Elaborative processing and recognition memory

by

Jackie Spear

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

University of Manitoba

Winnipeg, Manitoba

Abstract

In this dissertation, I present a comprehensive account of elaborative processing grounded in instance theory. My aim is to demonstrate that a variety of memory effects can be traced back to the same underlying mechanism: elaborative processing. The central focus of this research is the production effect, though the approach is also extended to two other well-established phenomena—the generation effect and the testing effect. In addition, I present a parallel investigation involving the production effect and directed forgetting. To achieve my aim, I conducted both experimental and computational analyses of these effects within the context of recognition. Building on the MINERVA 2 model, I introduced novel extensions to apply it to the production, generation, and testing effects. The overarching goal of this thesis is to demonstrate that these memory effects can be explained by two principles of memory, strength and distinctiveness, advancing our understanding of elaborative processing. This work will contribute to a deeper insight into the production effect and will enhance existing models of memory processing.

Keywords: production effect, generation effect, testing effect, recognition, MINERVA 2

Acknowledgements

First and foremost, I'd like to thank Randy Jamieson for his mentorship, support, and for challenging me to be a better researcher. You've helped shape the way that I think about thinking, and I'm grateful for all that you've taught me and the opportunities you've provided. I'd also like to thank my committee, Dr. Murray Singer, Dr. Jason Leboe-McGowan, and Dr. Sunmee Kim. Thank you for all your feedback and guidance. Thank you to my colleagues and friends, Dr. Chrissy Chubala, Dr. Matt Cook, Dr. Evan Curtis, Dr. Dominic Guitard, Dr. Nick Reid, Angus Ball, and Sarah McAmmond. Your support, collaboration, and friendship has made this journey all that much sweeter. Lastly, I'd like to thank my amazingly supportive family and of course, Jared Adams. You've all given me so much gratitude. Thank you.

Dedication

This dissertation is dedicated to both of my grandmothers, Nellie Trochim (1925-2008) and Rae Spear (1931-2019). Thank you for always pushing me to be the best version of myself and for teaching me what it means to be compassionate and resilient. Thank you for everything. You have given me a lifetime's worth of gratitude.

"*I was dropped from the moonbeams, and sailed on shooting stars*" - Sail to the Moon, Radiohead

Abstract ii
Acknowledgements iii
Dedication iv
List of Tables vii
List of Figures viii
Chapter 1: Elaborative Processing 12
Study Instruction Manipulations of Elaborative Processing 14
The Production Effect 16
The Generation Effect 16
The Testing Effect 17
Levels of Processing 17
Why Modeling? 20
Modeling the production effect 22
Structure of the Dissertation 22
Chapter 2: The Production Effect 24
Experiment 1: The Production Effect 28
Results and Discussion 32
Chapter 3: The Generation Effect 36
Experiment 2 39
Results and discussion 40
Chapter 4: The Testing Effect 45
Experiment 3 48
Chapter 5: A model of elaborative processing 55
Simulation of Experiment 1 60
Simulation of Experiment 1 – Strength and Distinctiveness 61
Simulation of Experiment 1 – Strength Only 63
Simulation of Experiment 2 – Strength and Distinctiveness 64
Simulation of Experiment 2 – Strength Only 66
Simulation of the testing effect 67
Simulation of Experiment 3 70

Chapter 6: The production effect and directed forgetting: Assessing strength and distinctiveness 73
Chapter 7: General Discussion 128
A Principled Approach 130
Uniting Siloed Approaches 130
Levels of Processing 133
Features – what do they represent? 134
Implications 138
Conclusion 140
References 142
Contributions of Authors 159
Appendix A 160
Appendix B 172

List of Tables

Table 2.1. Results from Experiment 1. Proportion "old" as a function of condition. 32

Table 3.1. Results from Experiment 2. Proportion "old" as a function of condition. 41

Table 4.1. Results from Experiment 3. Proportion "old" as a function of condition. 52

List of Figures

Figure 1.1. Experimental results from Jacoby (1983), reproduced in Brooks (1987). 19

Figure 4.1. A depiction of how participants were assigned to one of three conditions after the study phase, before they proceeded to the recognition test phase. In the mixed-list and pure-test conditions, participants studied blocks of four words and then were tested on two words, or four words, respectively. This cycle continued ten times until all forty words were studied, where participants were then given a standard yes/no recognition test. In the pure-read condition, participants studied all forty words and went straight to the recognition test phase. 51

Figure 5.1. An example of how distinctiveness is represented in the model. Produced items (first three rows) are assigned additional unique features (last two columns), whereas unproduced items (last three rows) receive no extra features. 57

Figure 5.2. Simulation results of the production effect in Experiment 1, integrating distinctiveness and strength. Mixed-list parameters: $L = 0.046$ for produced targets, $L = 0.040$ for read targets; the number of extra production features for produced targets was 250. The decision criterion was set to an unbiased value of 0.501. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 61

Figure 5.3. Simulation results of the production effect in Experiment 1, using only a strength mechanism. Parameters: $L = 0.0353$ for produced targets and $L = 0.0279$ for read targets.

Distinctiveness was not assumed; no additional features were added. The decision criterion was an unbiased value of 0.505. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 63

Figure 5.4. Simulation results of the generation effect in Experiment 2, integrating distinctiveness and strength. Mixed-list parameters: $L = 0.056$ for produced targets, $L = 0.040$ for read targets, and the number of extra generation features for produced targets was 250. The decision criterion was set to an unbiased value of 0.501. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 64

Figure 5.5. Simulation results of the generation effect in Experiment 2, using only a strength mechanism. Parameters: $L = 0.0500$ for generated targets and $L = 0.0254$ for read targets. Distinctiveness was not assumed so no additional features were added. The decision criterion was an unbiased value of 0.505. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 66

Figure 5.6. Simulation results of the testing effect in Experiment 3, assuming that the retrieved echo content with associated probe is encoded into memory after a 2nd iteration for all tested items. Parameters: $L = 0.0170$ for tested targets and $L = 0.0175$ for read targets. The decision criterion was an unbiased value of 0.48. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 70

Figure 6.1. Results from Experiment 1 in Spear et al. (2025). Error bars represent standard errors of the means. 86

Figure 6.2. Results from Experiment 2 in Spear et al. (2025). Error bars represent standard errors of the means. 90

Figure 6.3. An example of the two study procedures. The "after" study procedure was used in Experiments 1 and 2 in Spear et al. (2025), and the "during" study procedure was used in Experiment 3. The duration of stimulus presentation is displayed on the right whereas ISI is displayed on the left. 93

Figure 6.4. Results from Experiment 3 in Spear et al. (2025). Error bars represent standard errors of the means. 94

Figure 6.5. Simulation results of the production effect in Experiment 1 in Spear et al. (2025) integrating strength and distinctiveness. Mixed-list parameters: $L = 0.057$ for produced targets, $L = 0.057$ for read targets, and the number of extra production features for produced targets was 250. The decision criterion was set to a slightly conservative value of 0.4625. Error bars represent standard deviations. Dotted lines represent corresponding empirical means. 102

Figure 6.6. Simulation results of the production effect data in Experiment 1 in Spear et al. (2025) assuming only strength. Mixed-list parameters: $L = 0.0425$ for produced targets and $L = 0.0375$

for read targets. There were no extra production features in the second simulation, as we did not assume any contributions of distinctiveness. The decision criterion was set to a slightly conservative value of 0.4625. Error bars represent standard deviations. 104

Figure 6.7. Simulation results of the directed forgetting data in Experiment 2 in Spear et al. (2025). Parameters: $L = 0.050$ for R-cued targets and $L = 0.026$ for F-cued targets, and the number of extra elaborative features was 250. The decision criterion was set to a conservative value of 0.3875. Error bars represent standard deviations, and dotted lines represent corresponding empirical means. 106

Figure 6.8. Simulation results of the directed forgetting data in Experiment 2 in Spear et al. (2025). Parameters: $L = 0.0425$ for R-cued targets and $L = 0.0226$ for F-cued targets. The decision criterion was set to a conservative value of 0.3875. Error bars represent standard deviations, and dotted lines represent corresponding empirical means. 107

Figure 6.9. Simulation results of the mixed-list production effect data in Experiment 3 in Spear et al. (2025). Parameters: $L = 0.0425$ for produced targets and $L = 0.0375$ for read targets. The decision criterion was set to a slightly conservative value of 0.4625, as in Experiment 1. Error bars represent standard deviations, and dotted lines represent corresponding empirical means. 109

Chapter 1: Elaborative Processing

The quest to find ways to improve memory dates back to Francis Bacon in the 1600s and William James in the 1800s (Bacon, 2000; James, 1890; Roediger & Karpicke, 2006). Critically, James wrote that "things are impressed better by active than by passive repetition" (p. 646) (Kornell & Terrace, 2007). Since then, how to improve memory has been a driving force behind a range of research in the domain of cognitive psychology. However, at the same time, a number of experimental demonstrations of the same underlying mechanisms responsible for a benefit conferred to memory have been conducted. In this way, our discipline has lost sight of the forest for the trees, chasing effects and experimental demonstrations at the expense of developing investigative theories of memory and cognition.

A pioneer of the cognitive revolution, Herb Simon, started sounding the alarm regarding our science and the divisive nature of our discipline which has been divided into subdomains and different "schools of thought" (Simon, 1969; 1992). Even within the subdomains of our discipline, the divisive nature of carving processes up continues, with different systems of memory being one example. Largely, cognitive psychology has been organized as a series of effect-driven research, showing a large number of laboratory demonstrations of how to improve memory. For example, in a recent meta-analysis of the generation effect (McCurdy et al., 2020), the authors stated that the lack of a unified theory of the generation effect is likely due to different experimental procedures used to examine the phenomenon. This is surprising, given the size of the literature on the issue.

Newell's 1973 paper "You can't play 20 questions with nature and win..." has been cited over 1,500 times, and over 140 times since 2023. Clearly our discipline and other related fields

of research are still in a state that warrants discussion of the issues raised by Newell over fifty years ago. The search for unifying principles that govern human behavior and cognition ought to be at the heart of what drives current research.

What has been coined "the theory crisis" has been postulated to be one of the sources of the replication crisis in psychology (Oberauer & Lewandowsky, 2019). For example, Oberauer and Lewandowsky (2019) have argued that poor replicability is often caused by the weak logical link between theories and their corresponding empirical tests. They argue that not only is it our methods that need improving, but our attention to theory as well. Ideally, a prediction that is drawn is derived deductively from theory. However, in practice, thinking of the dichotomy between discovery-oriented research and theory-testing research as more of a continuum, rather than as two mutually exclusive options, is more practical. The closer a research question is to theory-testing on this continuum, the greater the strength of the inferential link between theory and the hypothesis.

In the extreme case, it has been argued that psychology has gathered a sufficient weight of experimental evidence to obviate the need of laboratory study and that we ought to work instead toward a purely theoretical analysis of that existing database (Griffiths, 2015). A less extreme argument is that psychology ought to find common principles that govern human behavior (Surprenant & Neath, 2013). It is in the spirit of this assertion that this dissertation proceeds. Instead of finding novel cognitive techniques or different laboratory demonstrations of effects under different conditions (e.g., does the production effect still hold under increasingly microscopic or exotic manipulations?), experimental analysis ought to be driven by theory, and a major aim of theory ought to be to unite the numerous laboratory examples that fill our academic

journals and literature. One such principle is elaborative processing that has been a cornerstone in memory research for decades.

