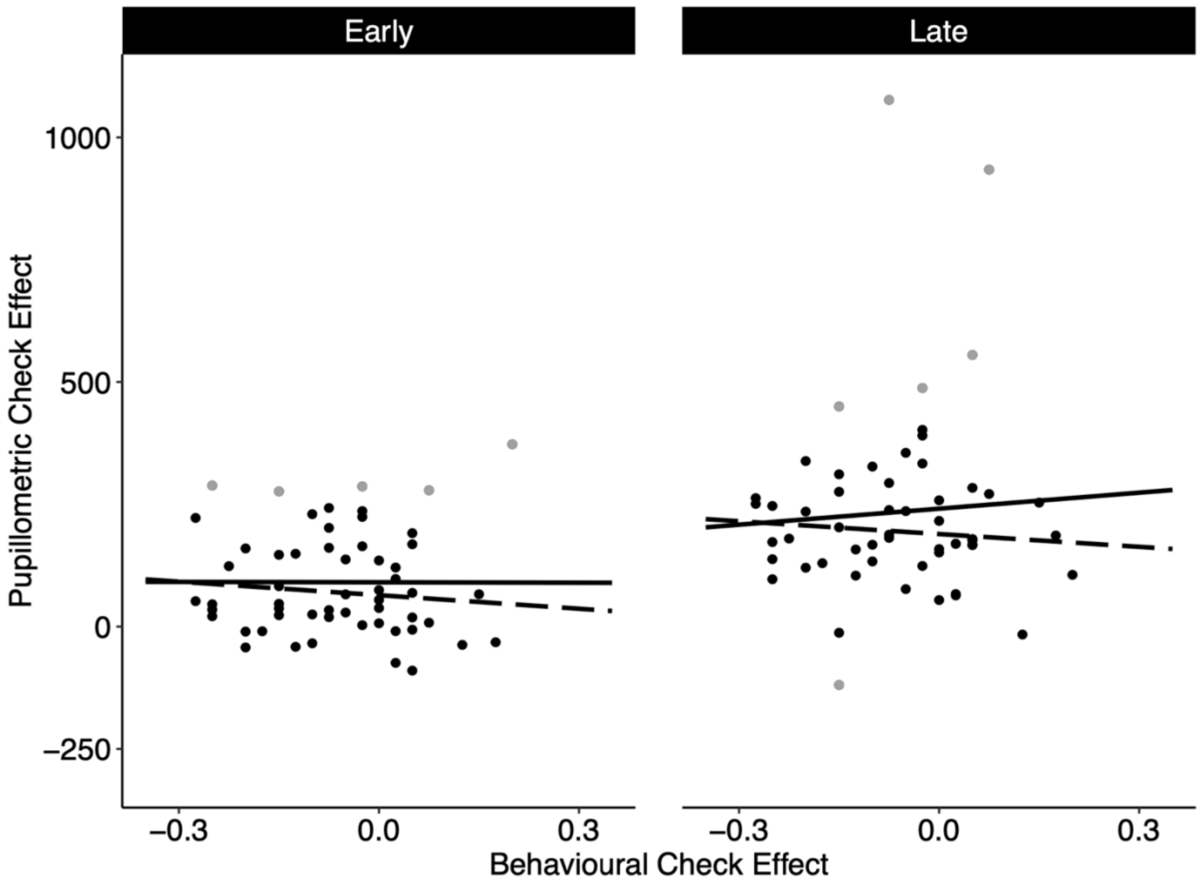

Figure 2.6: Relation of the pupillometric and behavioural check effect during early (1000-2000 ms) and late time-windows (2000-4000 ms). The solid regression lines demonstrate the relationship with all data included, the dotted regression lines represent the relationship after the removal of the outliers. Grey dots represent outliers.

2.5 Discussion

In Experiment 1 the study procedure was modified by adding a pre-cue prior to the stimulus to assess whether participants would respond differentially to the instruction alone. Specifically, differential responses were measured using pupillometry to assess changes in pupil dilation that occurred over the trial as a means of gauging cognitive effort. This was motivated by Tiller et al. (2018) who previously demonstrated that pupil size changed in response to production condition, but did not separate the cue from the stimulus or the productive act. By separating the cue from the stimulus, I could infer whether participants responded differently to the cue alone, even if there was no stimulus present. Pupillary differences observed during this pre-cue period would be unexpected from a distinctiveness account perspective and might indicate a role of attention in the production effect, an idea that was previously evidenced by other studies (Mama et al., 2018, Mama & Icht, 2018a). I expected that the remainder of the trial would replicate the previous pupillometric findings with pupil size being greater in the aloud and check conditions compared to the silent ones due to the additional effort of speaking aloud (Tiller et al., 2018). Lastly, I predicted that the correlation between the magnitude of the behavioural production effect and the pupillometric production effect would be replicated. This finding would strengthen the notion of a relationship between cognitive effort at encoding and the outcome on later memory performance. As predicted pupil size diverged during the pre-cue; unexpectedly, the pupil size in the check condition increased to a greater degree than in the aloud condition, and both were greater than the silent condition. A correlation between the pupillometric and behavioural production effect was also observed in both time windows; the correlation during the early time window was particularly surprising as participants were not actively encoding the stimulus at this time. These findings will be further elaborated on in the upcoming section.

2.5.1 Conclusions from behavioural data. First, a production effect was observed, evidenced by a greater proportion of "yes" responses to aloud words than silent words on the recognition test. This finding was expected as the production effect is robust in a within-subject design and has been replicated extensively (MacLeod & Bodner, 2017). Additionally, I observed a difference in the comparison of check and silent recognition scores with performance for silent items being significantly greater than check items. This finding was unexpected because MacLeod et al. (2010) had previously examined production with a repeated vocal response (saying ‘yes’) and obtained equivalent means between silent and check recognition scores. However, this discrepancy might have arisen because MacLeod et al. (2010) presented the stimulus and cue simultaneously whereas in the present study participants were presented with the instruction during a pre-cue. As a result, the current participants could prepare an exact response during the check condition. Since participants were already prepared to say “check” when the stimulus appeared on screen, their primary motive may have been to say “check”, and memorizing the stimulus could have become secondary to the task that they were already engaged in, resulting in less effort being used to encode the item. (This notion will be further explored in Section 4.1; Tyler, Hertel, McCallum and Ellis, 1979.) Related to this, after participants say “check” they may feel that they have completed the task and therefore did not return to encode the item that was presented, thereby not memorizing the item as well despite the cognitive effort exerted in planning and executing their non-specific vocal response.

A second difference between the present study and MacLeod et al. (2010) is that the current study included the repetitive verbal response as a third experimental condition whereas MacLeod et al. only compared silent words to a repetitive overt response. In fact, the presence of a third condition could have impacted recognition for the check items. Other studies that used a third condition have demonstrated mixed findings with some showing a similar trend as the present work, albeit not significant (Tiller et al., 2018; Fawcett et al., 2020); while others have observed a trend in the opposite direction (Roddick, Fawcett, Newman, Lambert, & Bodner, 2014). Adding a third condition changes the nature of the experimental task. For example, the addition of a third condition may increase cognitive load exerted on the task by invoking task switching costs combined with the extra effort of engaging in two unrelated mental tasks (i.e., processing the stimulus and saying “check”; Baddeley, Lewis, Eldrige &Thomson, 1984; Meiran & Chorev, 2005). While the outcome of having a third condition might have relevance in terms of its impact on experimental design, this finding does not have much significance in application. In practice, someone is unlikely to make an unrelated overt response when trying to memorize something.

2.5.2 Conclusions from pupillometric data. In Experiment 1, I had three hypotheses that were assessed using the pupillometric measure. First, I predicted that pupil size would differ in response to the instructional cue prior to both the presentation of the stimulus and production. Second, I predicted that these differences would be maintained for the duration of the trial and that I would replicate Tiller et al.’s (2018) observation that pupil size in the silent condition would decrease over the course of the trial. Lastly, I predicted that a second outcome from Tiller et al.’s prior study would be replicated in the present study; specifically, that I would also observe a relationship between pupil change and the production effect. This correlation was expected to demonstrate that the magnitude of the pupillometric effect was related to the magnitude of the resulting behavioural effect, implicating a relationship between the amount of cognitive effort used at encoding and the resulting memory performance. The presence of such a relationship would allow me to make a direct connection between cognitive effort and the production effect.

First, I will discuss the outcome of the early time window which spanned from 1000-2000 ms, including the pre-cue and the first 500 ms of the stimulus presentation (as mentioned in

Section 2.4.2, the pupil response takes time to respond and so it is unlikely that cognitive effort related to stimulus processing was captured in this window; Hoeks & Levelt, 1993). My windowed analysis confirmed that during the early time-window there were significant differences between all three conditions with pupil size in the check condition showing the greatest increase. Remarkably, task-related differences in mental effort emerged in response to the cue alone; since the stimulus was not on screen there was no distinctive information to be encoded. Further, during this early time window the pupillometric production effect was predictive of the behavioural production effect, meaning that change in pupil size predicted later memory performance even prior to the distinctive encoding elements. Importantly, the increase in pupil size prior to presentation of the stimulus is difficult to reconcile within a distinctiveness framework, suggesting instead that participants become somehow more alert or engaged in anticipation of production. The fact that these motivational differences were themselves predictive of the magnitude of the behavioural production effect supports the argument that participants are more attentive during aloud trials and that attention plays some role in the effect.

A related and unexpected finding was that pupil size in the check condition increased to a greater degree than the aloud condition during the pre-cue. I propose that during the check trial participants know that they will be saying “check” and so are preparing to say an exact response, whereas in the aloud trial they are just preparing to speak. This will be addressed in Experiment 2 where the encoding of the stimulus and production will occur at separate times. My outcome from Experiment 1 supports the notion that these early pupillary responses are preparatory in nature and that they might facilitate the production effect by focusing mental processes towards the word to be produced. Further evidence comes from the pupil increase during the aloud condition within the same period of time, indicating that pupil change reflects the cognitive effort used as participants prepare to speak aloud in the upcoming portion of the trial. Evidence for a link between pupil size and attention comes from studies on task engagement and mind-wandering which have noted that pupil size is greater when participants are on-task and completing a task that keeps their attention externally engaged (Unsworth & Robinson, 2018). The difference in pupil size between the two speaking conditions and the silent condition might similarly reflect differences in task engagement; knowing that they will be speaking, participants focus their attention to the task at hand, as evidenced by an increased pupil diameter.

Next, I will consider the pupillary differences that emerged in the late time window. This period related to my second hypothesis that pupil change would be sustained in the aloud and check conditions but drop in the silent condition. Furthermore, if saying the word aloud was more effortful, I expected that pupil size in the aloud condition would be greater than in the check condition. However, although average pupil diameter was numerically larger in the aloud than the check condition during the late time window, subsequent windowed analysis determined the difference was not significant. Nevertheless, this trend is in line with my explanation for the observed increase in the check conditions during the early time window; when participants were beginning to speak in the aloud condition during the late window they were preparing and giving an exact response, leading to an increase in pupil dilation above the check condition. Again, correlational analysis during the late time window confirmed a relationship between the pupillometric and the behavioural production effect. The greater increase in pupil dilation during the aloud condition compared to the silent condition was predictive of better encoding of aloud words compared to silent ones. This suggests that participants were exerting greater cognitive effort during aloud trials which facilitated encoding of the stimulus.