Study Instruction Manipulations of Elaborative Processing

In the last half century, many examples of elaborative processing have been investigated, all with subtle differences. One such demonstration is the enactment effect. The enactment effect is the memorial benefit conferred when an action verb is paired with performing an action in reference to it rather than simply reading it (e.g. pick up the pen; Engelkamp & Krumnacker, 1980; Cohen, 1981; Engelkamp & Zimmer, 1989). Since the 1980's the effect has been investigated in the context of order reconstruction (Engelkamp & Dehn, 2000), implicit and explicit memory (Engelkamp et al., 1995), paired-associate learning (Engelkamp, 1995), and free recall (Engelkamp, 1995). For example, in Engelkamp et al. (1995), enactment was shown to benefit an explicit recognition test but not an implicit verb identification test. In Engelkamp (1995), enactment led to better free recall than standard learning instructions.

The drawing effect is another phenomenon where a memorial benefit is observed for words that are pictorially drawn compared to written (Fernandes et al., 2018). For example, in Wammes et al. (2016), drawing the referents of words resulted in better recall of words in comparison to conditions where participants (a) wrote those words, (b) wrote descriptions of those words, (c) visualized the words, or (d) looked at pictures of the words. The drawing effect has also been used to examine age (e.g., old versus young adults; Meade, et al., 2018), stimulus type (e.g., Fernandes, Wammes, & Meade, 2018), categorical perception (Fan et al., 2015), and learning (Wammes et al., 2017). For example, in Wammes et al. (2017), participants were instructed to either to rewrite or to draw pictures that represented to be remembered definitions

from university textbooks. Results indicated that drawing conferred a larger memorial benefit than rewriting.

Computational analyses have recently been applied to the drawing effect, where Roberts and Wammes, (2021) examined the drawing effect using concrete and abstract words. In those studies, results showed that although the effect was slightly larger for concrete than for abstract words, it was still observed for all words. They further employed deep convolutional neural networks to analyze the drawings that people produced at study, where results suggested that the abstract drawings with more distinctive features (unlike other subjects' drawings of that word) conferred a larger memorial benefit. Conversely, prototypicality benefited concrete stimuli, suggesting that there is a strong connection between drawing and visual representation.

Another effect of elaborative processing has been coined the bizarreness effect (McDaniel & Einstein, 1986; Einstein & McDaniel, 1987; see Worthen, 2006 for a review). In a typical bizarreness effect experiment, participants are instructed to read bizarre or common sentences that include target capitalized nouns. A benefit to memory is conferred for target nouns embedded in bizarre sentences. The bizarreness effect has been demonstrated with pictures (Marchal & Nicolas, 2000) and in the context of free recall (Worthen & Roark, 2002), cued recall (Worthen & Loveland, 2003), and recognition (Thomas & Loftus, 2002; Worthen & Eller, 2002). For example, Worthen and Eller (2002) investigated the bizarreness effect for common and bizarre hand-drawn pictures. When participants were tested after a two-week delay, bizarre pictures were recognized better than common pictures.

The Production Effect

The production effect refers to the memorial advantage observed when words are actively produced—such as by speaking, singing, or typing—compared to passively engaging with the words, such as reading silently. In a keystone paper by MacLeod et al., (2010), the phenomenon was dubbed the production effect. The effect has been robust across a wide range of different materials and modalities and has been widely studied (see MacLeod & Bodner, 2017 for a brief review). The production effect is the first effect of elaborative processing that will be examined in this dissertation; thus, a more detailed description of the production effect is reserved for Chapter 2.

The Generation Effect

Yet another effect of elaborative processing is the generation effect. Like the production effect, the generation effect enhances memory by requiring participants to actively produce information rather than passively read it. However, instead of vocalizing or writing a complete word, participants must generate a missing part of a word from a given stem. For example, when presented with the cue-target pair "HOT – C_ _ _", the participant would generate "COLD".

The generation effect has been widely studied in memory research, with findings consistently showing that self-generated words are remembered better than those that are merely provided (Slamecka & Graf, 1978). This advantage is thought to arise from deeper cognitive processing (McCurdy et al., 2020). Like the production effect, the generation effect has been observed across various study conditions and appears to enhance both recognition and recall. The generation effect will be the second effect of elaborative processing examined in this dissertation. A more detailed review will be provided in Chapter 3.

The Testing Effect

The last effect of elaborative processing that will be examined in this dissertation is the testing effect. A more in-depth review of the testing effect will be provided in Chapter 4, however in short, the testing effect refers to the benefit of testing over and above restudying material. In a typical testing effect procedure, participants are given an opportunity either to restudy the material or to engage in a retrieval practice task, such as taking a test on the learned information. Research has consistently shown that retrieval practice due to the test enhances long-term retention more effectively than passive restudying (Rowland, 2014).

In relation to the production effect and the generation effect, the testing effect involves the most elaborative effort on a continuum (e.g., typing the word given the full cue and target word, typing the word given the full cue word and only the first letter of the target word, and typing the word given only the cue and none of the target word). The testing effect also serves to bridge the gap between basic research and the applied domain.

Levels of Processing

All the effects just discussed show that if an individual wants to learn something, merely reading or restudying material is not going to offer the utmost benefit. Rather, it is the act of engaging with material that matters, whether that be in the form of drawing, enacting, producing, generating, or testing. Thus, all of these encoding manipulations fit into a common theoretical framework known as levels of processing. Classically, levels of processing manipulations vary how deeply material is encoded at study (Craik & Lockhart, 1972). Generally, attending to surface features such as sound or appearance results in what is called “shallow” processing, whereas attending to meaning of words results in what is called “deeper” processing.

Additionally, sensory processing is thought to be more automatic whereas semantic encoding is thought to require more effortful engagement. The result that follows from the levels of processing manipulation is that material that has been processed in a deeper fashion is better remembered than is shallowly encoded information.

In a review paper, Lockhart and Craik (1990) further specified that the levels of processing framework include two different types of processing: type I, which refers to maintenance and rehearsal, and type II, which refers to elaborative rehearsal. In the realm of the production effect, the generation effect, or the testing effect, it is apparent that each procedure pits these types of processing techniques against each another.

There are a number of different ways that the levels of processing framework can be investigated in experimental designs. Diagnosticity of the task can be altered such that comparisons or encoding of stimuli is designed to emphasize similarity (Tversky, 1977). That is, the way in which the task is designed can lead to more contrastive or similarity-based comparisons, which in turn affect mental organization (e.g. Italy and Switzerland become more similar to one another in a task that also compares the similarity of South American countries instead of only European countries; Jacoby, 1979). Additionally, a closely related principle of memory is transfer appropriate processing (Morris et al., 1977). Transfer appropriate processing highlights the relation (usually congruency/incongruency) between encoding and retrieval and predicts that memory performance will be optimal when the encoding operations match the retrieval conditions.

In a classic paper by Jacoby (1983), incidental learning was examined under three different conditions: a no-context condition, a context condition, and a generate condition. In the

no context condition, a word was preceded by a mask (e.g., XXXX - DARK). In the context condition, a word was preceded by an antonym (e.g., HOT - COLD). Lastly, in the generate condition, participants were instructed to generate a word from an antonym (e.g., LOW - ????). When encoding and retrieval conditions were analogous (e.g. no context encoding with perceptual identification at retrieval), performance was at its best. However, that same pattern was reversed when participants were tested for recognition rather than perceptual identification at test.

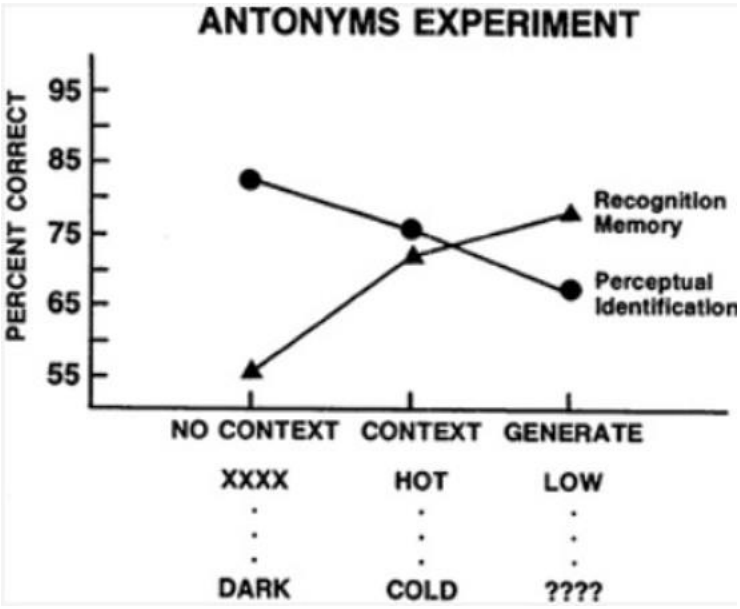


Figure 1.1. Experimental results from Jacoby (1983), reproduced in Brooks (1987).

Recently, Curtis (2019) used MINERVA to model the findings of Jacoby's classic demonstration. To model the different kinds of processing, additional assumptions about stimulus representation were made. In addition to the features of the memory trace that are typically used to represent an item, an extra set of features was added to each item to represent perceptual features or conceptual features of the stimulus. By assuming that a task can favor the activation of some of these features over others, MINERVA 2 captured this dissociation and

simulated the interaction between study and retrieval match/mismatch. That is, knowledge about the type of memory test, or the consideration of how memory is tested, can be useful for modeling memory performance, a condition that helps elucidate what might be encoded into memory. Based on that modeling exercise, Curtis managed to shed further light on a general principle of memory using a formal simulation of the 1983 Jacoby results. The representational assumptions adopted here are closely related to those used in Jamieson et al.'s (2016) model of the production effect and will be discussed in more detail later on in Chapter 5.

Although Jacoby's empirical demonstration and argument were sufficient to make the case that deep encoding versus shallow encoding cannot be objectively defined independent of a consequence of how memory is tested, Curtis'(2019) modeling exercise helps to bring that account under careful analytic scrutiny. More importantly for this thesis, it shows that elaborative processing may fall within the scope of modern computational accounts of memory and, if so, that this account might be able to help to better articulate a mechanistic and general account of how elaborative processing benefits memory.

Why Modeling?

Computational approaches to psychological processes are important for a number of reasons. Lewandowsky (1993) outlined several important aspects and benefits of computational modeling to the discipline of psychology. When the complexity of a problem challenges the bounds of human reasoning, models serve as a compensatory tool. They can take incomplete verbal theories and expose their weaknesses, demand precise specification, and be easy to replicate. Moreover, they can lead to unexpected discoveries, especially when they contradict expectations. Neath (1999) also argued that computational and mathematical models provide and

require precise specification of processes and assumptions, challenge human reasoning, and provide opportunities to discover novel findings. Farrell and Lewandowsky (2010) further argued that computational modeling should be a part of any psychologist's skillset because it forces one to fully specify all parts of a theory, avoids ambiguity when formulating theory, and aids in improved reasoning for complicated large-scale problems. Moreover, in the realm of computational neuroscience, model parameters can be linked to brain processes (Bowers, 2009). Most recently, van Rooij (2022) advocated for more models in our discipline, stating that the lack thereof hinders progress. Specifically, van Rooij argued that models ought to be included in the standard psychological curriculum and that there ought to be more theoretically driven research to accompany what she calls “effect driven naive empiricism”.

Of course, there have been criticisms of models, and when not designed properly they can be problematic or even misleading. One such problem is referred to as Bonini's paradox that arises when a model is designed to be so complex that it is impossible to identify the causal mechanism that produces a pattern of behavior (Dutton & Starbuck, 1971) and the temptation to choose a complex model over a simple one because of the belief that human beings are fundamentally complex (Loftus, 1985). Another issue that can arise is the inadvertent performance issue, which is when the output of a model does not necessarily follow from the modeler’s explicitly acknowledged assumptions (Neath, 1999; Dutton & Starbuck, 1971). However, it is clear that the benefits outweigh the costs, especially when researchers are mindful of the pitfalls of the modeling approach.

Importantly, as with experimental analysis, the aim of modeling should not explain dataset after dataset but rather should aim to explain processes that can be generalized beyond

single demonstrations (Estes, 1975). In the realm of machine learning, this is the issue of bias variance tradeoff (Geman, 1992). Ideally, a model should capture regularities and patterns in the data but also be able to generalize to new datasets. However, due to the inverse relation between bias and variance, it is impossible to do both well. Thus, the researcher must reach a productive compromise between the two that best meets the needs of the exploratory scope and target. Moreover, developing formal models and theories can help with issues associated with the replication crisis in our discipline regarding more rigorous theory testing rather than exploratory driven research (Oberauer & Lewandowsky, 2019).

Modeling the production effect

To date, there have been a few approaches to modeling the production effect, including the revised feature model (RFM; Saint-Aubin et al., 2021), REM (Kelly et al., 2022), attentional subsetting theory (AST; Caplan & Guitard, 2024) and MINERVA 2 (Jamieson et al., 2016; for more details on these models not including MINERVA 2, see Appendix A). However, in short, a commonality of all these models is the inclusion of additional features to represent the benefit of production. REM, AST, and MINERVA 2 are models of recognition, whereas the RFM is a model of recall. MINERVA 2 will be the model that is used in this dissertation to account for the results in Experiments 1-3. Additionally, it is the model that was used to account for the empirical results obtained in this dissertation and presented in Chapter 5. A more detailed description of the model will also be provided in Chapter 5.

Structure of the Dissertation

I will examine three effects of elaborative processing—the production effect, generation effect, and testing effect—over a series of three experiments. In addition, I will present

experiments that compare the production effect and directed forgetting in parallel investigations in Chapter 6. My goal will be to explain these effects using a computational model of human memory, MINERVA 2, that combines strength and distinctiveness mechanisms, my aim will be to show that, depending on the task at hand, the contribution of each of these mechanisms may vary. By doing so, I aim to investigate these effects of elaborative processing in parallel to show that common mechanisms can explain the benefit bestowed to memory in a principled framework. This work will aim to further delineate that the correlated concepts of strength and distinctiveness are not mutually exclusive and may work in tandem.

Chapter 2: The Production Effect

The first effect of elaborative processing that I will examine is the production effect—a phenomenon where producing information, such as saying it aloud, leads to better retention compared to silently reading the same information. Historically, the production effect has been studied using vocalization tasks, but research has also explored other modalities, including typing, imagining, singing, whispering, and mouthing (Murray, 1965; Hopkins & Edwards, 1972; Conway & Gathercole, 1987; MacLeod & Bodner, 2017; MacLeod et al., 2010; Jamieson & Spear, 2014; Forrin et al., 2012).

The production effect was first examined by Murray (1965), followed by Hopkins and Edwards (1972) and later by Conway and Gathercole (1987), before being popularized by MacLeod et al. (2010). It refers to the memory advantage for words that are actively produced during study compared to those that are merely read. In a typical production effect experiment, participants are instructed to produce some words (often through vocalization) while reading others silently. During the test phase, memory for the studied words is assessed, with produced words showing superior recognition and recall compared to unproduced words.

Although originally studied through vocalization, the production effect has since been observed in various other modalities, including writing (Forrin et al., 2012), typing (Jamieson & Spear, 2014), drawing (Wammes et al., 2018), whispering (Forrin et al., 2012), mouthing (Forrin et al., 2012), and singing (Quinlan & Taylor, 2013). Research has also explored the durability of the effect, showing that it can last up to a week (Grohe & Weber, 2018; Ozubko et al., 2012). The effect is robust across different memory tasks, appearing in both recognition and recall, and

is observed in both intentional and incidental study procedures (MacLeod et al., 2010; MacDonald & MacLeod, 1998).

Beyond its impact on memory, production has also been found to reduce false alarms to semantically related lures during testing, likely due to the way it enhances encoding (Dodson & Schacter, 2001). Although typically studied using mixed-list designs, the effect has also been observed in pure-list designs, albeit with a smaller magnitude (Bodner & Taikh, 2012; Fawcett et al., 2023). These mixed findings have sparked ongoing debate about the underlying mechanisms driving the production effect.

Two principal theories of what makes production beneficial have been debated – the strength account and the distinctiveness account. The strength account proposes that it is increased strength of the memory trace conferred by production that underlies the benefit, whereas the distinctiveness account postulates that it is the distinctive act of production that gives rise to enhanced performance. Mixed and pure-list examinations of the production effect have been leveraged to investigate the role of distinctiveness and/or strength in terms of a causal mechanism behind the production effect, and differing results from the two designs have been used to argue for/against each account. An investigation from Bodner et al. (2016) examined whether the strength or distinctiveness of an item renders the production effect effective by having participants type 0%, 20%, 50%, 80%, or 100% of the words. Results showed a statistical production effect, where a larger mixed-list production effect was found when only 20% of the items were produced than when 80% of the items were produced. In Experiment 2, the authors manipulated the strength of unproduced items by increasing the study time for these items, thereby equating produced and unproduced items in terms of strength. This attenuated the pure-

list production effect, but did not attenuate the effect in the mixed list condition, thereby lending support for influence of both distinctiveness and strength on the production benefit. Other examinations of the production effect have led to mixed findings, where pure-list production effects have been observed in some experiments but not others (MacLeod & Bodner, 2017; Fawcett & Ozubko, 2016). Thus, the strength versus distinctiveness debate continues.

Lastly, exceptions have been noted, where reverse production effects have also been observed, where the statistical distinctiveness of produced versus read words is varied (e.g. 80% of the words were read aloud and 20% were silent). In these cases, a memorial benefit for silent items has been observed in free recall (Icht et al., 2014).

Rare exceptions aside, and regardless of stimulus type or method, the production effect is robust. In MacLeod et al. (2022), the authors showed that across three experiments, the production effect was independent of word frequency effects, picture superiority effects, word versus nonword advantages, and consistent whether the type of memory test was recall or recognition. As such, the production effect appears robust to other factors that are known to affect memory performance.

Whereas many of the studies examining the production effect have relied on visual presentation of the stimuli, the production effect has also been extended using the auditory modality during the study phase (i.e., participants heard the to be remembered words during the study phase instead of being visually presented with them; Mama & Icht, 2016). A partial production effect has also recently been examined, where some or all of the letters of the produced word were typed, modulating the production effect, such that how much of an item is produced can moderate the size of the production effect (Kelly et al., 2022).

In a recent investigation, Fawcett et al. (2022) found evidence for the activation of conceptual content in memory. Reduction of the production effect was observed when synonym lures (e.g., poison - venom) were used, but not when homophone lures (e.g., toad - towed) were used (relative to unrelated lures). The authors suggested that the production effect enhances conceptual encoding, and that the production benefit is driven by the activation of other semantic features in memory, as well as the motoric (e.g., speaking) act of production. The authors postulated that the activation of semantic information is consistent with a recent neuroimaging study, that showed increased activation of the anterior temporal lobe regions that are known to be associated with semantic processing (Bailey et al., 2021). Based on the emerging picture, it would seem that production also leads to semantic elaboration of a studied item. However, a similar claim was made in parallel debates on the generation effect in the 1980s but was refuted by Johns and Swanson (1988). In their study, the authors were able to observe a generation effect with nonwords, such that recognition rates were higher for generated than read nonwords.

Yet a more recent debate returned, where the production effect was studied in both true and false memory, using the DRM paradigm. Lu et al. (2024) found increased false memory to semantically related lures and suggested that their pattern of findings was due to enhanced relational or gist processing. More specifically, the authors posited that the act of production may add both semantic features and contextual features. Thus, the debate is ongoing where a mixed pattern of results has been obtained over the years.

The production effect has also been shown to be consistently robust over a range of materials (MacLeod et al., 2022). For example, high and low frequency words in recognition, pictures and words in free recall, and incidental learning of words and nonwords in recognition.

Yet a recent investigation provided evidence that the production effect may be somewhat material independent, where a benefit of production is still observed even if the production does not match study items, coined the mismatching production effect (i.e., producing "fence" when presented with the target "table"; Kelly et al., 2024).

Even with all the progress that has been made over the last half century in this domain, we are still studying the ways in which we can explain and modulate the production effect. More generally, with all of the other areas of effect-driven research that have been discussed thus far (specifically the generation effect and the testing effect), we are also still studying the ways in which elaborative processing affects memory and information processing. In sum, all the aforementioned effects are driven by doing something above and beyond simply reading a word: passive versus active engagement of study material. Next, the first experiment in this dissertation on the production effect serves as a foundation for a standardized procedure, enabling a systematic investigation of multiple elaborative processing effects.

Experiment 1: The Production Effect

Experiment 1 examined the production effect using a standard recognition procedure. The standard design was used to provide a clear test of the model, and to compare results with other effects of elaborative processing – the generation effect and the testing effect. Moreover, word-pairs were used as the stimuli, so that parallel investigations for the generation effect and testing effect could be conducted using the same stimuli.

Method

Participants. According to a two-tailed 80% power analysis conducted with G*Power 3.1 (Faul et al., 2009), a minimum of 15 participants was needed, with alpha set to 0.05. We used an effect size of $d = 0.81$, obtained from Jamieson and Spear (2014). However, as the current design involved three different conditions (a mixed-list, a pure-produce, and a pure-read condition) we sought to exceed this amount by more than threefold, as we included an additional procedural change of collecting our data online. As such, data were collected from a total of 97 participants. From the analysis, seven participants were excluded. Four were excluded on the basis of self-reporting that they were distracted while doing the experiment (e.g., in class or listening to music) and three were excluded for not complying with the experimental instructions at least 80% of the time. Non-compliance to instructions was defined as not typing at least 80% of the time on produce trials, or typing on read trials. Of the 90 participants included in the analysis, 60 were female, 29 were male, and one undisclosed. The mean age of participants was 20 years (range = 17–56, SD = 5.62 years). Data were collected to ensure that there were at least 30 participants in each of the 3 conditions (the mixed, pure-produce, and pure-read conditions).

Materials. Each participant was presented with a total of 40-word pairs during the study phase. This number allowed for recognition performance above chance, but not at ceiling, based on previous pilot work. Participants were then tested for their recognition of the second word in each pair, one at a time, along with an equal number of new, unstudied lures (80 words in total). The new unstudied lures presented in the test phase did not rhyme with the old studied words. The word pairs that were used were four letter rhyming pairs (e.g. hear - fear) and were taken from Johns and Swanson (1988). All word pairs were constructed from different common three

letter stems with first letter substitutions (see Appendix B for stimuli). For example, as seen in the first row of Appendix B, there are three words and nonwords that are all constructed from the same common stem “ack” (e.g. words: tack, back, and pack, and nonwords: cack, nack, and gack).

The 80-word pairs were randomly selected from each rhyming word-triplet. Each set of words had an equal chance of being selected. Each word pair was randomly assigned to serve as an old study item or new lure presented during the test phase.

Procedure. Participants were tested online using jsPsych version 6.3.1, a JavaScript library for running behavioral experiments via a web browser (https://jspsych.org; de Leeuw, 2015). The experiment was hosted online using GitHub Pages. At the beginning of the experiment, participants were randomly assigned to a mixed-, pure-produce, or a pure-read condition.

All participants provided informed consent which was followed by instructions. The instructions asked participants to turn off any background audio and to remain in full screen for the duration of the experiment. After this, they were presented with instructions for the experiment. The production task involved typing the word that was presented. Participants were instructed that if they saw a word appear in green, their job would be to type that word, and if they saw a word appear in red, their job would be to read that word silently. Also embedded within these instructions was an attention check. Participants were told that when asked to “provide an answer” on the following screen, to provide the answer to “2+2 = __”. If the participant typed in anything but the number “4”, the experiment looped back to the instruction screen until the correct answer was provided.