Subsequently, I will consider the pupil trends observed in the silent condition across both early and late time windows. Initial exploratory analysis suggested that during the silent trials pupil size initially increased but then dropped over the time course of the trial. In fact, near the end of the trial, pupil size in the silent condition dropped below the baseline measurement. Windowed analysis again confirmed this initial observation; not only was the change in pupil size consistently smaller than the aloud and check conditions in both time windows, but there was a marked decrease in pupil size in the silent condition between the two windows (see Figure 2.4). This observation is consistent with the assumption that participants do not fully attend to or stay fully engaged with the stimulus for the entire trial. If participants were for example, rehearsing the word after they read it I would expect that pupil size would be maintained over a greater time period. The drop in pupil size during the silent trials is consistent with the "lazy reading hypothesis" (Begg & Snider, 1987; MacLeod et al., 2010).

Alternatively, the smaller pupil size in the silent conditions, especially during the later time window, might indicate less effortful processing. That is, in the aloud and check trials, pupil size increases because encoding the word and speaking aloud is more effortful. In the silent condition participants are using less effort by simply reading the word silently. This difference in observed effort does not necessarily indicate that attentional or effortful processes are driving the memory outcome. If the cognitive effort indicated by pupil increase in the two speaking conditions was driving memory performance, then I might also expect better memory for the check words. Because check words were not remembered as well as aloud words (indeed, they were remembered more poorly than the silently read words), this suggests that the observed production effect was still influenced by the distinctive processing of the aloud words. My correlational analysis further supported this notion showing no relationship between the magnitude of pupil change during check trials and later memory performance. Similarly, Fawcett et al. (2020) also used a third motor control condition and found greater activation of brain areas associated with production without observing improved memory for "check" items. They posited that production did not benefit control stimuli the same way as aloud stimuli since the word "check" was not useful as a retrieval cue as it was a repetitive verbal response. Furthermore, their brain imaging revealed greater activation in speech processing areas during aloud conditions suggesting that these words were better encoded. Nevertheless, the difference in pupil sizes which emerged during the early window are difficult to explain within a pure distinctiveness framework and suggest an attentional component that might facilitate or provide additional processing on top of the distinctiveness processing which occurs.

Chapter 3: Experiment 2

3.1 Overview

Experiment 2 was designed to parse the mental processes that occur during the production effect from the motoric act. Whereas in Experiment 1 the trial was designed so that mental effort could be assessed prior to the presentation of distinctive information, in Experiment 2 I used a delayed production design (Hassall et al., 2016; Mama & Icht, 2018b). Participants viewed both the stimulus and cue simultaneously but were asked to hold their response for a separate "Go" signal. Note that similar to the design used by Mama and Icht (2018b), the word disappeared from the screen during the Go cue meaning that participants had to use their immediate memory of the word to produce it. By using this design, I can examine mental processes in response to distinctive information (i.e., the stimuli) separately from the motor processes which occur during the productive act. Hassall et al. (2016) previously demonstrated a greater amplitude of the P300b signal (see Section 1.3.2) in response to the cue and stimuli in a similarly designed delayed production procedure. Given that, and the finding in Experiment 1 that pupil size differed between the three conditions (aloud, check, silent) prior to motoric production, I expect that a similar divergence in the pupillary response will be observed in Experiment 2 even prior to the productive act itself.

Similar to Experiment 1, I split my pupillary analysis across time-windows to compare different processes occurring across the course of the trial. For Experiment 2 three time-windows were chosen: (1) an early window spanning the initial presentation of the stimulus to assess initial encoding processes; (2) a middle window prior to production to capture processes that might occur to prepare for speaking; (3) a late time window including the verbal response to investigate processes related to motoric production. In Experiment 1, I observed greater pupil dilation in the check and aloud conditions in response to the cue alone compared to the silent condition, an effect which remained for the duration of the trial. I predict that in Experiment 2 a similar change in pupil size will occur with greater pupil dilation during the aloud and check conditions compared to the silent condition both before and after the participant initiates motoric production. Since I previously observed a response to the cue alone, I should see a similar preparatory response in this experimental design. However, in Experiment 1 participants saw the cue alone, prior to the distinctive information from the stimulus and I saw an increase in pupil size for the check conditions above the aloud condition. In Experiment 2 I hypothesize that the change in aloud pupil size will be equivalent or greater than in the check condition because participants will be not only preparing to speak but also encoding the distinctive elements of the stimulus.

Next, I hypothesize that there will be a divergence in pupil dilation during the Go cue. Specifically, I predict that the increase in pupil size in the aloud and check conditions will continue to grow at the onset of the Go cue in response to the effort used to produce the word (or say "check"). Since during the aloud condition a distinct response will be generated, pupil size will be greater in the aloud condition and sustained for a longer duration. Finally, in the silent condition I hypothesize that pupil diameter will drop at the onset of the Go cue, since participants will not have to make a response. These findings would be in line with observations made by Tiller et al. (2018) who observed a steep decrease in pupil diameter during silent trials whereas, aloud and check pupil change was more sustained; they attributed this to the possibility that participants were disengaging more quickly with silent items. I surmised that if this is the case, there should be a similar drop in pupil size during the delayed production task. If participants are less engaged with the silent items, I would expect them to disengage with the item by the time the Go signal appears since they will be able to read and encode the item at the outset of the trial. By using a delayed production procedure in Experiment 2 I aim to expand on my observations made in Experiment 1, namely to determine if attention or effort is an underlying cognitive process of the production effect.

3.2 Methods

3.2.1 Participants. Thirty-five students enrolled at Memorial University were either recruited through the university's psychology pool and completed the study in exchange for one course credit or were recruited by poster advertisement and paid $10.00 for an hour of their time. Participants were asked to refrain from wearing makeup. Data were preprocessed using the guidelines outlined in Section 2.2. The same cut-off criteria of 50% remaining data after removing trials to lost pupil signal was used. As all participants met this criterion the final sample consisted of 35 participants.

3.2.2 Materials. The experimental setup and stimuli were identical to those used in Experiment 1, with the exception of the addition of a new instruction symbol '=' used as the 'Go' signal. As depicted in Figure 3.1, this symbol was also represented by a box (720 pixels x 540 pixels) matched for luminance in centre screen and was used to cue the participant to speak after the word presentation.

3.2.3 Procedure. The general procedure for Experiment 2 closely resembled that of Experiment 1. The experiment consisted of the same four phases and the eye tracker was set up following the same procedure as the previous study. Whereas in Experiment 1 the participant was told to speak while the word was on the screen, in Experiment 2 they were instead asked to hold their response (on 'check' and 'aloud' trials) until the second screen when the Go cue appeared. During the silent condition participants were told to read the word to themselves silently and to focus on the center screen for the entirety of the trial. The remainder of the procedure was identical to that used in Experiment 1. For that reason, trial events were identical between these experiments for the setup, familiarization phase and test phase. Changes to the practice and study phase are detailed below.

Practice Phase An identical list of three-letter words was presented to the participant following the same protocol as Experiment 1. However, prior to presentation the participant was alerted that there would be a fourth cue, '=', that would follow the presentation of the words and was notified that verbal responses should be withheld until the symbol appeared.

Study Phase As depicted in Figure 3.1 the study phase for Experiment 1 differed from that presented in Experiment 2. Again, trials began with two blank 200 ms screens appearing before and after a 1000 ms blink screen. Then cue *and* the target item were presented concurrently for 3500 ms followed by the equal symbol for 1500 ms during which the participant could make their verbal responses. As in the previous study, an interstimulus was presented before and after the stimuli to control for changes in luminance.

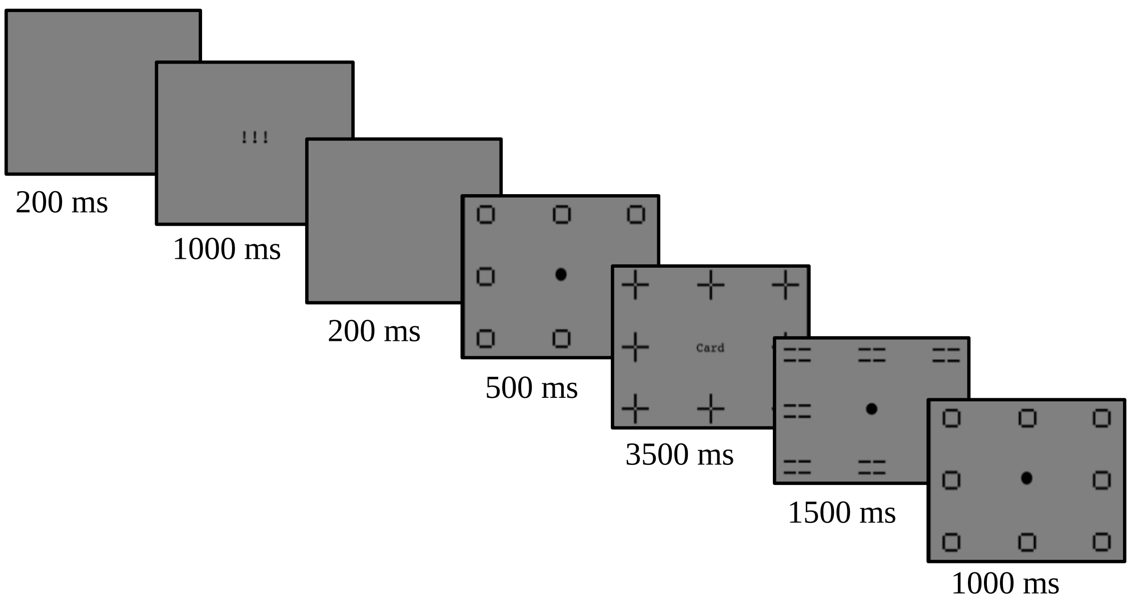

Figure 3.1: Outline of the screens presented to participants during a typical trial in Experiment 2. This experiment used a joint cue and item presentation followed by the second ‘Go’ cue represented by the “=”. The “+” symbols displayed in the present figure indicate that the participant should read the word aloud. The three exclamation marks denote that the participant should blink at that point. Duration is indicated below each event.

3.3 Pupil Processing

Experiment 2 followed the same preprocessing procedure as Experiment 1 (see section 2.2). Again, baseline was collected for the last 200 ms of the interstimulus placeholder appearing between the blink instruction and the stimulus onset (see Figure 3.1). There were no other differences in the procedures used.

3.4 Results

Statistical analysis was completed using the procedural guidelines and software outlined in Experiment 1. As before, I began with an analysis of the behavioural data, to establish the presence of a standard production effect.

3.4.1 Behavioural analysis. Following protocol from Experiment 1, test-phase confidence rating responses were collapsed into a binary response (yes/no)[5]. After, the proportion of "yes" responses for each of the test-phase conditions (aloud, check, silent, foil) were then averaged and compared using a repeated measures ANOVA. My analysis revealed a main effect of condition, $F(3, 102) = 161.29$, $MSE = <0.01$, $p < .001$, $\eta_G^2 = 0.58$. Consistent with Experiment 1, post-hoc comparisons depicted in Figure 3.2 demonstrated that saying the words aloud improved recognition relative to all remaining conditions. Contrary to Experiment 1, reading the word silently was not indicative of better performance than saying 'check', although 'yes' responses to silent words was still numerically greater than check. Additionally, in line

[5] Following Experiment 1 procedures confidence ratings were collected for later analysis outside the scope of the thesis.

with my previous findings, participants were more likely to respond “yes” to both of these items compared to foil items.