Next, the participants were given a short practice phase that included “produce” and “read” trials and that provided feedback on their performance. The practice study phase consisted of 16 practice trials, 8 of which were “produce” trials, and 8 of which were “read” trials, presented in random order. For the “produce” trials, if participants typed the second word correctly, they were presented with feedback that read “Correct! You typed correctly!”. If participants typed anything except the exact second word, they were presented with feedback that read “Incorrect. You must type the exact word”. For the “read” trials, if participants did not type anything, they were presented with feedback that read “Correct! Thank you for reading!”. Conversely, if the participant typed anything on “read” trials, they were presented with feedback that read “Incorrect. Please do not type.”. This was followed by a practice test phase in which participants also received feedback on their recognition performance.

Once the study phase commenced, each study word pair appeared on the screen one at a time for 3,000 ms, in either red or green, easy-to-read 48px font on a white background. After each study word pair, the screen was cleared for 500 ms, after which the next study word pair appeared. All 40 study pairs were presented in random order. After all of the 40 study word pairs were presented, instructions for the test phase were presented.

After the study phase was over, participants were presented with test instructions, which informed them to press ‘y’ if they recognized a word, and ‘n’ if they did not. During the test phase, participants were tested on their recognition of all previously studied targets, along with an equal number of unstudied lures. Each test word appeared in the center of the screen, in black 48px font, on a white background, until the participant pressed ‘y’ or ‘n’. Once the participant responded, the screen was cleared for 500 ms, after which the next test word appeared. This

procedure was repeated until all 80 test words has been presented in random order. After the test phase finished, participants were presented with a demographic questionnaire, followed by a question that asked whether they were doing anything else while completing the experiment, and finally a debriefing screen.

Results and Discussion

All analyses were conducted in R (R Development Core Team, 2022; version 4.2.1, 2022). The results of the mixed-list (first row), pure-produce (second row), and the pure-read (third row) conditions are displayed in Table 2.1. Hit rates for the targets are shown in the first two columns; the rates of false alarms are shown in the third column. The production advantage in both the mixed- and pure-list conditions is displayed in the last column.

Table 2.1. *Results from Experiment 1.*

	Targets		Foils	PE
Condition	Produced	Read		
Mixed	0.74 (0.02)	0.67 (0.02)	0.31 (0.01)	7%
Pure Produced	0.75 (0.01)		0.25 (0.02)	10%
Pure Read		0.65 (0.02)	0.27 (0.03)	

Note. Targets = studied words, Foils = unstudied words, and PE = production effect. Proportion *old* with SE in parentheses is displayed.

For the mixed-list condition, the first comparison revealed that participants responded "yes" significantly more to old items than to new ones, $t(29) = 9.33$, $p < 0.05$, $d = 0.64$. A mixed-list production effect was also obtained, $t(29) = 2.28$, $p = 0.03$, $d = 0.11$, with a production advantage of 7%. The pure-list production effect was also significant using the Welch correction for unequal variances, $t(48.60) = 2.29$, $p = 0.03$, $d = 0.59$, with a production advantage of 10%.

Strength or Distinctiveness? If it is strength driving the results in the production effect, then the size of the production effect in the mixed- and pure-list conditions should be of similar size (although we typically see a larger effect mixed- than pure-lists). However, if it is primarily due to distinctiveness, then the size of the production effect ought to be significantly larger in the mixed-list condition compared to the pure-list condition. To assess the role of distinctiveness, I next applied an Erlebacher analysis.

The current design that was used had both independent and dependent variables, where this mixed type of comparison evades traditional methods. However, the Erlebacher technique allows for the comparison of the independent variable (produced or unproduced) in the within- versus between-subjects designs by means of a modified ANOVA. Specifically, the Erlebacher method provides an unbiased estimate of the effects of design type, as well as the interaction between design type and the independent variable of interest (production vs. reading). The threshold of rigor that was used to determine statistical significance for all analyses was $p < 0.05$.

A 2 (item type: produced vs. read) × 2 (design: mixed vs. pure) Erlebacher ANOVA was conducted. The ANOVA revealed a main effect of production, $F(1, 58) = 10.01$, $p < 0.01$, $\eta^2 = 0.06$, but no main effect of design type, $F(1, 58) = 0.01$, $p = 0.96$, $\eta^2 < 0.01$, the latter of which indicated that targets were not recognized more overall in the pure-list conditions (70%)

compared to the mixed-list condition (71%). Critically, there was no significant interaction, $F(1, 58) = 0.57$, $p = 0.50$, $\eta^2 < 0.01$, as there was no significant difference in the size of the production effect in the pure-list conditions (10%) compared to the mixed-list condition (7%).

A cost or a benefit? Lastly, in terms of costs and benefits of production, it is assumed that a benefit to produced items is observed if the hit rate for produced items in the mixed-list condition is larger than in the pure-produce condition. Conversely, it is assumed that there is a cost to unproduced items if the hit rate for read items in the mixed-list condition is lower than the hit rate for read items in the pure-read condition. A benefit for produced items would support the notion that the act of production strengthens memory, whereas a cost would suggest that read items suffer a disadvantage when they compete with produced items in a mixed-list condition.

The hit rate for produced items in the mixed-list condition was 74%, whereas it was 75% in the pure-produce condition (a 1% difference). The hit rate for the read items in the mixed-list condition was 67%, whereas it was 65% in the pure-read condition (a 2% difference). Thus, the current results indicate that there was neither a cost nor a benefit given the negligible differences.

Summary. Experiment 1 was a standard production effect experiment which was conducted in both mixed- and pure-list designs. In the mixed-list condition, the standard production effect was obtained, with a higher hit rate for produced items over read items. In the pure-list condition, the sometimes elusive pure-list production effect was also obtained, with a higher hit rate for the pure-produced items over the pure-read items. To assess the roles of strength and/or distinctiveness, the size of the production effect in the mixed-list condition was compared against the size of the production effect in the pure-list conditions. The analysis revealed no significant difference, indicating that distinctiveness did not play a significant role in the current set of

results. Thus, in a production-by-typing procedure, the current set of results indicate that strength plays a larger role than distinctiveness when driving the production effect. The next goal of this dissertation was to investigate other effects of elaborative processing, and to assess the roles of strength and distinctiveness beyond the production effect. Experiment 2 will investigate the generation effect using the same design and materials as in Experiment 1.

Chapter 3: The Generation Effect

The next effect of elaborative processing that I examined was the generation effect. The generation effect is yet another laboratory demonstration that elaborative processing aids memory. Dating back to Jacoby (1978) and Slamecka and Graf (1978), the generation effect is the memorial benefit for words that a person generates during study. In a typical generation effect procedure, a participant is presented a word and a completion, where participants are instructed to generate the completion based on a rule (e.g., HOT - C _ _ _?) in the generate condition. In the read condition, participants are instructed to simply read the completed word pair (e.g., HOT - COLD). The rule for generating can take on many forms, such as antonyms, synonyms, categories, rhymes (Slamecka & Graf, 1978), semantic associates (Begg et al., 1989), translation (Slamecka & Katsaiti, 1987), and definitions (Forrin et al., 2014). The generation effect has also been examined in a number of different experimental contexts, including math problems (McNamara & Healy, 2000), pictures (Kinjo & Snodgrass, 2000), learning (Lutz, 2003), aging (Java, 1996), and free recall (McDaniel et al., 1990). Research has also investigated important differences in the generation effect for internal or external generation (i.e. experimenter-generated or self-generated words; McFarland et al., 1980), and when they are applied to causal relations when reading scientific texts (Abel & Hanze, 2019). In a particularly intriguing study, the generation effect was investigated with monkeys (Kornell & Terrace, 2007), although this finding has since been challenged (Staniland et al., 2015).

Begg et al. (1991) found that generation and reading had equal memorial benefit when participants imagined the read items and that generating represents a particular kind of discriminative encoding:

> "Reading does not ensure that items will be processed in a way that is appropriate for the test. Generating has a better chance of doing just that. Although there is some question about whether the good thing that generating does is to distinguish the items, or to engage other useful skills that subjects have in their bag of cognitive tricks, it seems clear that it is the result of applying these tricks, not the reason for doing so, that is important. Generating might make items distinctive, and might even guarantee it, but reading can do every bit as good a job, although it is unlikely to do so spontaneously." p. 494

As with the other effects discussed thus far, it is the act of doing something above and beyond reading that confers a benefit to memory performance. It is the active engagement with material that is important, instead of merely being exposed to it (Tulving, 1962).

The generation effect has even been examined using words and nonwords (Johns & Swanson, 1988). Here the authors were able to obtain a generation effect with the nonwords, which challenged the current prevailing wisdom that stimuli must contact semantic memory to elicit a generation effect. Effectively, by presenting the target at the end of each trial, nonwords were able to become familiar enough to remove the confound of the familiarity of words over nonwords. This study in particular will be pertinent to the experiments that are presented later.

The generation effect is more robust in mixed than pure-list study designs (Slamecka & Katsaiti, 1987). Although there are differences between the production effect and the generation effect, there are strong similarities. In a recent meta-analysis of 310 experiments examining the generation effect (McCurdy et al., 2020), the mechanism of the generation effect again came into question, where seven different theories of the generation effect were identified.

Of the theories presented, there were two types: item memory theories and context memory theories. Of the item memory theories, mental effort, selective displaced rehearsal, semantic (lexical) activation, and two-factor/multi-factor were put forth. For example, the multifactor theory posits that two mechanisms are responsible for improving memory for self-generation: enhanced item-specific processing and enhanced relational processing (i.e., processing of the cue-target relational information).

The context memory theories included associative strengthening, item-context tradeoff, and a processing account (McCurdy et al., 2020, Mulligan, 2004). For example, the processing account postulates that the processes required for self-generation and reading are different, thereby leading to differential context memory effects. That is, generation effects can be found for conceptual details like source and cue words, but not for perceptual details like font color, and vice versa for reading.

Thus, authors showed support for some theories, and not others. However, no one general or cohesive account was offered. As with other related effects, a theoretical account of the generation effect remains debated but unresolved. Despite the lack of a singular theoretical explanation, the generation effect remains a well-documented memory phenomenon, with broad implications for learning and instruction. Its robustness across various experimental paradigms

suggests that self-generation enhances memory through multiple, complementary mechanisms rather than a single underlying process.

Further, the interaction between the generation effect and other elaborative processing effects, such as the production effect and the testing effect, raises important questions about their shared contributions to memory performance. While both the production and generation effects involve active engagement with material, they differ slightly in the nature of that engagement—production emphasizing encoding via motor action (e.g., typing), whereas generation requiring the active retrieval or construction of a target item based on a cue (but in the present investigation, also by typing).

Building on previous foundational research, the next experiment proposed next will examine the generation effect using the same materials and design as used in Experiment 1 for the production effect to examine the contributions of strength and/or distinctiveness. By systematically manipulating the elaborative processing task at hand, these experiments aim to further elucidate the mechanisms driving the generation effect and its relation to broader principles of memory and learning.

Experiment 2

The purpose of Experiment 2 was to assess the generation effect procedure using the same materials and design as used in the production effect in Experiment 1. This consistency allowed for a direct comparison of the production and generation effects. Both mixed- and pure-lists were again used to evaluate the roles of strength and distinctiveness in the generation effect. As before, the signature of a distinctiveness-based generation effect is a larger mixed-list

generation effect than a pure-list generation effect (McDaniel & Bugg, 2008). Conversely, the signature of a strength-based generation effect is equal sizes of generation effects across mixed- and pure-lists (i.e., no difference in the size of the benefit of generated items over read items across the two designs).

Method

Participants. Participants were recruited in the same way as in Experiment 1. Data were collected from a total of 105 participants. From the analysis, 15 participants were excluded. Eight were excluded on the basis of self-reporting that they were distracted while doing the experiment (e.g., in class or on the phone) and three were excluded for not complying with the experimental instructions at least 80% of the time. Of the participants that were included in the analysis, 50 were female and 40 were male. The mean age of participants was 19.86 years (range = 18–42, SD = 4.38 years).

Materials. The materials were the same as in Experiment 1.

Procedure. The procedure was identical to Experiment 1, with the difference that participants were instructed to generate the words from cues presented in green but to read the words presented in red. As before, participants were randomly assigned to one of the three between-subjects conditions: mixed, pure-generate, and pure-read.

Results and discussion

All analyses were conducted in R (R Development Core Team, 2022; version 4.2.1, 2022). The results of the mixed-list (first row), the pure-generate (second row), and the pure-read

(third row) conditions are displayed in Table 3.1. Hit rates for the targets are shown in the first two columns, and the rates of false alarms are shown in the third column. The generate advantage in both mixed- and pure-list conditions is displayed in the last column.

Table 3.1. *Results from Experiment 2.*

	Targets		Foils	GE
Condition	Generated	Read		
Mixed	0.83 (0.02)	0.58 (0.02)	0.24 (0.01)	25%
Pure Generated	0.82 (0.01)		0.24 (0.02)	16%
Pure Read		0.66 (0.01)	0.29 (0.02)	

Note. Targets = studied words, Foils = unstudied words, and GE = generation effect. Proportion *old* with SE in parentheses is displayed.

The first comparison revealed that participants responded "yes" to items that they studied more than to items that they did not (i.e. old vs. new items), $t(29) = 20.21$, $p < 0.001$, $d = 0.75$. More critically, a mixed-list generation effect was observed, $t(29) = 6.82$, $p < 0.001$, $d = 0.40$, with a generate advantage of 25%. A pure-list generation effect was also observed using the Welch correction for unequal variances, $t(57.11) = 4.67$, $p < 0.001$, $d = 1.21$, with a 16% advantage for generate items.

Strength or Distinctiveness? As in Experiment 1, if strength is driving the results in the generation effect, then the generation effect in the mixed- and pure-list conditions should be of similar size. However, if it is primarily due to distinctiveness, then the size of the generation effect ought to be significantly larger in the mixed-list condition compared to the pure-list condition. To assess the role of distinctiveness, I next applied an Erlebacher analysis.

A 2 (item type: generated vs. read) × 2 (design: mixed vs. pure) Erlebacher ANOVA was conducted. The ANOVA revealed a main effect of generation, $F(1, 58) = 66.35$, $p < 0.01$, $\eta^2 = 0.33$, but no main effect of design type, $F(1, 58) = 1.52$, $p = 0.22$, $\eta^2 < 0.01$, the latter of which indicated that targets were not recognized more overall in the pure-list conditions (74%) compared to the mixed-list condition (71%). Critically, although there was a 9% difference between the mixed- and pure-list generation effects, the interaction was not significant, $F(1, 58) = 2.68$, $p = 0.11$, $\eta^2 < 0.01$.

A cost or a benefit? Lastly, in terms of costs and benefits of generation, it is assumed that a benefit to generated items is observed if the hit rate for generated items in the mixed-list condition is larger than that in the pure-generate condition. Conversely, it is assumed that there is a cost to read items if the hit rate for read items in the mixed-list condition is lower than the hit rate for read items in the pure-read condition.

The hit rate for generated items in the mixed-list condition was 83%, whereas it was 82% in the pure-generate condition (a 1% difference). The hit rate for the read items in the mixed-list condition was 58%, whereas it was 66% in the pure-read condition (an 8% difference). Thus, the current set of results indicate that there was a cost for read items in the mixed-list condition, but no sizeable benefit for produced items given the negligible 1% difference.

Experiment 1 and Experiment 2. As the generation effect was larger than the production effect, an analysis of the two types of targets across the two procedures in the mixed-list condition was conducted. A mixed ANOVA with item type (produced/generated vs. read) as the within-subject factor was used to assess this difference. The analysis revealed that this difference was not statistically significant, with no main effect of experiment, $F(1, 58) = 0.01$, $p = 0.93$, $\eta^2_g < 0.01$.

However, there was a main effect of item type, $F(1, 58) = 45.40$, $p < 0.01$, $\eta^2_g = 0.19$, and a significant interaction between Experiment and item type, $F(1, 58) = 14.98$, $p < 0.01$, $\eta^2_g = 0.07$, indicating that the elaborating encoding benefit (or cost) in the generation effect (25%) exceeded that of the production effect (10%).

Summary. Thus far, I have examined two effects of elaborative processing: the production effect and the generation effect. Both effects involve actively engaging with material to enhance memory, yet they differ slightly in their methods. The production effect refers to the memorial benefit gained when a person actively produces information, such as reading aloud or typing, whereas the generation effect occurs when an individual generates an item from a given cue, such as completing a word stem. Despite their similarities, the current set of results indicates that generating an item from a stem leads to a more substantial memory advantage compared to merely producing the item, even though the difference between the two procedures is subtle (e.g., typing the word or typing part of the word when part of the target-stem completion is missing).

Building on the question of whether generating an item from a stem provides a greater memorial benefit than producing it, the next step in my investigation was to explore whether generating the entire word—rather than completing a partial cue—would produce an even larger benefit. If generating a partial word strengthens memory compared to simple production, it may follow that generating the full word might amplify this effect by requiring even more elaborative processing.

Thus, I turned my attention to another influential effect of elaborative processing: the testing effect, which I examined in Chapter 4. The testing effect is a well-established

phenomenon in which retrieving information during testing leads to significantly better retention compared to merely re-studying the material. Like generation, testing leverages active engagement with the material, but it does so by requiring retrieval. By comparing these different effects of elaborative processing, my research aimed to uncover how distinct types of active cognitive engagement—generation and retrieval—contribute to improved memory performance. The next effect of elaborative processing that will be examined in Chapter 4 is the testing effect.

Chapter 4: The Testing Effect

To bridge the gap between basic research and the applied domain, a testing effect procedure was used to examine yet another demonstration of elaborative processing that confers a benefit to memory. Again, although there are subtle differences, the testing effect, the production effect, and the generation effect are all examples of elaborative processing that involve more effortful engagement with the to be remembered material. In the production effect, participants produce material during the study phase. In the generation effect, participants produce part of the material during the study phase. In the testing effect, participants produce material during the interim study/retrieval phase.

> "A curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning (by heart, for example), when we almost know the piece, it pays better to wait and recollect by an effort within, than to look at the book again. If we recover the words the former way, we shall probably know them the next time; if in the latter way, we shall likely need the book once more" (James, 1890, p. 646).

Although there have been arguments that suggest the generation effect and the testing effect are different, much like the argument that production and generation are different, there is recent evidence that suggests they are similar when controlling for design effects. Mulligan et al. (2018) argued that an important difference between the testing effect and the generation effect is that the generation effect involves retrieval from semantic memory, whereas the testing effect involves retrieval from episodic memory (Mulligan et al., 2018; Karpicke & Zaromb, 2010). However it has also been argued that there are important similarities, where both are susceptible

to design effects, where both effects show similar patterns of results in recognition and free recall, where conditions that produce negative effects in free recall, both produce positive effects in recognition memory (Burns, 1992), and where both disrupt order information in memory (Karpicke & Zaromb, 2010; Mulligan et al., 2018). Interestingly, the same can also be said for the production effect (Forrin & MacLeod, 2016). Furthermore, the testing effect is larger in mixed-list than in pure-list conditions, a stable finding regarding both the production effect and generation effect (Mulligan et al., 2019). The testing effect has been examined in several different experimental paradigms, however the design that will be used to examine the testing effect here is one that closely resembles a generation effect procedure in the domain of recognition.

The testing effect is the finding that testing people's memory of studied materials enhances retention of to be learned material above and beyond restudying that material. In a typical testing effect procedure, the testing condition is compared across three different conditions. All three conditions commence with a study phase. Then a no test control group does nothing, a restudy control group restudies the list, and a testing group is tested on that list. Then, all three receive a final memory test. Results show a largest memory benefit following the test manipulation, even when there is no feedback given for the test.

For example, in a study by Karpicke and Roediger (2008) the testing effect was examined by presenting Swahili-English word pairs (e.g., mashua - boat) across four different conditions. During the test phase, participants were given the first item in each word pair, the Swahili word, to test their memory of the corresponding English word (e.g., mashua - ?). The first condition consisted of a simple study - test design. In the three other conditions, if the item was recalled, it

either (a) remained in the next study phase but was dropped from subsequent test phase, (b) remained in the study phase but was dropped from testing, or (c) was dropped from both study and test. When participants were tested again a week later, results showed that repeated testing conferred the largest benefit, larger than the benefit of restudying.

Roediger and Karpicke, in their 2006 review paper of the testing effect, refer to the testing effect as a small version of the Heisenberg uncertainty principle in the field of psychology: “Just as measuring the position of an electron changes that position, so the act of retrieving information from memory changes the mnemonic representation underlying retrieval – and enhances later retention of the tested information” (Roediger & Karpicke, 2006, p. 182). More simply, testing can be thought of as a kind of modulation of memory, enhancing memory through repeated retrieval.

One caveat to the testing effect is that there must be some study opportunities before any benefit can be observed. However, the testing effect challenges the common simplifying assumption that tests are merely assessment devices, and that learning only occurs during study trials.

The theoretical explanations behind the testing effect often suggest that strength is the underlying mechanism. Carpenter (2009) found evidence that retrieval practice strengthens semantic relation between concepts. Similarly, Roediger and Butler (2011) discovered that testing leads to deeper conceptual processing compared to merely rereading the same material. Pyc and Rawson (2010) demonstrated that retrieval practice strengthens connections between new and prior knowledge, such as learning a new Spanish word for an already known English word.

Moreover, research indicates that the magnitude of the testing effect is not necessarily dependent on retrieval practice performance. For instance, Chan et al. (2024) found that conditions favoring better retrieval practice did not always produce a greater testing effect. Rickard and Pan (2018) proposed a dual memory theory, which postulated that separate memories form as a consequence of study and test events. This theory helps explain the strength of the testing effect by suggesting that retrieval practice creates more robust memory representations compared to restudy conditions.

Experiment 3

In Experiment 3, the same set of stimuli used in Experiments 1 and 2 was used to enable direct comparisons across procedures and to examine how the pattern of results might differ within the context of a testing effect procedure. Since the testing effect is another form of elaborative processing, it is expected to yield results similar to the production and generation effects, further emphasizing the importance of active engagement in enhancing memory. As in Experiments 1 and 2, participants were assigned to one of three conditions: a pure-read condition, a pure-test condition, or a mixed-list condition. Following this assignment, participants proceeded to a recognition test phase. During this phase, they were presented with a list of words that included all the items they previously studied as well as an equal number of new, unstudied words. The recognition test phase was a standard yes/no format, requiring participants to identify whether each word was part of the original study set.

Participants. Participants were recruited in the same way as in Experiment 1 and 2. Data were collected from a total of 112 participants. From the analysis, 22 participants were excluded. Seven were excluded for self-reporting that they were distracted while doing the experiment

(e.g., in class), and 15 were excluded for not complying with the experimental instructions at least 80% of the time. Of the participants that were included in the analysis, 53 were female, 36 were male, and one undisclosed. The mean age of participants was 19.59 years (range = 17–37, SD = 2.64 years).