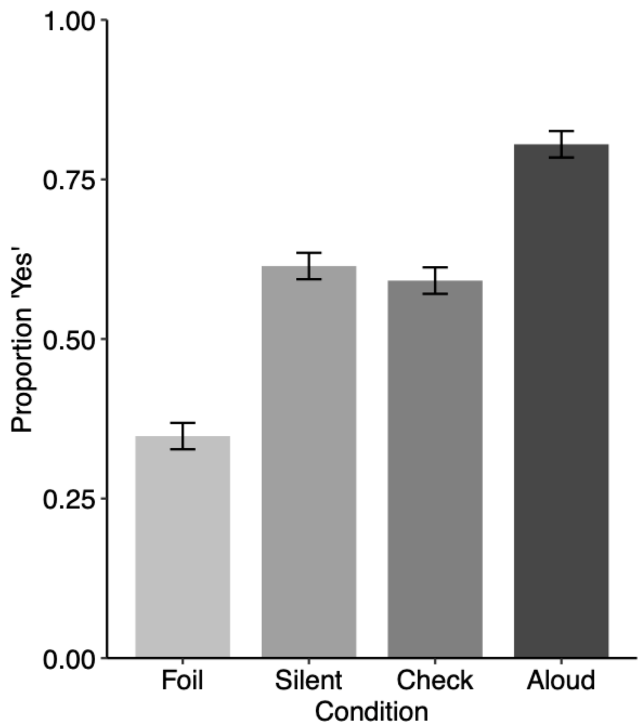

Figure 3.2. Mean proportion of ‘yes’ responses to items by condition (foil, silent, check, aloud), Experiment 2. Error bars represent ½ FLSD.

3.4.2 Pupillary analysis. Emulating the procedure used in Experiment 1, I then examined the pupillometry data. Again, I averaged change in pupil diameter collected during the preprocessing phase (see Section 2.2) across participants for the three conditions (aloud, check, silent) the resulting waveforms are depicted in Figure 3.3. Note that for the time comparisons a value of zero represents the onset of the stimulus.

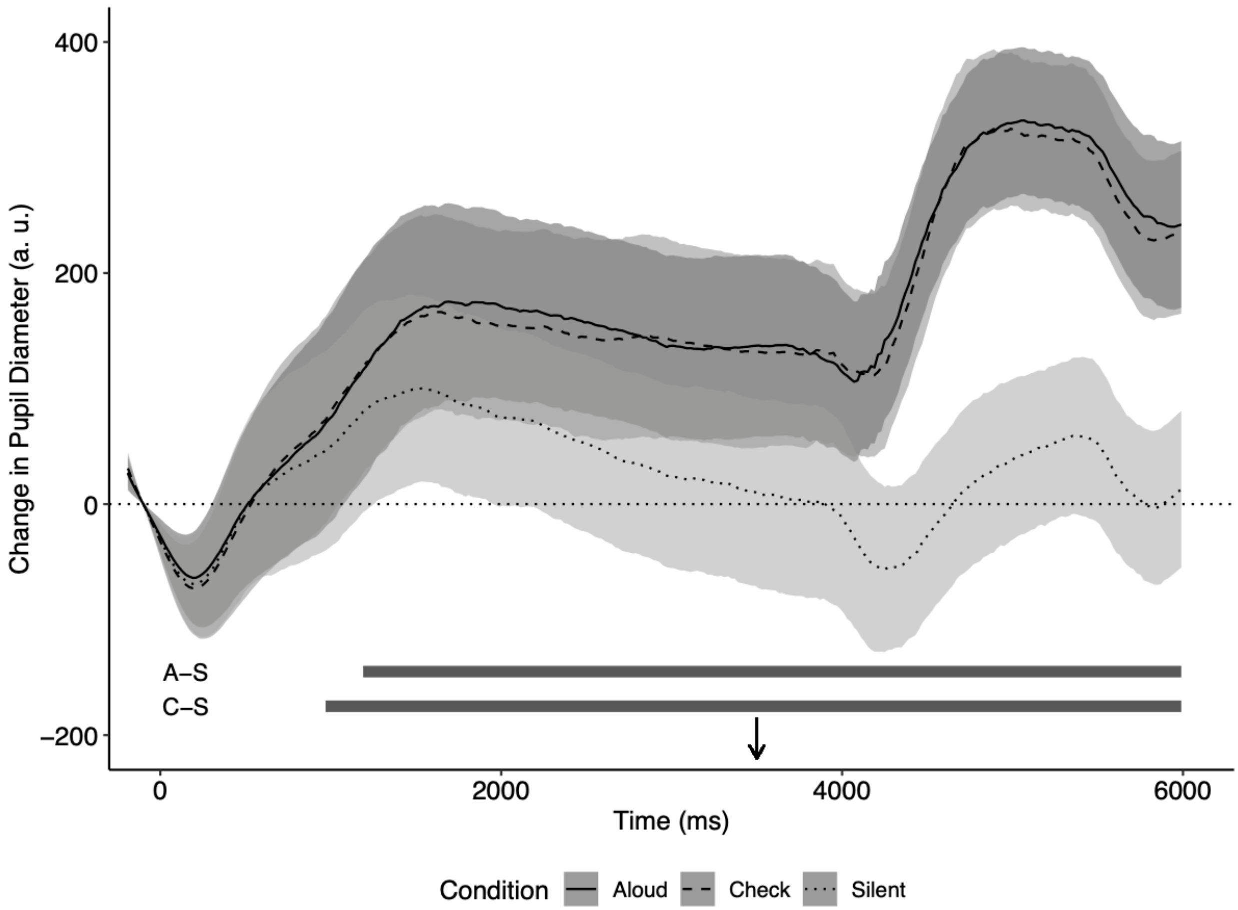

Figure 3.3. Mean change in pupil diameter (a.u.) as a function of conditions (aloud, check, silent). On the x-axis, the value 0 indicates the beginning of the stimulus presentation. The dotted horizontal line represents average pupil size during the baseline period. Error bars show 95% confidence intervals. The bars running along bottom of the graph reflect periods for which the appropriate conditions (e.g., A-C = Aloud – Check) differ significantly ($p < .05$). The aloud and check conditions did not differ significantly within any of the included time samples. Arrow indicates onset of ‘Go’ cue.

Maintaining consistency with Experiment 1, I completed an exploratory mass univariate analysis on the entire time series using t-tests comparing each condition at every sample within the time series (300 tests per time series). Error rate was again adjusted using the FDR method. Figure 3.3 displays the results from these comparisons, with significant differences being represented by the lines running horizontally across the bottom of the figure.

The comparisons revealed a similar trend when comparing both the aloud and check conditions to the silent condition. The increase in pupil size in both comparisons reached significance around 1000 ms and continued to grow as the trial progressed. These observations mirror those observed in Experiment 1; a larger increase in the aloud and check pupil change compared to silent is consistent with the notion that more mental effort was used on these trials. However, the third comparison between the aloud and check conditions suggested no significant differences between the two groups. The results from this final comparison lead to a different conclusion than Experiment 1 where I observed differences between the aloud and check conditions. This may be because in Experiment 1 participants were able prepare an exact response in the check condition only, but in Experiment 2 participants could prepare their response in both conditions.

I next proceeded to the planned time window analyses, this time using three predetermined windows of interest: (a) an “early” time window subtending 1000-2500 ms, (b) a middle time window subtending 2500-4500 ms, and (c) a late time window subtending 4500-6500 ms. Here, the early window was selected to capture the initial cognitive processing that would occur in response to the word and cue presentation. This window was analogous to the late window in Experiment 1, but combined cognitive processing in response to the word and instruction, while separating them from vocal production. The middle and late time windows were chosen to observe activity leading up to and following production, respectively. If there are preparatory processes that occur prior to production, I would expect them to emerge during the former, whereas the latter should capture processes specifically related to the productive act itself.

Mean change in pupil diameter was analyzed as a function of condition (aloud, check, silent) and time window (early, middle, late) using a repeated measures factorial ANOVA. Replicating Experiment 1, I observed a significant difference for condition, $F(2, 68) = 52.85$, $MSE = 16478.00$, $p < .001$, $\eta_G^2 = 0.10$ and a significant interaction, $F(4, 136) = 26.09$, $MSE = 4355.08$, $p < .001$, $\eta_G^2 = 0.03$. Diverging from Experiment 1, my analyses also revealed a significant main effect of time, $F(2, 68) = 16.06$, $MSE = 20109.92$, $p < .001$, $\eta_G^2 = 0.04$. Post-hoc analyses are displayed in Figure 3.5; as depicted, a similar trend emerged across all three time windows, with a larger increase in pupil size for aloud and check trials than for silent trials. However, the difference between the aloud and check conditions was non-significant during each time period. The interaction is driven by a significant increase in pupil size for both the aloud and check conditions from the middle to the late time window, while there was no change in the silent pupil size between these to time windows. Comparatively, pupil change in the silent condition is significantly greater in the early window compared to the middle and late windows, which do not differ. Considering this, it appears that during aloud and check conditions participants were allocating an increasing amount of mental effort to initially examine the word, prepare for a vocal response, and produce the word (or say “check”). In the silent condition a steady decrease in pupil size suggests that mental effort lessened over the course of the trial.

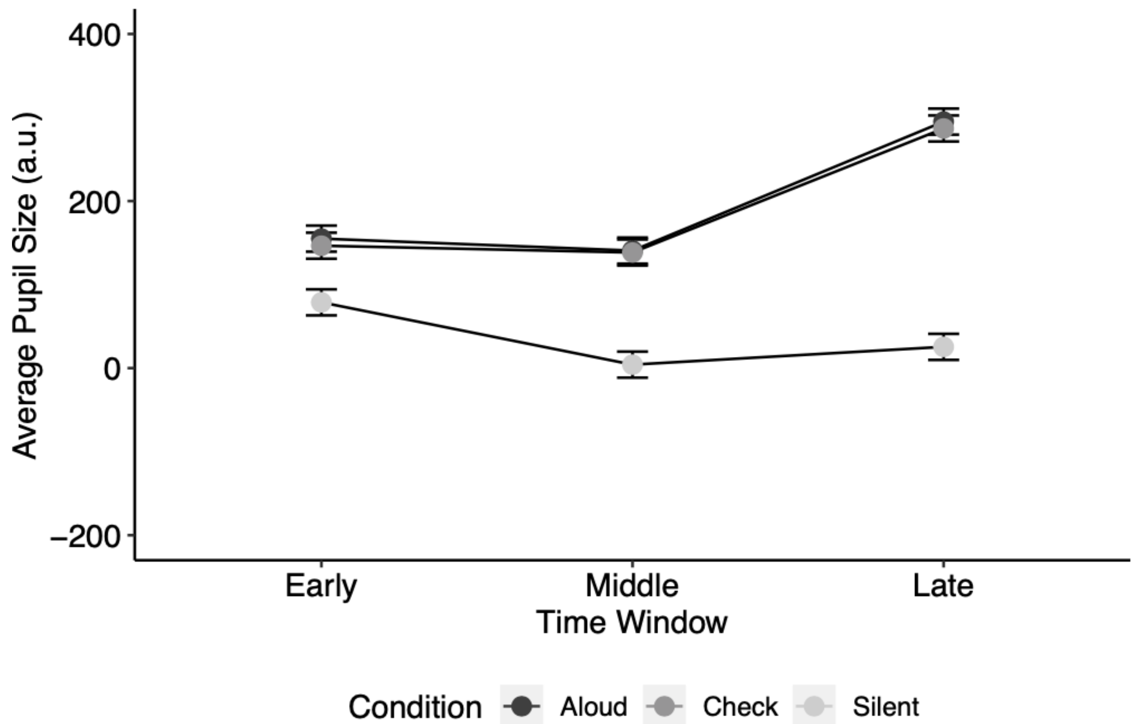

Figure 3.4: Average change in pupil size during early (1000-2500 ms), middle (2500-4500 ms) and late time-windows (4500-6000 ms). Error bars represent ½ FLSD.