Materials. The materials were the same as in Experiment 1 and 2. However for the testing effect, the word pairs were scrambled such that each pair no longer rhymed. This was done so that when participants were asked to do recall practice in the testing effect procedure, they were in fact recalling a word and not using a rhyme rule to produce their answer.

Procedure. As with the production effect and the generation effect procedures, during the study phase, participants were presented with pairs of words for 2000 ms. After the study phase, one third of subjects were assigned to a mixed-list test condition, one third of the participants were assigned to a pure-test condition, and one third of the participants were assigned to a pure-read condition.

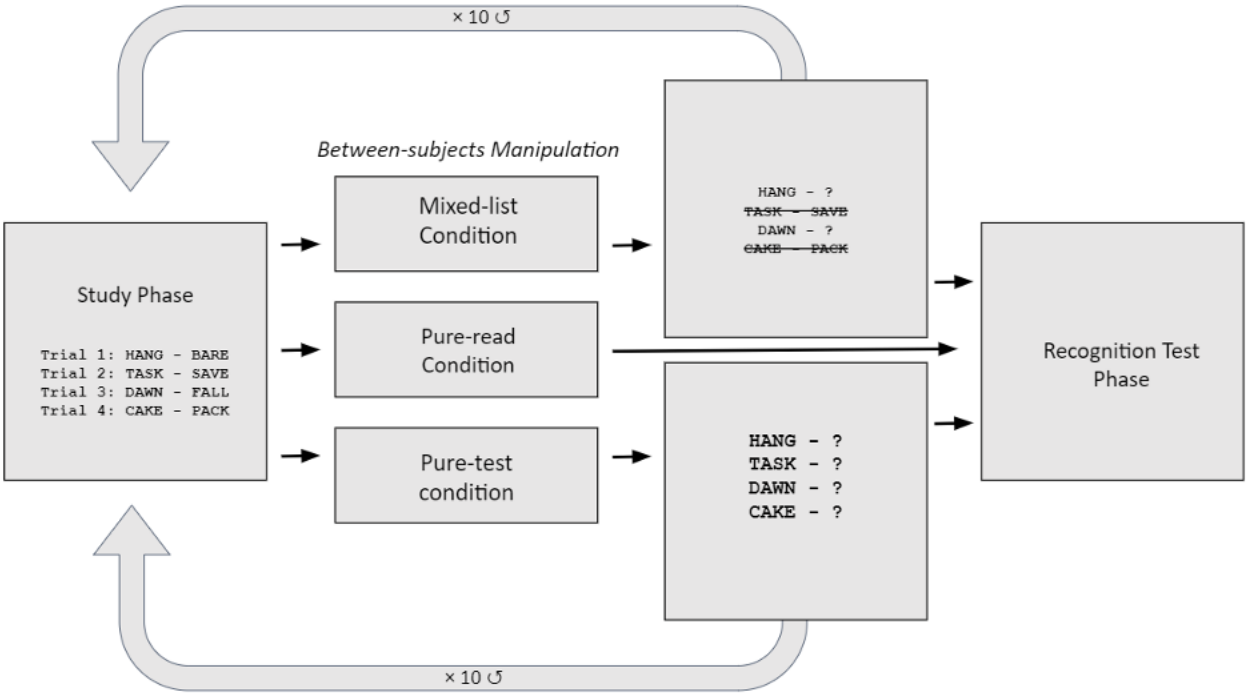
In the pure-read condition, in the study phase, participants were presented with all the 40 intact cue-target pairs (e.g., task - save), where they then proceeded straight to the recognition test. In the pure-test condition, the study phase was blocked, such that participants saw four intact cue-target pairs and then were given a test phase for the second word in each of the four pairs (e.g., task - ?). Study and test blocks were independently randomized in terms of stimuli order. Participants were asked to type the corresponding word for each pair in the space provided. This study/test cycle of presenting four cue-target pairs, and then testing the second word in each pair, repeated ten times, such that the total number of study cue-target pairs was 40. Once this cycle was completed, participants proceeded to the standard yes/no recognition test.

In the mixed-list test condition, the study phase was again blocked, such that participants saw four intact cue-target pairs, but participants were then presented with only half of the incomplete cue-target pairs (two), where they were asked to recall and type the second word in the space provided (e.g., park - ?). The two (out of the four) word-pairs that were selected for testing were selected randomly for each block. Again, this study/test cycle of presenting four cue-target pairs, and then testing two of the four cue-target pairs, was repeated ten times, such that the total number of cue-target pairs that were studied was 40, and the total number of cue-target pairs that were tested was 20. Once this cycle was completed, participants then proceeded to the standard yes/no recognition test.

In all three conditions, when participants proceeded to the recognition test phase, they were presented with the second word in each pair that was studied, along with an equal number of unstudied single word lures. Test words were presented one at a time, in the center of the screen, in black font. Each trial persisted until the participant responded “yes” or “no”. See Figure 2 for the complete study design.

Figure 4.1

Testing effect study design.



Note. A depiction of how participants were assigned to one of three conditions after the study phase, before they proceeded onto the recognition test phase. In the mixed-list and pure-test conditions, participants studied blocks of four words and then were tested on two words, or four words, respectively. This cycle continued ten times until all forty words were studied, where they were then given a standard yes/no recognition test. In the pure-read condition, participants studied all forty words and went straight to the recognition test phase.

Results and discussion

All analyses were conducted in R (R Development Core Team, 2022; version 4.2.1, 2022). The results of the mixed-list (first row), the pure-tested (second row), and the pure-read (third row) conditions are displayed in Table 4.1. Hit rates for the targets are shown in the first

two columns, and the rates of false alarms are shown in the third column. The testing advantage in both mixed- and pure-list conditions are displayed in the last column.

Table 4.1. *Results from Experiment 3.*

	Targets		Foils	TE
Condition	Tested	Read		
Mixed	0.72 (0.01)	0.67 (0.01)	0.29 (0.02)	5%
Pure Tested	0.70 (0.01)		0.26 (0.02)	6%
Pure Read		0.64 (0.01)	0.28 (0.02)	

Note. Targets = studied words, Foils = unstudied words, and TE = testing effect. Proportion *old* with SE in parentheses is displayed.

I started by examining the results of the mixed-list condition. The first comparison revealed that participants responded "yes" significantly more to old items than to new ones, $t(29) = 9.81$, $p < 0.001$, $d = 0.62$. A mixed-list testing effect was also obtained, $t(29) = 2.45$, $p = 0.02$, $d = 0.10$, with a testing advantage of 5%. The pure-list testing effect was not significant using the Welch correction for unequal variances, $t(55.61) = 1.68$, $p = 0.10$, $d = 0.43$, although there was a memorial testing advantage of 7%.

Strength or distinctiveness? As in Experiment 1 and 2, if there is a strength mechanism driving the results in the testing effect, then the size of the testing effect in the mixed- and pure-list conditions should be of similar size. However, if it is primarily due to distinctiveness, then the size of the testing effect ought to be significantly larger in the mixed-list condition compared to the pure-list condition. To assess the role of distinctiveness, I next applied an Erlebacher analysis.

A 2 (item type: tested vs. read) × 2 (design: mixed vs. pure) Erlebacher ANOVA was conducted. The ANOVA revealed a main effect of testing, $F(1, 58) = 7.73$, $p = 0.01$, $\eta^2 = 0.04$, but no main effect of design type, $F(1, 58) = 0.48$, $p = 0.49$, $\eta^2 = 0.01$, the latter of which indicated that targets were not recognized significantly more overall in the pure-list conditions (67.5%) compared to the mixed-list condition (69.5%). Critically, there was no significant interaction, $F(1, 58) < 0.01$, $p = 0.98$, $\eta^2 < 0.01$, as there was no significant difference in the size of the testing effect in the pure-list conditions (7%) compared to the mixed-list condition (5%).

A cost or a benefit? Lastly, in terms of costs and benefits of testing, it is once again assumed that a benefit to tested items is observed if the hit rate for tested items in the mixed-list condition is larger than that in the pure-test condition. Conversely, it is assumed that there is a cost to read items if the hit rate for read items in the mixed-list condition is lower than the hit rate for read items in the pure-read condition.

The hit rate for tested items in the mixed-list condition was 72%, whereas it was 71% in the pure-test condition (a 1% difference). The hit rate for the read items in the mixed-list condition was 67%, whereas it was 64% in the pure-read condition (a 3% difference). Thus, the current results indicate that there was neither a cost nor a benefit given the negligible differences.

Experiment 1, Experiment 2, and Experiment 3. In the mixed-list conditions, there was a larger generation effect than a production effect (a 25% advantage vs. a 7% advantage. Yet there was not a larger testing effect (5%) than a generation effect (25%). An analysis of the two types of targets across the three procedures in the mixed-list conditions was conducted. A mixed ANOVA with item type (produced/generated/tested vs. read) as the within-subjects factor was used to assess this difference. The analysis revealed no main effect of experiment, $F(2, 87) = 0.03$, $p =$

0.97, $\eta^2_g < 0.01$, but a main effect of condition, $F(1, 87) = 50.25$, $p < 0.01$, $\eta^2_g = 0.14$, and a significant 3-way interaction, $F(2, 87) = 11.91$, $p < 0.01$, $\eta^2_g = 0.07$, indicating that the interaction between elaborative processing (produced/generated/tested vs. read) and design type (mixed vs. pure) differed based on procedure type (i.e., a production, generation, or testing procedure). Specifically, it was observed that the largest elaborative processing benefit was in the generation effect procedure and the smallest elaborative processing benefit was found in the testing effect procedure.

With all the empirical results in hand, a model of elaborative processing is presented next. In the next chapter, I will present a model based on MINERVA 2 that is able to account for the empirical results found in this dissertation.

Chapter 5: A model of elaborative processing

To account for the results of three effects of elaborative processing, the production effect, the generation effect, and the testing effect, a formal modeling based on the MINERVA 2 model of human memory is presented.

A Model of Elaborative Processing

MINERVA 2 is a global matching model and accounts for memory storage, retrieval, and decision. As previously mentioned in the introduction, the MINERVA 2 model of the production effect was utilized and expanded upon for this project (Jamieson et al., 2016). The goal was to show that similar mechanisms can account for three effects of elaborative processing: the production effect, the generation effect, and the testing effect. All three effects were explained under a common theoretical framework. Under this common theoretical framework, all three effects were accounted for by similar mechanisms, where the act of elaboration, whether that be by producing an item, generating an item, or producing an answer when being tested on an item, confers a benefit to the memory trace whereby additional features are added to represent this process, and that can then be used to aid retrieval.

MINERVA 2 has had a wide range of success in a number of domains, including category learning, levels of processing effects, word identification, decision making, artificial grammar learning, associative learning, implicit rule learning, the production effect, semantic knowledge, sentences and metaphors, directed forgetting, and verbal short tern memory (Hintzman, 1984; 1986; 1988; Collins et al., 2020; Goldinger, 1988; Kwantes & Mewhort, 1999; Doughterty et al., 1999; Jamieson & Mewhort, 2009; Jamieson et al., 2012; Chubala et al, 2016;

Jamieson et al., 2016; Jamieson et al., 2018; Reid et al., 2023; Reid & Jamieson, 2023; Guitard et al., 2025; see also Jamieson et al., 2022).

The MINERVA 2 model assumes that memory is a matrix, where each row represents an item, and each column represents the specific features of each item. When words are encoded into memory, they are encoded to a unique row in memory, but with some degree of noise. This noise is controlled by the parameter *L*, which is the learning parameter in the model.