Finally, emulating the analysis previously completed in Experiment 1 and by Tiller et al. (2018), I completed a correlational analysis assessing the relationships of the magnitude of the behavioural production effect (improved memory for aloud words over silent) with the magnitude of the pupillometric production effect (increased mental effort as evidenced by an increased difference in pupil size for aloud compared to silent items). A separate analysis was completed for each of the three-time windows. Following the same protocol as in Experiment 1, I identified outlier cases in the data using the MCD. This analysis confirmed six cases as outliers; of these two were outliers only in the early time window, one only in the late window, and the other three were present in both the middle and late time windows. The latter cases reflected points with a pupillometric production effect of close to zero, despite a large behavioural production effect. One of these participants had already been flagged earlier due to issues surrounding data recording (i.e., their thick eyeglasses made them difficult to calibrate), but they were ultimately included because inspection of their pupillometric data following recording revealed it to not be *particularly* noisy. Due to the presence of these outliers, I conducted the correlational analyses twice – once including and once excluding those participants.

The first analysis including all participants revealed no significant relationship between the pupillometric and behavioural production effect across the early $r = .10$, $t(33) = 0.55$, $p = .59$, *95% CI* [-0.25, 0.42], middle, $r = .01$, $t(33) = 0.07$, $p = .94$, *95% CI* [-0.32, 0.34], or late, $r = .10$, $t(30) = 0.59$, $p = .56$, *95% CI* [-0.24, 0.42], time windows.

The subsequent analysis excluding the previously detected outliers revealed a significant positive correlation in the late time-window only, $r = .44$, $t(29) = 2.70$, $p = .01$, *95% CI* [0.11, 0.69], although the middle time window was marginally significant, $r = .33$, $t(30) = 1.88$, $p = .07$, *95% CI* [-0.03, 0.61], and the early time window trended in a similar direction, $r = .17$, $t(31) = 0.94$, $p = .35$, *95% CI* [-0.19, 0.48]. After removing the outliers the correlational analysis still revealed a smaller relationship than expected. Importantly, the present sample size is lower than had been intended owing to recruitment difficulties and early closure of the university due to a global pandemic, leaving my analyses underpowered to detect correlations aligned with those reported in the earlier experiments (power to detect $r = .3$ is only 41.47%). Furthermore, it is also possible that the use of three different time windows diluted the measure: In Experiment 1 the time trial was split into two windows. Furthermore, Tiller et al. (2018) sampled one window which was selected 1000 ms after cue and stimulus onset, until after verbal production was completed, capturing the entire production process. Importantly, in both of these cases the processing of the stimuli and production occurred at the same time, possibly strengthening the relation to subsequent performance. In the current experiment initial encoding processes (i.e., viewing the stimulus) and production (i.e., saying the word aloud) were split into two time-windows. Since production and encoding occurred at different times the peak amount of mental effort used might have been less and more dispersed than when production and encoding happen simultaneously.

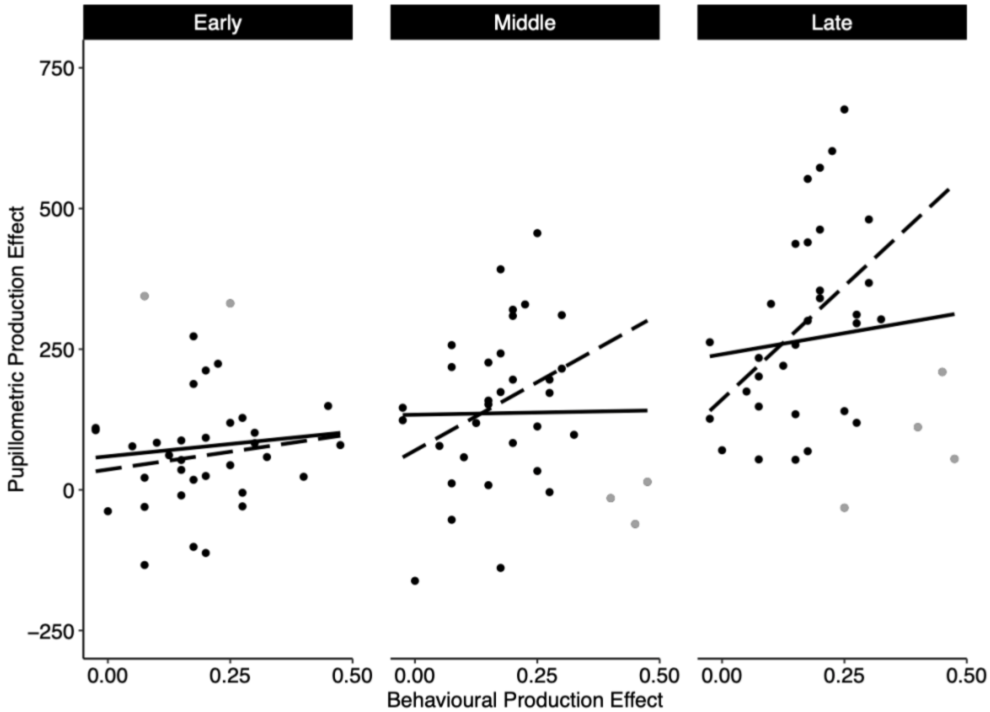

Figure 3.5: Correlation between pupillometric and behavioural production effect across the early (1000-2500 ms), middle (2500-4500 ms) and late time-windows (4500-6000 ms). The solid line represents linear regression with the inclusion of all data points. The dashed line demonstrates linear regression after the removal of the outliers. The outlier points are represented by the grey points.

Next, I completed a second correlational analysis to determine if there was a relationship between increased pupil size during check conditions and later memory performance. This time I assessed the relation of the pupillometric check effect (difference in the magnitude of the pupil signal between silent and check) and the behavioural check effect (difference in recognition between silent and check items). Once more, outlier analysis revealed one participant whose data was indicated as an outlier in all three time windows. Two separate participants were outliers in either the early or late time windows only.

With the inclusion of all data points, the analysis revealed no significant relationship across the early r = -.10, t(33) = 0.60, p = .55, *95% CI* [-0.42, 0.23], middle, r = -.05, t(33) = -0.29, p = .77, *95% CI* [-0.37, 0.29], or late r = -.02, t(33) = -0.14, p = .89, *95% CI* [-0.36, 0.31], time windows. Replicating the results in Experiment 1 the outcome of the analysis suggests that there was no relationship between pupil change and later memory during the check condition. Subsequent analysis following the removal of the outlying data points revealed still insignificant correlations across the early, r = .13, t(31) = 0.78, p = .44, *95% CI* [-0.21, 0.46], middle, r = .12, t(32) = 0.71, p = .49, *95% CI* [-0.22, 0.44], and late r = .07, t(31) = 0.37, p = .71, *95% CI* [-0.28, 0.40], time windows. The results from this correlational analysis are depicted in Figure 3.6.

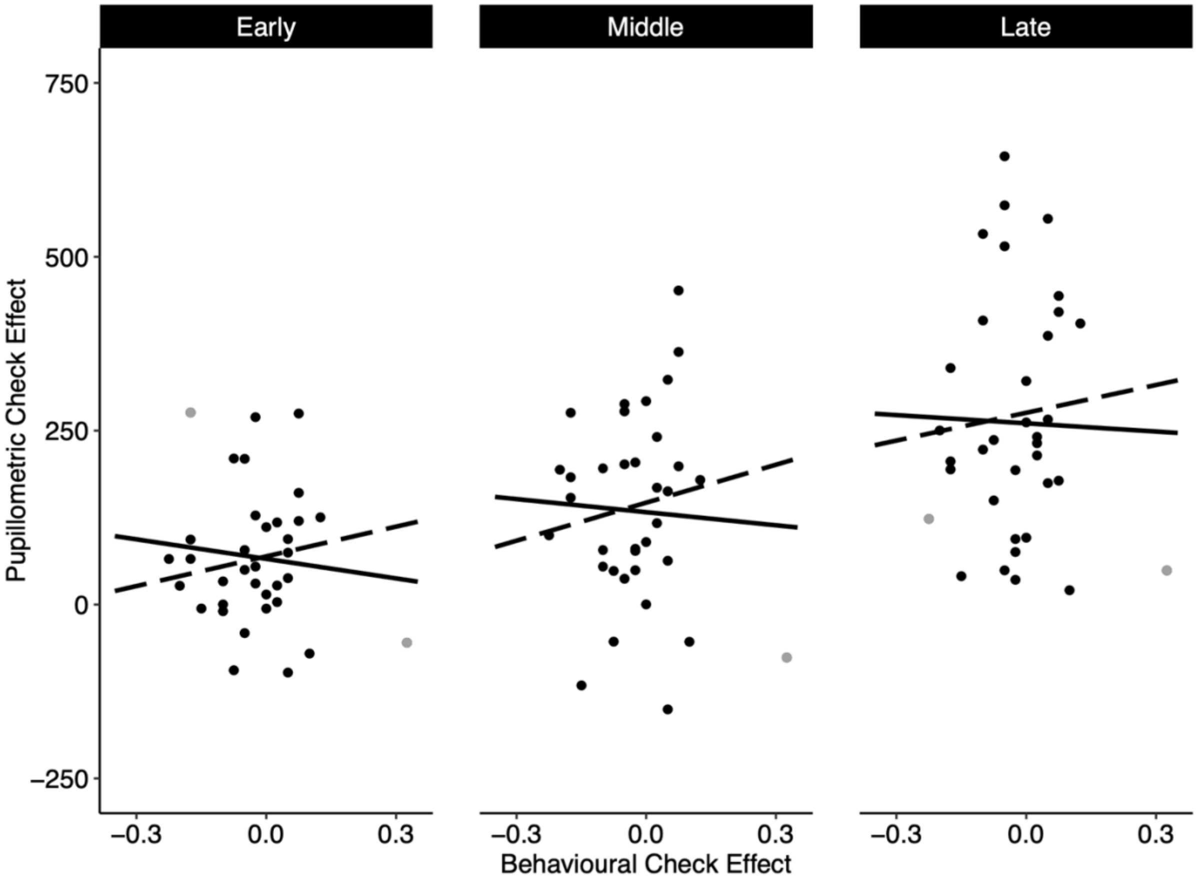

Figure 3.6: Correlation between pupillometric and behavioural check effect across early (1000-2500 ms), middle (2500-4500 ms) and late time-windows (4500-6000 ms). The solid line represents linear regression with the inclusion of all data points. The dashed line demonstrates linear regression after the removal of the outliers. The outlier points are represented by the grey points.

3.5 Discussion

The goal of Experiment 2 was to measure cognitive responses to not only the instruction (as in Experiment 1) but also to the distinctive aspects of the word, still separated from the productive act itself. To do this, I implemented a post-cue during the study phase where participants had to wait until a later “Go” cue was presented to produce the word. Using this procedure I was able to elaborate on my finding of a preparatory process in Experiment 1 and to observe if mental effort would differ by condition when encoding the stimulus was separated from production. In Experiment 1 a preparatory effect was evidenced, such that pupil size deviated in response to the instruction alone; I took this to mean that participants were engaging in attentional or effortful processes to prepare themselves for the upcoming trial. Considering this, and that other studies have also evidenced cognitive differences prior to production when the stimulus was pre-cued (Hassall et al., 2016), I expected to see a similar pupil response in the present experiment prior to production.

Furthermore, by using a post-cue procedure I could address the idea that participants are less attentive during silent trials. If this is the case then I expected participants would readily disengage with the material, especially if they have already encoded the word. This could shed light on motivational differences in the silent condition: If participants do not stay engaged or try to rehearse the item after it leaves the screen it would suggest that participants are less motivated to memorize the silent items. Lastly, since the previous experiment and Tiller et al. (2018) both demonstrated a positive relationship between the pupillometric and behavioural production effects; I expected to observe a similar trend across all three time windows reflecting a relationship between the amount of cognitive effort being used during the aloud trials and the memory benefit of aloud words over silent ones.

In Experiment 2 I observed a divergence in pupil size in the two vocal conditions from the silent condition early in the trial, before participants spoke aloud, supporting my prediction that I would observe differences in pupil size prior to production. I suggest that this change represents preparatory processes that reflect the effort of preparing to speak aloud. Additionally, I observed a decrease in pupil size during the silent condition over the course of trial, suggesting that participants were disengaging in the trial after encoding the word. Finally, I found a relationship between the pupillometric production effect and the behavioural production effect in the late window only. These observations will be discussed in greater detail in the coming sections.

3.5.1 Conclusions from behavioural data. Once more I observed a production effect with a greater proportion of “yes” responses to items that were said aloud than to ones that were read silently. The finding of better memory for silent items over check items in the first experiment was not replicated in the second experiment. However, numerically the means trended in a similar direction. The previously discussed implications related to the impact of the inclusion of a third condition on the performance for check items (see Section 2.5.1) cannot be ruled out. Again, I observed a trend towards poorer memory for the check items compared to silent ones which is in agreement with the previously mentioned point that check items may incur a cost related to being cognitively engaged in two unrelated tasks or because of task switching costs (Baddeley et al., 1984; Meiran & Chorev, 2005). However, the relationship between change in pupil size and memory performance was found to be not related, suggesting that check memory performance was not influenced by cognitive effort.

3.5.2 Conclusions from pupillometric data. In Experiment 2, pupillometry was used to assess three hypotheses. First, I had predicted that a divergence in pupil size would occur prior to production, in line with Experiment 1 and Hassall et al.'s study (2016), both of which observed a difference in cognitive processing prior to the act of production. Second, I predicted that pupil size in the aloud condition would be greater than in the check condition both before and after the Go cue because preparing a specific response should be more effortful than saying a repetitive word. Third, I again expected to find a relationship between the pupillometric effect and the behavioural effect, such that a larger increase in pupil size during aloud trials compared to silent trials was related to better memory for aloud words compared to silent words, replicating both Experiment 1 and Tiller et al. (2018). Finally, for my fourth hypothesis I proposed that once again I would observe a smaller pupil increase during silent trials compared to both aloud and check trials, especially in the latter half of the trial after the stimulus disappeared.

To assess the first hypothesis, that changes in pupil size would occur prior to production, I examined data during the first two time windows. The windowed analysis confirmed that across the early and middle time windows, the increase in pupil size during the silent trials was significantly lower than both the aloud and check trials, which did not differ from one another. Part of my hypothesis was supported; differences were observed prior to production. However, pupil size in the aloud condition was not greater than the check condition across either time window. This finding did support my theory proposed in Experiment 1: That pupil size was greater in the check condition during the pre-cue period because participants were preparing an exact response. In Experiment 2, participants knew what they would be saying in both conditions, so it was expected that the pupil increase in the check condition would not be higher. However, a greater pupil change in both the speaking conditions was observed compared to the silent condition even though during these time periods participants were engaged with the same amount of information (i.e., reading the stimulus and being alerted to the condition). An increase in pupil dilation in the aloud and check conditions compared to silent trials suggests that participants were preparing to speak even early in the trial. This notion lends support to the hypothesis that there is a preparatory component to the production effect and again, may reflect differences in attention or cognitive effort related to preparing an aloud response. Partially addressing my third hypothesis, unlike in Experiment 1, my correlational analysis was not sufficiently powered to provide strong support that these processes were related to later memory performance. However, the correlations did trend in the expected direction with the middle correlation being marginally significant. Furthermore, the late time window revealed a significant positive correlation after the removal of the outliers. Additionally, analysis of the relationship between pupil size and memory for check items also revealed an outcome that was not significant. Moreover, as previously mentioned my correlational analysis was underpowered due to early interruption of data collection because of university closure in response to a global pandemic.

My second hypothesis proposed that pupil size for aloud words should be larger than check words across all three time windows, since in the aloud conditions participants would be preparing a unique response and that should be more effortful. Subsequent analysis of the time windows revealed that pupil change did not differ between the aloud and check conditions across any time window. This finding is in agreement with the distinctiveness account and suggests that speaking in either of the two conditions was equally effortful. However, only memory for aloud items benefited from vocalization suggesting that reading the specific word aloud enabled more effective encoding of the unique features of that word (Forrin et al., 2012); the check response was not unique and so those words did not benefit from the extra effort of speaking aloud (MacLeod et al., 2010). However, the results supported my third hypothesis: That there would be a relationship between the pupillometric and behavioural production effect. In the late time-window (when participants produced the word) the results did suggest a moderate positive relationship between the pupillometric and behavioural production effect. This provides a challenge for the notion that cognitive effort did not play a role in the production effect since pupil changes reflect cognitive effort and the correlation with later memory performance suggests that some of the effortful processing that was observed during the aloud trial was related to memory for those words. Conversely, I found no evidence suggesting that the increase in pupil size during check trials was related to memory performance. However, an alternate explanation would be that pupil dilation reflects an increase in distinctive processing: participants who are more engaged with the word might better encode the distinctive features of the stimulus resulting in improved memory.

Finally, I will discuss my final prediction that the pupillometric trends observed during the silent condition would decrease across the course of the three time windows. The windowed analysis confirmed that the change in pupil size during the silent condition was consistently smaller than the other two conditions during all three time windows. Furthermore, comparing changes in silent pupil size over the three time windows mirrored my Experiment 1 results with pupil size being greater in the early window compared to the other two windows which did not statistically differ. The second part of my prediction – that pupil size would drop after the production cue – was partially confirmed. As depicted in Figure 3.5, pupil size for silent words during the middle time window was close to baseline, and the waveforms plotted in Figure 3.3 reveal that silent pupil size dropped below baseline shortly after presentation of the Go cue. This observation suggests that after the cue was displayed participants may have felt that they had finished the task and did not stay engaged. However, after this cue pupil size slightly increased a second time during the late time window. Keeping this in mind, I propose that during the silent conditions, when the Go cue is presented participants are initially relieved that they do not have to speak (dip below baseline) but then are preparing for the next trial (increase to above baseline). A second explanation could be that participants are engaging in maintenance rehearsal and reprocessing the word; however, retrieving the word from memory would be expected to improve memory for those words which was not reflected in the behavioural data (see Greene, 1987 for a discussion on maintenance rehearsal).

Overall the results from Experiment 2 were largely consistent with a distinctiveness account of production with little divergence between the aloud and check pupil sizes observed. Nevertheless, some of the observations were consistent with those in Experiment 1. Namely, the decrease in the silent pupil signal indicating that participants used less effort to encode those words and that they disengaged with the words earlier in the trial then in the speaking conditions. Next, the emergence of pupil change prior to production in both experiments indicated that speaking aloud was more effortful and suggested that participants were allocating cognitive resources to prepare their vocal response early in the trial. Lastly, a relationship between the pupillometric and behavioural production effect was observed in both time windows in the first experiment and the late window in the second experiment, demonstrating that cognitive effort was predictive of later memory performance. Furthermore, I did not find evidence that increased pupil size was related to memory for check items. These findings are consistent with a distinctiveness account but are also indicative of differences in attention or cognitive effort between aloud and silent trials. In the final chapter I will attempt to reconcile the results within these frameworks.

Chapter 4: General Discussion

Producing content to be memorized has been touted as a simple mnemonic device as it requires little knowledge and practice. However, effective use requires an integrated understanding of the production effect and its underlying processes, which remain under debate. Heretofore, research has focused on whether distinctiveness adequately explains the mnemonic benefit of the act of production; other processes that could occur in addition to or separately from the act of production are less examined. Through the preceding experiments I took a new perspective to examining the underlying cognitive processes of this classic paradigm by applying a psychophysiological measure – pupillometry – to acquire a measurement of cognitive effort. The two experiments completed were modeled on a previous study completed by Tiller et al. (2018) and were modified to the end of replicating and elaborating on those previous findings. The outcome of the present studies evidenced the role of distinctive processing while also suggesting a role of attention or cognitive effort. While I do not believe that the current results contradict the distinctiveness account, I do posit that they provide evidence for a supporting process that can enhance or modulate the effectiveness of distinctive encoding processes.

4.1 Interpreting the Pupillary Signal

Across two studies I demonstrated the importance of distinctiveness, attention and cognitive effort in eliciting the production effect. In Experiment 1 a pre-cue procedure was completed with the primary goal of determining if participants would respond differentially to the instructional cue when presented alone. The experimental outcome confirmed that participants did respond differently to the cue even when it appeared before the word, as evidenced by different pupil sizes in response to the cue alone. I believe these processes to be preparatory in nature and that they exhibit mental effort used to focus one’s attention towards the experiment when participants know they will be completing a more demanding upcoming task. In comparing the two speaking conditions to the silent condition, the speaking conditions both elicited a larger pupil size before speaking, whereas during the silent conditions pupil change was much smaller in comparison. In support, Unsworth and Robinson (2018) previously revealed that pupils changed prior to completing a cognitive task and that an increase in pupil diameter meant that participants were focused on preparing for the upcoming task; the authors confirmed that participants were focused on the task by including thought probes during this time period. The early divergence in pupil change prior to the presentation of distinctive information is difficult to accommodate within the current distinctiveness framework. Nevertheless increased focus or engagement on the task might still facilitate memory through distinctive processing; this idea is explored further in the proceeding section.

Next, in Experiment 2 a delayed production procedure was used to elaborate on the preparatory responses observed in Experiment 1 (see Mama & Icht, 2018b; Thoms et al., 2020). Again, Experiment 2 revealed that pupil dilation changed significantly prior to production, indicating that participants engaged in differential processing depending on the experimental condition, this time while the stimulus was on screen. However, this difference was only observed between the silent condition and the two conditions which involved speaking. The aloud and check conditions did not significantly differ from one another, suggesting that the larger pupil size in these two conditions reflected cognitive responses that were related to preparing to speak aloud. Importantly, these changes occurred early in the trial well before the onset of the Go cue, which indicates that extra cognitive processes were recruited well before the onset of speaking and were used to prepare for the upcoming vocal response. Although I cannot directly infer that these differences are reflecting attentional processes, the observation of a divergence in the pupil signal during Experiment 2, prior to production is in line with the speculation that pupil change prior to vocal production is caused by preparatory processes. An idea previously proposed in response to the early pupil change observed in Experiment 1.

Pupil size was consistently smaller in the silent condition compared to the other two conditions across both experiments, a finding that corroborates earlier pupillary findings by Tiller et al. (2018). Furthermore, in both studies pupil diameter decreased over the course of the trial. A smaller pupil size indicates that encoding silent words was less effortful than encoding aloud or check words (Hess & Polt, 1964). In Experiment 1 the finding that pupil size was smaller while encoding the stimulus was not unexpected as this occurred during the same time that participants spoke in the aloud and check conditions. Therefore, it is unsurprising that encoding while speaking was more effortful than encoding the word in the absence of speech. The observations made in Experiment 2 are more difficult to explain; since the stimulus was presented before production in the aloud and check conditions, participants in all three conditions were reading the word silently to themselves during the early period of the trial. Thus, participants in the check and aloud conditions should not have been using more mental effort to encode the word; any additional effort was therefore most likely related to preparing to vocalize. It is possible that knowing they would be speaking, participants allocated more attentional resources to the word. Hassall et al. (2016) previously demonstrated differences in brain activity prior to production when the stimulus preceded speaking. Across both studies, pupil size decreased over the course of the trial in the silent condition; this observation was also thought to indicate changes in attentional processes. Tiller et al. (2018) also observed constriction of the pupil size during the silent condition over the course of the trial and suggested that this process reflected attentional disengagement, leading to lower memory performance for silent words.