With the current tasks at hand —the production effect, the generation effect, and the testing effect—depending on whether results follow a distinctiveness account or a strength account, under the MINERVA framework, the results can be modeled using two different mechanisms. If it appears that it is strength that is driving the results, then the results can be modeled by varying the learning parameter, *L*, in the simulations, so that elaborated items are stored with more intact information than unelaborated items. Specifically, varying *L* determines how much information is encoded into memory, such that higher levels of *L* encode more information, thereby increasing the strength of an item, and where decreasing *L* decreases the amount of information that is encoded, thereby decreasing the strength of an item. Therefore, varying *L* for the elaboratively processed items can implement the assumption that strength is increased for these items, and it is this process that is responsible for the benefit conferred by production/generation/testing.

However, if it is distinctiveness that is driving results, then the results can be modeled using the addition of extra features onto each item that is processed in an elaborative way (see Figure 5.1). By adding extra features to each item that is processed in an elaborative way, this implements the assumption that the act of production/generation/testing adds distinctive information to a memory trace (against a backdrop of no distinctive information added to

unelaborated items), where this distinctive information can be used at retrieval to benefit recognition. However, if it is a mix of strength and distinctiveness that is driving results, then both mechanisms can be adjusted in tandem, to account for the influence of both.

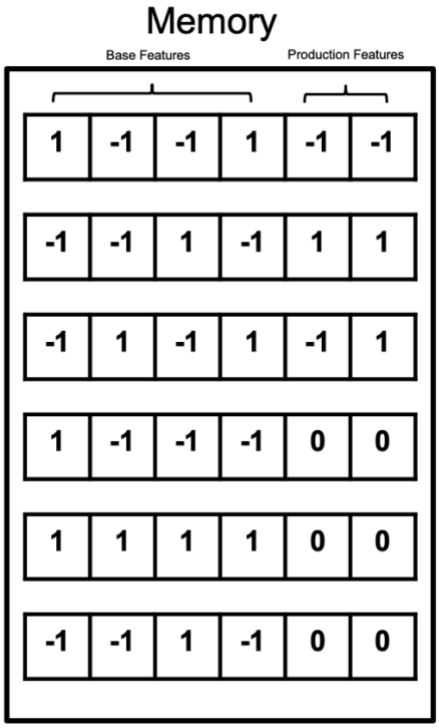


Figure 5.1. An example of how distinctiveness is represented in the model. Produced items (first three rows) are assigned additional unique features (last two columns), whereas unproduced items (last three rows) receive no extra features.

Retrieval. If an item is elaboratively encoded, it will contain extra non-zero features. However, it is assumed that the initial probe that is presented to memory does not contain these distinct features, but rather these features must be retrieved from memory through an iterative retrieval process. When a probe is presented to memory, activation is similarity-based and is calculated on a feature-to-feature basis. This process operates in parallel, where all traces in memory are contacted simultaneously. These activated traces are then represented in a structure called an

echo, where this echo is comprised of two key properties: echo content and echo intensity. Echo content, c, is a vector consisting of the sum of all the traces that are activated:

$$c_{j=\sum_{i=1}^{m} a_i \times M_{ij}} \quad \{for\ each\ j = 1 \ldots n\}$$

where c_j is the j^{th} element of the echo, a_i is the activation for the i^{th} trace in memory, M_{ij} is the j^{th} element of the i^{th} trace in memory, m is the number of traces in memory, and n is the number of elements in each vector. The activation for each trace in memory is computed as the cosine similarity between the probe and that trace raised to the exponent of three (Hintzman, 1986; 1988).

The retrieval process operates iteratively, using the test probe as a retrieval probe to retrieve an echo three times. During the first iteration, the probe consists solely of the base features. In the second iteration, if the word is successfully produced, the retrieved echo content (which becomes the new probe) incorporates additional information about the extra features that were encoded during the study phase.

Following three iterations, the probe's familiarity is computed as an echo intensity, which is the sum of the activation elicited by the probe:

$$f = \sum_{i=1}^{i=m} \left(\frac{\sum_{j=1}^{j=d} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=d} p_j^2} \sqrt{\sum_{j=1}^{j=d} M_{ij}^2}} \right)^3$$

where a familiarity (*f*) value is calculated based on the probe's similarity to all traces in memory, *M*, where specifically, a cosine similarity calculation is used, p_j is feature *j* of the probe, M_{ij} is feature *j* of trace *i* in memory, *m* is the number of memory traces, and *d* is the dimensionality of these traces. The similarity calculation is transformed into activation by raising it to the power of three, which amplifies the signal-to-noise ratio for all computed familiarity values. These

similarity values are then aggregated to produce an overall familiarity index, denoted as f, also referred to as echo intensity in other applications of the MINERVA model. Distinctiveness in the model is represented and utilized through the encoding of additional nonzero features during the study phase, combined with the iterative retrieval process that retrieves these features.

The model simulates decision-making by leveraging all computed familiarity values (for both old and new words) to establish a criterion based on a selected percentile that best fits the data. For example, if the decision criterion is set to the 55th percentile, this implies that the top 45% of echo intensities with the highest familiarity would be labeled as "OLD" (corresponding to a decision criterion of 0.45), while the remaining echo intensities would be classified as "NEW." In this scenario, the criterion reflects a slightly conservative approach, suggesting that the model has a slight tendency to favor classifying items as "NEW," given that the criterion is just above the median.

Encoding. Each word is represented by a unique vector, where values are drawn from a normal distribution with $M = 0$ and a $\sigma = 1/\sqrt{(d)}$, where d is the dimensionality of each word vector (see Murdock, 1982; Jamieson & Hauri, 2012; Jones & Mewhort, 2007). In the simulations to follow, d was set to 300 for each word in the production and generation effects.

However, for the testing effect, to represent each word-pair, the total dimensionality was increased to 600. This was done to account for the fact that on testing trials in the testing effect procedure, no portion of the target was presented (only a “?” appeared). Thus, only the cue was presented to memory as a probe in these simulations, where on subsequent iterations, the cue will serve to pick up the target features in each word pair, as well as the elaborative features added, that represent the elaborative act of testing.

The current word representations are slightly different from classical instantiations of MINERVA 2 but set the model up to deal with other types of representations in future investigations, such as those derived from natural language processing models (e.g. LSA; Landauer & Dumais, 1999).

Simulation of Experiment 1

For each simulated subject, 80 randomly generated vectors were constructed (40 to serve as studied targets and 40 to serve as unstudied lures). For the mixed-list condition, 40 vectors were stored to memory, where half of these vectors (20) represented the produced items and contained extra information to represent the act of production, and half (20) represented the read items, which contained no information. The model also assumed that "produced" targets were more strongly encoded into memory with a higher value of L than the words that participants were instructed to read.

For the pure-list condition, two sets of simulations were run. In the pure-produce condition, all 40 targets contained extra information to account for distinctiveness. In the pure-read condition, no extra information was added. When each item was encoded into memory, L was also varied to account for contributions of strength.

Familiarity was computed using the iterative retrieval function as previously described. These echo intensities were then converted to a "yes" or "no" decision based on a criterion that was chosen to best fit the data.

Each simulation consisted of 1000 runs. Model performance and comparison between the empirical results and simulated results were compared using *RMSE*.

Simulation of Experiment 1 – Strength and Distinctiveness

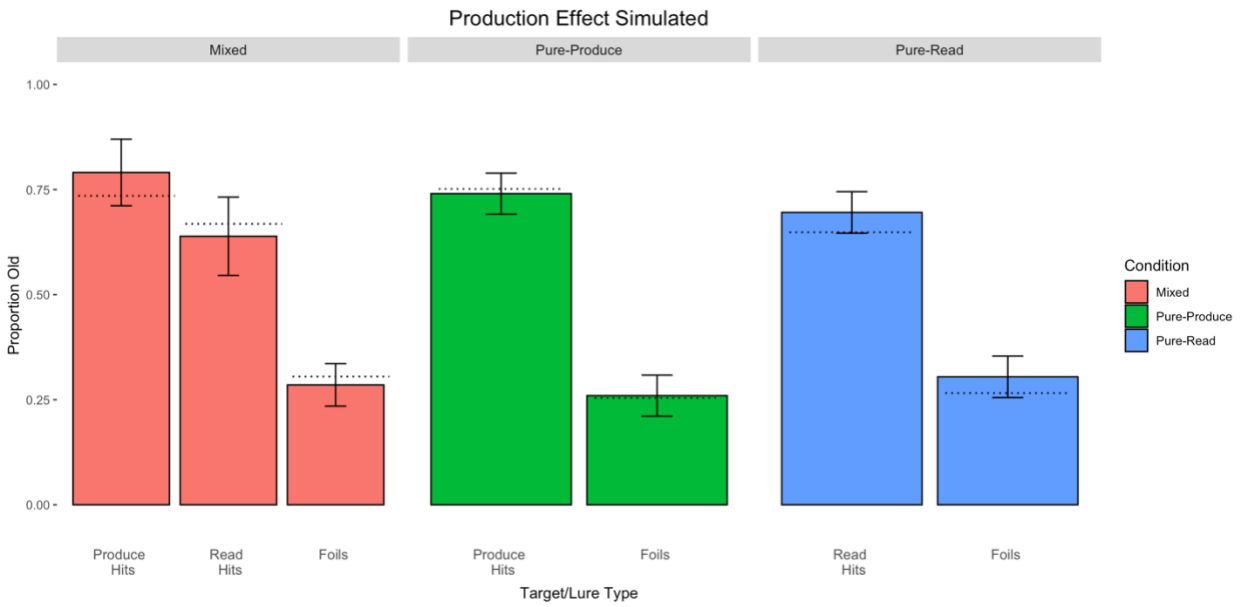


Figure 5.2. Simulation results of the production effect in Experiment 1, integrating distinctiveness and strength. Note. Mixed-list parameters: $L = 0.046$ for produced targets, $L = 0.040$ for read targets, and the number of extra production features for produced targets was 250. The decision criterion was set to an unbiased value of 0.501. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

Figure 5.2 shows the simulation results of the mixed-list condition on the left-hand side, the pure-produce condition in the middle, and the pure-read condition on the right. By assuming that produced targets are more strongly encoded than read targets and that produced targets were encoded with some extra distinctive information, I was able to generally reproduce the pattern of results that I obtained in Experiment 1 across the three conditions, *RMSE* = 0.0343.

Although the current model fits the data fairly well, the model missed some important aspects of the data. In the mixed-list condition, the model predicted a higher hit rate for produced

items but predicted a lower hit rate for the read targets, as well as the rate of false alarms for unstudied lures. In the pure-list conditions, the model closely resembled the empirical data in the pure-produce condition, however, predicted a higher hit rate for read items in the pure-read condition as well as for the new unstudied foils in that condition.

Although the current model fit the data well, we sought to explore whether we could obtain a good fit to the data using only a strength mechanism. In the obtained empirical results, we did not see a strong signature of distinctiveness, as signaled by the nonsignificant interaction between the mixed- and pure-list production effects. This suggests that the current results were likely driven by a strength mechanism. A simulation using only a strength mechanism is presented next, where no additional production features were added, and where the iterative retrieval mechanism was omitted.