Similarly, I suggest that my findings are supportive of the previous notion and reflect participants both engaging less with silent words and more readily disengaging from them. This finding supports the "lazy reading hypothesis" and suggests that participants do show differential engagement on silent trials (MacLeod et al. 2010).

A final trend observed across both studies was a similarity in pupil diameter between the check and aloud conditions which only significantly diverged during the pre-cue in Experiment 1. This similarity indicates that a comparable amount of mental effort was used when participants read the word, prepared to speak and during their vocal response during aloud and check trials. Despite pupil measures indicating that similar amounts of cognitive effort were used in these two conditions when the word was on screen, only the aloud condition showed improved memory performance; indicating that it was probably the distinctive information strengthened by the unique vocal response that was better encoded during the aloud trials and subsequently boosted memory for aloud words. Evidence for this notion comes from Fawcett et al. (2020) who observed similar differences between check and aloud trials in brain activation. As in the present study there was no memory improvement for the check conditions leading the authors to conclude that it was the distinctive information encoded on the aloud trials that improved memory for those items. The correlational analysis confirmed that pupillary change was related to the production effect during aloud trials but was not during check trials. The parallels between these two studies and the conclusions that can be made about memory performances during aloud and check trials will be discussed further in the upcoming section.

4.2 Implications for the Distinctiveness Account

The distinctiveness account of the production effect posits that producing words renders those words distinctive compared to words that are not produced and that these words benefit from additional unique information added by the act of production (MacLeod et al., 2010). Furthermore, it is thought that participants heuristically use this information at test to inform decisions about whether a word had been previously studied (Dodson & Schacter, 2001). Although the main findings were largely consistent with a distinctiveness account, some of my findings provide a challenge to a pure distinctiveness account and suggest that there are other active mechanisms that improve memory for aloud words during the production task.

First, the observed similarities in the pupil signal during the aloud and check conditions were comparable in both studies, only diverging during the early window in Experiment 1. These results are compatible with a distinctiveness account. If pupil size was reflecting mental effort related to the task and to later memory performance then I would have expected to see a greater increase in pupil diameter in the aloud condition compared to the check condition. The absence of a difference suggests that the effort that was used in these conditions was indicative of the effort related to speaking and not necessarily the encoding processes. Similarly, if cognitive effort improved memory by increasing the strength of produced words (Bodner & Taikh, 2012), then I would have expected to see an improvement in memory for the check condition as well, since pupil size also increased during these trials indicating more mental effort was used during encoding. However, memory performance for the check condition did not benefit from the observed cognitive effort and in fact was lower than the performance for silent words in Experiment 1. This indicates that aloud words benefited from additional distinctive processing that increased memory for those words at test (MacLeod et al., 2010). Although, note that mental effort in the check condition may have reflected cognitive effort allocated to other areas of the task, for example, engaging in two unrelated tasks might have contributed to mental effort in the check conditions; this idea will be explored further in the following section.

Additionally, the finding that memory performance improved for aloud words but not check words might be due to differences in participants' ability to use those responses to guide their decisions at test using the distinctiveness heuristic (Dodson & Schacter, 2001). Speaking was equally effortful in both conditions but only the aloud words were encoded better. The only difference between the two conditions was that in the aloud condition participants read the stimuli aloud – a unique response, whereas in the check condition participants said "check" – a repetitive verbal response. MacLeod et al. (2010) surmised that for production to improve memory the produced word had to be unique, otherwise it would not benefit from the additional distinctive processing. This suggests that memory performance for aloud words was improved because aloud words benefited from greater distinctive processing, resulting in a more elaborate encoding trace in one's memory. However, another factor influencing the performance difference between the aloud and check conditions might have been the ability to use this information to guide decisions at test. In the check condition participants produced a repetitive verbal response, therefore the response would be less useful to use heuristically at test. A similar inference was made by Fawcett et al. (2020) who demonstrated increased activation in the motor and auditory cortices during study in the aloud and check conditions but only observed a performance improvement in the aloud condition. Interestingly, they also observed activation of the auditory and motor cortices for control words during test, suggesting that participants were trying to access a production trace but since the trace did not have any defining information it did not aid memory performance. In conclusion, the similarities observed in pupil size in the aloud and check conditions in the present studies without an observed outcome in memory performance provide support to the distinctiveness account, suggesting that memory performance is largely the outcome of encoding the distinctive nature of the aloud words which are produced during study.

In Experiment 1, I demonstrated that cognitive effort differed by condition even prior to the presentation of distinctive information. This finding is difficult to explain with the original distinctiveness account since cognitive processes that occurred during that time could not have been related to distinct information. Providing a further challenge, my correlational analysis revealed that these differences were related to memory performance at test but only for aloud words. I suggested that these differences were likely reflecting preparatory processes, meaning that participants increased their attention to the task knowing the upcoming task would be more effortful. However, a theory that is more in line with a distinctiveness account is that differences in attention might not impact memory directly but rather contribute to the effect indirectly through supporting distinctive processing. For example, attention might facilitate production by enabling participants to encode more or better encode the distinctive elements of the stimulus, thereby enhancing the memory trace. The amount of attention that is allocated to the word would modulate the degree to which memory is improved.

Further evidence for additional active cognitive processes during production comes from studies relating dual-process theory to the production effect. This theory is based on the notion that recognition memory is driven by two discrete memory processes: recollection and familiarity (Yonelinas, 2002). Recollection is thought to be a process of retrieval where one makes a decision at test based on specific episodic traces about the encoding of the memory (e.g., tone of voice), whereas familiarity is a trace feeling that something has been seen but in the absence of specific details (e.g., a sense of having seen the word before but no details about the way it was presented). Since distinctive encoding involves storing specific information and using that information to inform decisions at test, it follows that recollection should be the driving process in production if the memory boost is purely the outcome of distinctive encoding (Fawcett & Ozubko, 2016; Ozubko, Gopie et al. 2012). In contrast, familiarity reflects a process akin to encoding strength or attention where memory is not guided by a specific recollective episode but a broader notion that something has been studied.

Ozubko, Gopie et al. (2012) were the first to investigate recollection and familiarity within the production paradigm. Across two studies they obtained separate measures for recollection and familiarity and by comparing these measures across silent and aloud conditions they concluded that producing words improved both types of memory. They observed an increase in recollection, highlighting the important role that distinctive processing plays in the production effect. However, they also observed a significant increase in familiarity for words that had been produced. Importantly, familiarity for aloud words remained even when silent words were also strengthened through repetition. This led to the conclusions that familiarity was unlikely reflecting a difference in encoding strength, instead the authors suggested that attention might be at play, bolstering the notion previously put forth that attention might play a supporting role in the production effect (see MacDonald & MacLeod, 1998) while also showing evidence for dual-processing.

Later, Fawcett and Ozubko (2016) replicated the finding that production increased both familiarity and recollection again evidencing dual-processing in the production effect. Unlike Ozubko et al. (2012), they proposed that familiarity was reflecting strength processes but added that increased attention mediated encoding strength, such that increased attention would elicit stronger encoding and thereby improve memory. Interestingly, the authors also completed a meta-analysis including studies that had measured recollection and familiarity in both within-and between-subjects designs. They observed that in a between-subjects design the production effect was driven by familiarity, suggesting that in a within-subject design, production boosts both familiarity and recollection, but in a between-subjects design memory is benefited by familiarity alone. This suggests that in a within-subject design the encoding of distinctive information causes the memory boost but in the between-subjects design memory improvement is driven by encoding strength. In the following section I will expand on the notion that attention or other processes could play a more central role in the production effect and discuss the pupillometric findings within a broader range of cognitive theory.

4.3 Evidence Supporting Alternative Accounts of Production

In Experiment 1, I observed a divergence in pupil size as a response to the experimental condition prior to the presentation of any distinctive information. I previously examined these findings within the framework of attention facilitating the process of distinctive encoding. Now I would like to explore the idea that attention could play a more central role in the production effect. An alternate interpretation of the role of attention in the production effect is that attention mediates the effect. A mediation framework of attention in production would suggest that attention can facilitate distinctive processing as previously discussed, but also that attention can directly impact memory performance such that words that are better attended to are better encoded. This second component could explain why the pupillary response to the cue in the early window was related to the magnitude of the production effect. If this early pupillary response reflects differences in attention, then the amount of attentional engagement that was allocated during the pre-cue might predict the advantage that extra attention gave to encoding. Similarly, Fawcett and Ozubko (2016) posited that Hassall et al.'s (2016) observation of the P300 signal prior to production could be reflecting attentional resources allocated to facilitate production.

The increase in pupil size observed during the aloud and check conditions prior to production in both Experiment 1 and 2 might reflect a similar process. Whether attention directly impacts later memory or if it merely facilitates distinctive processing, my observation of pupil differences in the pre-cue in Experiment 1 suggests that attentional processes play a role.

Another possible indicator of attention in the current studies was the consistent finding of a smaller pupil size during the silent condition which decreased in size over the course of the trial. This pattern can be interpreted two ways. First, encoding the silent stimulus may not be as effortful as encoding the aloud or check stimulus; participants might stay focused on the word but use less effort to rehearse it. Second, participants may not attend to silent words as well and become disengaged with the task, perceiving it to be complete after reading the word silently. The fact that silent pupil size dropped below baseline by the end of the trial in both Experiment 1 and 2 (see Figures 2.3 and 3.3) suggests that participants were not actively focusing on the word throughout the entire trial. If attention was sustained, I would expect that maintaining focus over the course of the trial would be somewhat effortful (see Unsworth & Nash, 2008). That participants do not attend to silent items as well is compatible with a lazy reading hypothesis (MacLeod et al., 2010). The results are in agreement with previous research suggesting that participants do not attend to silent items as effectively as they do to aloud ones, based on subjective reports from participants (Fawcett & Ozubko, 2016; Ozubko et al., 2020) and evidence that reading passages aloud reduces mind-wandering (Varao Sousa et al., 2016). However, it should be noted that MacLeod et al. (2010) provided evidence challenging the lazy-reading hypothesis, demonstrating that the production effect held even when silent words were strengthened by generation. On that note, Fawcett (2013), proposed a “lively reading” hypothesis: that participants might not be less engaged with or motivated to encode the silent items, but rather that they were attending to the aloud items more, resulting in a benefit to those items. The results from the current experiments are also compatible with this viewpoint as I observed consistently a larger pupil diameter in the aloud condition compared to the silent condition.

Further evidence that attention might contribute to the production effect comes from studies where attention was directly manipulated within the production paradigm. First, Mama et al. (2018) demonstrated that the production effect was eliminated in a noisy environment when linguistic or fluctuating background noise was present. They argued that the elimination of the production effect was due to the noise interrupting attentional resources during the task. Interestingly, noise impacted memory for the aloud words only – memory for silent words was similar across conditions. One might argue that noise interrupted distinctive processing and not attentional resources by reducing the capacity to encode auditory feedback. However, across all noise conditions participants did not receive auditory feedback including a steady-state noise condition which resulted in a production effect comparable to that in the no noise condition. This indicated that it was not just the absence of auditory feedback that eliminated the effect with linguistic or fluctuating noise. Later Mama and Icht (2018a) again addressed the role of attention in the production effect by comparing a group of participants who were diagnosed with ADHD before and after taking MPH. Similar to Mama et al. (2018), improving attention (through administration of MPH) only improved memory for the produced words; silent words were equivalently recalled. This cohesion of results provides evidence that attention is fundamental in production as only those words that had been produced were impacted when attention was reduced.

Similarly, Slaney (2015) showed that divided attention (by having participants complete a secondary task where they responded to differing tones) eliminated the production effect for recall when the distraction occurred during the test phase. Unlike Mama et al. (2018), the author did not find an impact of divided attention during the study phase and suggested that distinctive information was automatically encoded even when attention was divided. However, these differences may have arisen due to methodological differences; in Slaney's study the tones appeared across differing sub-trials, whereas Mama et al. had a blanket noise condition that persisted through the entire study phase. Although arriving at different conclusions about the role of attention in production, studies that have manipulated attention suggest that attention does play an important role in the production effect and that the effect is reduced if not eliminated when attention is decreased (Mama et al., 2018; Mama & Icht, 2018a; Slaney, 2015).

As discussed earlier, pupil size during check conditions changed mostly to the same degree as the aloud conditions across both experiments but with no memory benefit to check words, which I determined was consistent with a distinctiveness account. However, I also mentioned that an alternative perspective is that equivalent mental effort is used on both trials, but in the aloud trials this effort is related to encoding the word while in the check trials it is in response to a different task demand. For example, it is possible that when participants had to say "check" aloud that they were primarily focused on the vocal response and that memorizing the stimulus became a secondary task. Previously, Tyler, et al. (1979) had participants complete a secondary task (e.g., responding to a tone) while they were engaged with a primary task, which involved choosing the correct word for an anagram or to complete a sentence. The primary task had a low effort and a high effort (e.g., more difficult anagrams or sentence completion) condition, and participants were told that this task was more important than performance on the secondary task (responding to the tone). They found that effort impacted performance on the secondary task but only in the high effort condition; when participants were using more cognitive effort on the primary task, response times for the secondary task were longer. This supports the assumptions made in the present study; if participants are using more cognitive effort (as evidenced by pupil dilation) but not showing an improvement for memory for those words, then participants could be perceiving memorizing the stimulus as a secondary task and thus performance is being impacted by the primary task – saying "check" aloud. When making a repetitive verbal response participants have to engage in two different processes – encoding the stimulus and saying an unrelated word aloud – this process may be effortful in its own right. Participants are required to hold two phonological representations (i.e., "check" and the silently read word) in their mind which might result in the word not being encoded as well because of the cognitive load of having to perform two unrelated tasks during encoding (Baddeley et al., 1984).

A second possibility is that the increase in pupil size before speaking is caused by performance anxiety. Evidence that participants might feel performance anxiety when speaking aloud during the production task was put forth by Forrin et al. (2019). These findings are especially relevant to the current study because Forrin et al. also alerted participants to the upcoming trial instruction and determined that memory performance differed depending on the upcoming condition. A connection between anxiety and cognitive effort has been previously measured using pupillometry; Hepsomali, Hadwin, Liversedge, Degno and Garner (2018) showed that participants who have higher trait anxiety might compensate by allocating more cognitive effort towards highly demanding cognitive tasks. When participants with high trait anxiety completed a cognitively demanding Go/No-Go task they showed greater pupil dilation than a low-trait anxiety group, indicating more cognitive effort being used. However, performance measures did not differ between the two groups, suggesting that cognitive effort was being used as an effective strategy to compensate for task related deficits due to anxiety. Although this study was making a comparison between participants with low and high trait anxiety, the results indicate that when participants are anxious they might compensate with mental effort. If Forrin et al.'s speculation about performance anxiety during production is correct, then it could be the case that in the speaking conditions some of the pupil change is reflecting mental effort being allocated in response to performance anxiety about the upcoming task. However, the correlation of the increase in pupil size during aloud trials compared to silent ones with the behavioral production effect suggests that pupil size is reflecting task-specific mental effort – at least during aloud and silent trials. Nevertheless, it is possible that performance anxiety partially contributes to the increase in pupil size observed.

4.5 Future Directions

Presently, in my lab we have obtained preliminary data extending the findings of Thoms et al. (2020) who studied the delayed production effect using varying stimulus onset asynchrony timings. Specifically, Thoms (2020) modified the previous study by applying pupillometry to measure cognitive effort during the delayed production effect across time intervals of 500 ms and 1500 ms. They also included catch trials – where the Go signal to produce the word never appeared, thus participants thought that they were going to be producing the word but never did; these catch trials also occurred less frequently than the other two conditions so that participants would be less likely to adapt to them. Previously, Thoms et al. (2020) observed increasing recognition accuracy for aloud and silent words as the trials lengthened, suggesting that a longer trial increased encoding strength. They also made the novel observation of a production effect during the catch trial when no word was produced, similar to the offline production effect but with the key difference being that participants were not told to imagine saying the words (Jamieson & Spear, 2014). By applying pupillometry to the previous behavioural delayed production design we will be able to make further inferences about whether cognitive effort is sustained over longer trials and if this does indeed relate to better memory for these words. My own observations in Experiment 2 suggest that pupil size will be sustained over longer trials since both the aloud and check conditions showed sustained pupil size when a delayed production procedure was used.

A limitation present in both the present Experiment 2 and Thoms (2020) is that neither of these studies included an immediate production condition. In previous investigations of the delayed production effect Mama and Icht (2018b) included an immediate production condition in which participants read the word aloud while it was on screen. Including an immediate production condition would allow for a direct comparison of memory improvement across the course of the trial while providing the additional measure of pupillometry. A future study could modify the delayed production procedures used in the present Experiment 2 and by Thoms (2020) so that there is also an immediate production condition. This would provide further corroboration for the theory that the delayed production condition resulted in better memory performance because the words were strengthened over the delayed period; this should be reflected in sustained cognitive effort over the course of the delay.

The present study demonstrates that pupillometry is an effective tool for measuring the cognitive processes that underlie the production effect. Certainly, my most challenging finding was the divergence of pupil diameter being different in all three conditions during the pre-cue in Experiment 1. I suggested that this was caused by an attentional component causing a greater pupil increase in the two speaking conditions, reflecting the effort it took participants to focus or upregulate their attention to prepare for speaking aloud. Future studies should find a way to adjust the current paradigm to determine if there is an attentional preparatory component to the production effect in a way that is more similar to the classic paradigm (i.e., simultaneous presentation of the instruction cue, stimulus, and production). Notably, by using a pre-cue procedure I fundamentally changed the experimental procedure from the classic paradigm and therefore, I can only infer that these processes would also occur without modifications. Demonstrating preparatory differences without using a pre-cue could provide further evidence for this process in the production paradigm, for example, participants might still upregulate attentional processes prior to speaking. Alternatively, I previously mentioned that another explanation for the increase in pupil size prior to production might be that it is an indicator of performance anxiety. Forrin et al., (2019) previously suggested that memory for silent words was worsened by performance anxiety elicited from participants being aware that they would be speaking aloud on an upcoming trial. A second physiological measure of anxiety, such as skin conductance could be applied to the present pre-cue paradigm to assess whether participants do indeed get anxious when they are about to speak aloud (Weerts, & Lang, 1978). This manipulation might also inform as to whether pupillometry can be used as a valid measure to assess performance anxiety.

4.6 Final Conclusions

Reading words aloud renders them more memorable than words that are read silently. There is substantial evidence that produced words are perceived as distinctive from unproduced words and that subsequent distinctive processing and related processes (i.e., use of a distinctiveness heuristic) are the main drivers of the production effect (Dodson & Schacter, 2001; MacLeod et al., 2010). Nevertheless, some studies have provided evidence that production also works by enhancing attentional processes (Forrin et al., 2019; Mama et al. 2018; Mama & Icht, 2018a) or by strengthening the encoding trace (Mama & Icht, 2018b). Additionally, attention has been cited as a mechanism that might inhibit the encoding of silent words (i.e., a "lazy reading hypothesis"; Fawcett & Ozubko, 2016; MacLeod et al. 2010; Ozubko et al., 2020).

The current experiments provided further corroboration of distinctiveness as a central process in the production effect. However, my results provided evidence that challenge a pure distinctiveness account; in particular, the finding of a pupillary response prior to the presentation of distinctive information supports earlier findings that attention might play a role in the production effect. Future studies should aim to further examine whether attention plays a critical or supportive role in improving memory via production. Further evidence for attention in the production effect was evidenced by smaller pupil changes during silent trials, consistent with the "lazy reading account" of production, which suggests the production effect may be in part due to a cost to silent items (MacLeod et al., 2010). Finally, correlations between the difference in cognitive effort measured on aloud trials compared to silent ones and the resulting production effect in recognition suggest that alternative processes, such as attention and cognitive effort are related to later memory. Furthermore, the current studies highlighted pupillometry as a useful metric for examining the processes underlying the production effect. Future research should continue to use pupillometry to further examine the cognitive underpinnings of the production effect. Given its simple implementation and effectiveness at improving memory it is important to understand the production effect as its applications in areas, such as academia and the workplace are enumerable. Only through understanding the underlying cognitive processes can the paradigm be applied in a truly effective manner.

References

Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 4*, 518-540. doi: 10.1037/0096-3445.113.4.518

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experiment Psychology: Learning, Memory and Cognition, 13*, 553-563. doi: 10.1037/0278-7393.13.4.553

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*, 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. T., Bernstein, D.M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*, 93-98. doi: 10.1037/cep0000082

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory and Cognition 38*, 1711-1719. doi: 10.1037/a0028466

Bodner, G. E., Taikh, A., & Fawcett, J.M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin and Review*, *21*,149-154. doi:10.3758/s13423-013-0485-1

Brenner, M. (1973). The next-in-line effect. *Journal of Verbal Learning and Verbal Behaviour*, *12*, 320-323. doi: 10.1016/S0022-5371(73)80076-3

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association, 68,* 66-74. doi:10.2307/2284140

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behaviour, 11*, 671-684. doi: 10.1016/s0022-5371(72)80001-x

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 33*, 497-505. doi: 10.1080/14640748108400805

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*, 341-361. doi: 10.1016/0749-596X(87)90118-5

Delacre, M., & Klein, O. (2019). Routliers: robust outliers detection. R package version 0.0.0.3. https://CRAN.R-project.org/package=Routliers

Dodson, C. S., & Y Schacter, D. L. (2001). "If I had said it I would have remembered it: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin and Review*, *8,* 155-161. doi: 10.3758/bf03196152

Experiment Builder (Version 2.2.299) [Computer software]. Mississauga, Ontario, Canada: SR Research

Eyelink 1000 [Apparatus and software]. (2010). Mississauga, Ontario, Canada: SR Research

Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paideía*, *23*, doi: 10.1590/1982-43272356201314

Fabiani, M., & Donchin, E. (1995). Encoding processes and memory organization: A model of the von Restorff effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 224-240. doi: 10.1037/0278-7393.21.1.224

Faraone, S. V., Bierderman, J., Spencer, T., Wilens, T., Seidman, L. J., Mick, E., & Doyle, A. E. (2000). Attention-deficit/hyperactivity disorder in adults: an overview. *Biological Psychiatry, 48*, 9-20. doi: 10.1016/S0006-3223(00)00889-1

Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology, 6*, 363-379. doi: 10.1016/j.stamet.2008.12.005

Fawcett, J. M. (2013). The production effect benefits performance in between-subjects design: A meta-analysis. *Acta Psychologica*, 142, 1-5. doi:10.1016/j.actpsy.2012.10.001

Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory, 20*, 655-666. doi:10.1080/09658211

Fawcett, J. M., Bodner, G. E., Bailey, L., Stewart, B., Roddick, K., O'Neil, K.,… & Newman, A. (2020). The neural correlates of the production effect: An fMRI study. Manuscript submitted for publication.

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 70*, 99-115. doi: 10.1037/cep0000089

Fuermaier, A. B. M., Tucha, L., Koerts, J., Aschenbrenner, S., Kaunzsinger, I., Hauser, J., …& Tucha, O. (2015). Cognitive impairment in adult ADHD--perspective matters! *Neuropsychology, 29*, 45-58. doi: 10.1037/neu0000108

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J.W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology, 66*, 2289-2294. doi: 10.1080/17470218.2013.858170

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborate vs non-elaborative processing. *Memory, 22*, 470-480. doi: 10.1080/09658211.2013.798417

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further for the item-order account. *Canadian Journal of Experimental Psychology/Revue Canadienne de psychologie expérimentale, 70*, 125-138. doi: 10.1037/cep0000088

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*, 1046-1055. doi: 10.3758/s1321-012-0210-8

Forrin, N. D., Ralph, B. C. W., Dhaliwal, N. K., Smilek, D., & MacLeod, C. M. (2019). Wait for it… performance anticipation reduces recognition memory. *Journal of Memory and Language, 109*, 1-13. doi:10.1016/j.jml.2019.104050

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavioural Research Methods,* 43, 1171-1181. doi: 10.3758/s13428-011-0109-5

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition, 16*, 110-119. doi: 10.3758/BF03213478

Geller, J., Still, M. L., & Morris, A. L. (2016). Eyes wide open: Pupil size as a proxy for inhibition in the masked-priming paradigm. *Memory & Cognition, 44*, 554-654. doi: 10.3758/s13421-015-0577-4

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 15*, 870-878. doi: 10.1006/nimg.2001.1037

Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Occulometric variations during mind wandering. *Frontiers in Psychology, 5*, 1-10. doi: 10.3389/fpsyg.2014.00031

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*, 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x

Hassall, C. D., Quinlan, C. K., Turk, D. J., Taylor, T. L., & Krigolson, O. E. (2016). A preliminary investigation into the neural basis of the production effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 70*, 139-146. doi: 10.1037/cep0000093

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behaviour Research Methods, 50*, 107-114. doi: 10.3758/s13428-017-1008-1190

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem solving. *Science, 143*, 1190-1192. doi: 10.1126/science.143.3611.1190

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A qualitative system analysis. *Behaviour Research Methods, Instruments, & Computers, 25*, 16-26. doi: 10.3758/bf03204445

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behaviour, 11*, 534-537. doi: 10.1016/S0022-5371(72)80036-7

Hourihan, K. L., & Churchill, L. A. (2019). Production of picture names improves picture recognition. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale,74,* 35-43. doi: 10.1037/cep0000185

Hourihan, K. L., & MacLeod, C. M. (2008). Directed forgetting meets the production effect: Distinctive processing is resistant to intentional forgetting. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 62*, 242-246. doi: 10.1037/1196-1961.62.4.242

Hubert, M., & Debruyne, M. (2009). Minimum covariance determinant. *WWEREs Computational Statistics, 2*, 36-43. doi: 10.1002/wics.61

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory*. (pp. 3-25). Oxford: Oxford University Press

Jamieson, R. K., Mewhort, D. J. K., & Hockley, (2016). A computational account of the production effect: Still playing Twenty-Questions with nature. *Canadian Journal of Psychological Science, 70*, 154-164. doi: 10.1037/cep0000081

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 68*, 20-28. doi: 10.1037/cep0000009

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 40*, 300-305. doi:10.1037/a0033337

Jonker, T. R., Levene, M., & MacLeod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 441-448. doi: 10.1037/a0034977

Kamp, S. -M., Forester, G. R., Murphy, A. R., Brumback, T., & Donchin, E. (2012). Testing a distinctiveness explanation of the primacy effect in free recall using event-related potentials. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34*, 539-544. Retrieved from https://escholarship.org/uc/item/7rk363fw

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, 154, 1583-1585. doi: 10.1126/science.154.3756.1583

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology, 38*, 557-577. doi: 10.1017/S0048577201990559

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). The production effect in long-list recall: In no particular order? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 70*, 165-176. doi: 10.1037/cep0000086

Lawrence, M. A. (2016). ez: Easy analysis and visualisation of factorial experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

Lin, O. Y. H., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale,66*, 212-216. doi: 10.1037/a0028309

Macdonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica, 98*, 291-310. doi: 10.1016/S001-6918(97)00047-4

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*, 390-395. doi: 10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology, 33*, 671-685. doi: 10.1037/a0018785

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods, 50*, 94-106. doi: 10.3758/s13428-017-1007-2

Mama, Y., Fostick, L., & Icht, M. (2018). The impact of different background noises on the production effect. *Acta Psychologica, 185,* 235-242. doi: 10.1016/j.actpsy.2018.03.002

Mama, Y., & Icht, M. (2018a). Production effect in adults with ADHD with and without Methylphenidate (MPH): Vocalization improved verbal learning. *Journal of the International Neuropsychological Society, 00*, 1-6. doi: 10.1017/S1355617718001017

Mama, Y., & Icht, M. (2018b). Production on hold: Delaying vocal production enhances the production effect in free recall. *Memory, 26,* 589-602. doi: 10.1080/09658211.2017.1384496

Meiran, N., & Chorev, Z. (2005). Residual task-switching cost. *Experimental Psychology, 52*, 109-124. doi: 10.1027/1618-3169.52.2.109

Morehead, K., Rhodes, M. G., & Delozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*, 257-271. doi: 10.1080/09658211.2014.1001992

Ozubko, J. D., Bamburoski, L. D., Carlin, K., & Fawcett, J. M. (2020) Distinctive encodings and the production effect: Failure to retrieve distinctive encodings decreases recollection of silent items. *Memory, 28*, 237-260. doi: 10.1080/0965811.2019.1711128

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory and Cognition*, *40*, 326-338. doi: 10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory*, *20*, 717-727. doi: 10.1080/09658211.2012.699070

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect, *Memory, 22,* 509-524. doi:10.1080/09658211.2013.800554

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. Journal of Experimental Psychology, 36, 1543-1547. doi: 10.1037/z0020604

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H.,… & Lindeløv, J. (2019). PsychoPy2: Experiments in behaviour made easy. Behaviour Research Methods. 10.3758/s13428-018-011-y

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118,* 2128-2148. doi: 10.1016/j.clinph.2007.04.019

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory, 21,* 904-915. doi: 10.1080/09658211.2013.766754

Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology/ Revue canadienne de psychologie expérimentale, 73*, 254-264. doi: 10.1037/cep0000179

R Core Team (2018). R: A language and environment for statistical computing (Version 3.5.1) [Computer Software]. Retrieved from https://www.R-projects.org/

Roddick, K., Fawcett, J. M., Newman, A., Lambert, A. M., & Bodner, G. E. (2014). [Pilot experiment exploring the neural basis of production]. Unpublished raw data.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *American Psychological Association, 7*, 504-523. doi: 10.1037/1082-989X.7.4.504

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., & Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS ONE, 6*, 1-8. doi: 10.1371/journal.pone.0018298

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science, 21*, 786-789. doi: 10.1177/09567976

SR Research Ltd. (2010). *EyeLink 1000® Installation Guide*. Mississauga, Ontario, Canada.

Slaney. B. (2015). *The production effect: The role of attentional resources* (Master's thesis). Retrieved from https://research.library.mun.ca/11661/1/thesis.pdf

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology/ Revue canadienne de psychologie expérimentale, 70*, 186-194. doi: 10.1037/cep0000083

Thoms, C. (2020). *The delayed production effect: Using pupillometry to explore the mechanisms of production* (Unpublished honours thesis). Memorial University of Newfoundland, St. John's, Canada.

Thoms, C., Fawcett, J. M., Hourihan, K. L., & Willoughby, H. V. (2020). *The Contributions of Response Preparation to the Production Effect in a Delayed Production*. Manuscript in preparation.

Tiller, J. C., Hourihan, K. L., & Fawcett, J. M. (2018). Production is in the eye of the beholder: A pupillometric investigation of the production effect. Poster presented at: Canadian Society of Brain, Behaviour and Cognitive Science; 2018, July 4-7; St. John's, NL.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 607-617. doi: 10.1037/0278-7393.5.6.607

Unsworth, N., & Robison, M. K. (2018). Tracking arousal state in mind wandering with pupillometry. *Cognitive, Affective & Behavioural Neuroscience, 18*, 638-664. doi: 10.3758/s13415-018-0594-4

Weerts, T. C., & Lang, P. J. (1978). Psychophysiology of fear imagery: Differences between focal phobia and social performance anxiety. *Journal of Consulting and Clinical Psychology, 46*, 1157-1159. doi: 10.1037/0022-006X.45.5.1157

Winn, M., Indt, D., Koelewijn, T., & Kuchinsky, S. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing, 22*, 1-32. doi: 10.1177/2331216518800869 van Rij, J., Henderiks, P., van Rijn, H., Baayen, R.H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Pupillometry in Hearing Science: Original Article, 23,* 1-22. doi: 10.1177/2331216519832483

Varao Sousa, T. L., Carriere, J. S. A., & Smilek, D. (2013). The way I encounter reading material influences how frequently I mind wander. *Frontiers in Psychology, 4*, 1-8. doi: 10.3389/fpsyg.2013.00892

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology, 52*, 23-36. doi: 10.1016/j.ijpsycho.2003.12.003

Yu, K., Prasad, I., Mir, H., Thakor, N. & Al-Nashash, H. (2015). Cognitive workload modulation through degraded visual stimuli: A single-trial EEG study. *Journal of Neural Engineering, 12,* doi: 10.1088/1741-2560/12/4/046020

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517. doi: 10.1006/jmla.2002.2864

Appendix

MEMORIAL UNIVERSITY

Interdisciplinary Committee on
Ethics in Human Research (ICEHR)

St. John's, NL Canada A1C 5S7
Tel: 709 864-2561 icehr@mun.ca
www.mun.ca/research/ethics/humans/icehr

ICEHR Number:	**20200706-SC**
Approval Period:	August 27, 2019 – August 31, 2020
Funding Source:	NSERC [RGCS# 20131132; Hourihan] [RGCS# 20171303; Fawcett]
Responsible Faculty:	Dr. Jonathan Fawcett Psychology
Title of Project:	*Attention and Memory for Words*

Title of Parent Project:	*Attention and Human Memory*
ICEHR Number:	*20171484-SC*

August 27, 2019

Hannah Willoughby
Department of Psychology, Faculty of Science
Memorial University of Newfoundland

Dear Hannah Willoughby:

Thank you for your submission to the Interdisciplinary Committee on Ethics in Human Research (ICEHR) seeking ethical clearance for the above-named research project. The Committee has reviewed the proposal and agrees that the proposed project is consistent with the guidelines of the *Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans* (TCPS2). *Full ethics clearance* is *granted to* August 31, 2020. However, the recruitment documents are missing. Please complete the **ICEHR - Post-Approval Document Submission form** and upload these documents. ICEHR approval applies to the ethical acceptability of the research, as per Article 6.3 of the *TCPS2*. Researchers are responsible for adherence to any other relevant University policies and/or funded or non-funded agreements that may be associated with the project.

The *TCPS2* **requires** that you submit an Annual Update to ICEHR before August 31, 2020. If you plan to continue the project, you need to request renewal of your ethics clearance and include a brief summary on the progress of your research. When the project no longer involves contact with human participants, is completed and/or terminated, you are **required** to provide an annual update with a brief final summary and your file will be closed. If you need to make changes during the project which may raise ethical concerns, you must submit an Amendment Request with a description of these changes for the Committee's consideration prior to implementation. If funding is obtained subsequent to approval, you must submit a Funding and/or Partner Change Request to ICEHR before this clearance can be linked to your award.

All post-approval event forms noted above can be submitted from your Researcher Portal account by clicking the ***Applications: Post-Review*** link on your Portal homepage. We wish you success with your research.