Simulation of Experiment 1 – Strength Only

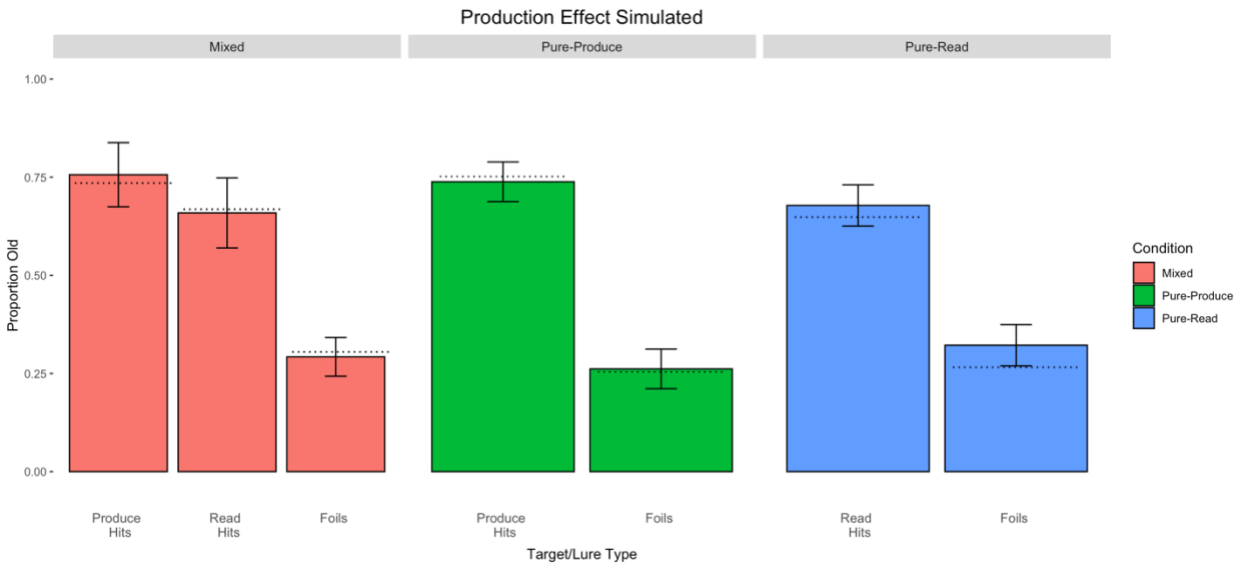


Figure 5.3. Simulation results of the production effect in Experiment 1, using only a strength mechanism. Parameters: $L = 0.0353$ for produced targets and $L = 0.0279$ for read targets. Distinctiveness was not assumed; no additional features were added. The decision criterion was an unbiased value of 0.505. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

Figure 5.3 shows the simulation results of the mixed-list condition on the left-hand side, the pure-produce condition in the middle, and the pure-read condition on the right. By assuming that produced targets are more strongly encoded than read targets, we were able to closely reproduce the pattern of results that we obtained in Experiment 1 across the three conditions, RMSE = 0.0266.

The model captured the data well; however, it again missed some key aspects of the data. In the mixed-list condition, the model predicted a slightly higher hit rate for produced and read targets, as well as for the rate of false alarms to unstudied foils. However, in the pure-produce

condition, the model predicted a lower hit rate for produced targets, as well as a slightly lower false alarm rate for unstudied foils. Additionally, the model predicted a lower rate of hits and false alarms in the pure-read condition. Despite these misses, the model captures the pattern of the empirical data well, with a strong model fit.

Using the same approach, I next present simulation results from the generation effect in Experiment 2 across the 3 conditions: the mixed-, pure-generate, and pure-read conditions.

Simulation of Experiment 2 – Strength and Distinctiveness

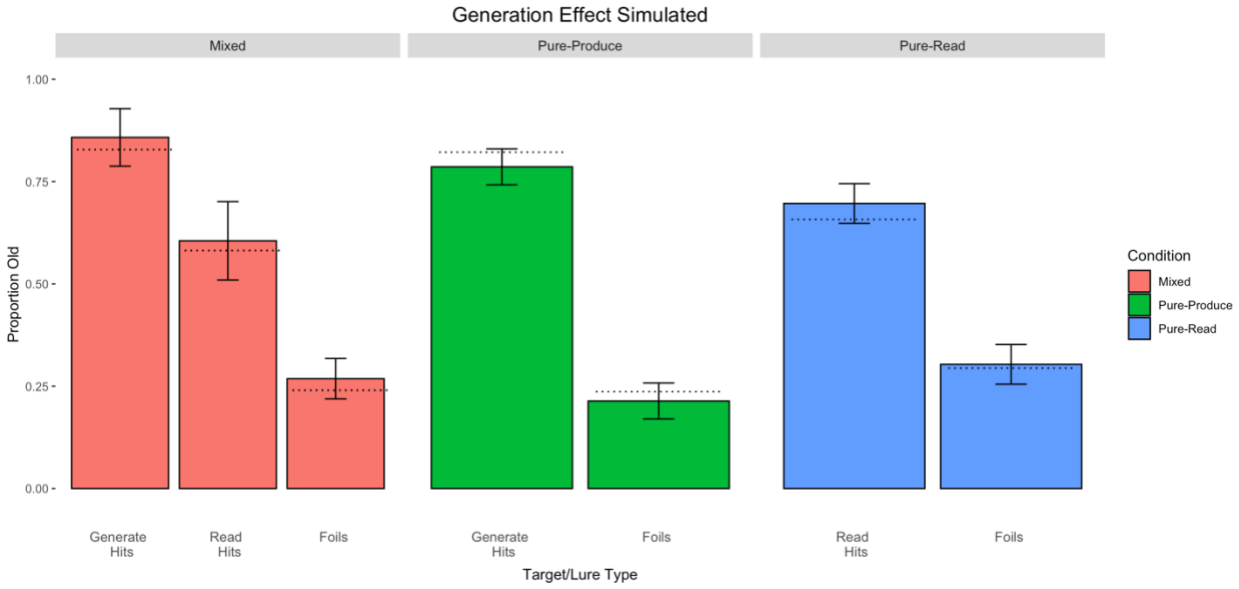


Figure 5.4. Simulation results of the generation effect in Experiment 2, integrating distinctiveness and strength. Note. Mixed-list parameters: $L = 0.056$ for produced targets, $L = 0.040$ for read targets, and the number of extra generation features for produced targets was 250. The decision criterion was set to an unbiased value of 0.501. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

The first set of simulations were conducted in the same way as the first simulation in Experiment 1. Figure 5.4 shows the simulation results of the mixed-list condition on the left-hand side, the pure-generate condition in the middle, and the pure-read condition on the right. By assuming that generated targets are more strongly encoded than read targets and that generated targets were encoded with some extra distinctive information, I was able to generally reproduce the pattern of results that I obtained in Experiment 2 across the three conditions, *RMSE* = 0.0284.

Although the current model fits the data well, the model did not predict the general pattern perfectly. In the mixed-list condition, the model predicted a higher hit rate for generated and read targets, as well as the rate of false alarms for unstudied lures. In the pure-list conditions, the model predicted a lower hit rate for generated items, as well as a lower false alarm rate for unstudied foils. In the pure-read condition, the model predicted a higher hit rate for read items but closely reproduced the rate of false alarms for the new unstudied foils.

As in the empirical results for the production effect, the analysis of the mixed- versus pure-list generation effects did not reveal a significant difference in the size of the generation effect between the two designs, which is a signature of a strength mechanism. Thus, I again sought to explore how a strength only model might account for the data.

Simulation of Experiment 2 – Strength Only

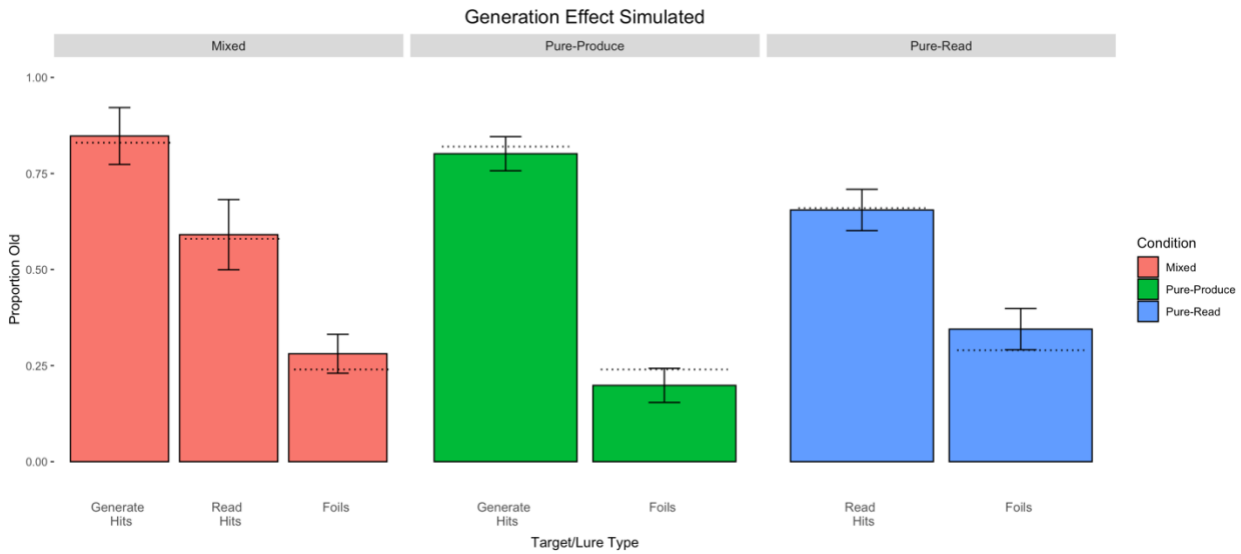


Figure 5.5. Simulation results of the generation effect in Experiment 2, using only a strength mechanism. Parameters: $L = 0.0500$ for generated targets and $L = 0.0254$ for read targets. Distinctiveness was not assumed; no additional features were added. The decision criterion was an unbiased value of 0.505. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

Figure 5.5 shows the simulation results of the mixed-list condition on the left-hand side, the pure-produce condition in the middle, and the pure-read condition on the right. By assuming that produced targets are more strongly encoded than read targets, I was able to closely reproduce the pattern of results that we obtained in Experiment 2 across the three conditions, RMSE = 0.0306.

The model captured the data well; however, it had some minor misses. In the mixed-list condition, the model predicted a slightly higher hit rate for generated targets, read targets, and false alarms for the unstudied foils. In the pure-generate condition, the model predicted a slightly

lower hit and false alarm rate. In the pure-read condition, the model closely matched the hit rate for read targets but predicted a higher false alarm rate for unstudied foils.

Although the strength-based model had a slightly worse fit than the combined model (RMSE = 0.0306 vs. RMSE = 0.0284), given the negligible difference, and the explanatory power of the more parsimonious model (i.e., there was a distinct strength-based signature in the empirical data), the current set of results suggests that the strength model is the better account of the data in hand.

Thus far, the modeling demonstrations of the production and generation effects show that a more parsimonious strength-based model can account for the empirical results found across the mixed- and pure-list conditions. Moreover, the strength-based model of the production and generation effects corroborates the empirical findings found in this dissertation, given that I did not find evidence for distinctiveness in either effect. Next, I sought to account for the results of the testing effect, but with a slightly different model, as necessitated by the slightly different procedure used in the testing effect (i.e., participants had to recall a target from a cue given no part of the target word; participants were shown only a "?" on testing trials).

Simulation of the testing effect

The simulations of the testing effect in Experiment 3 were conducted in a slightly different way than they were for the production and generation effects. Namely, as there was no part of the target that appeared on the intermediate testing trails in the testing effect, both the cue and the target word were encoded into memory.

As before, to represent the first word in each pair, I used a random vector of values drawn from a normal distribution with $M = 0$ and $\sigma = 1/\sqrt{(d)}$, where $d = 300$ dimensions. The second word in each pair was constructed by selectively sampling values from the first word's vector and incorporating new values. This process was governed by a similarity parameter within the model, which determined the likelihood of retaining the original value from the vector or replacing it with a newly drawn value from the same distribution. The newly constructed vector to represent the second target word in each pair was then concatenated onto each first word, resulting in a 600-dimensional word-pair trace. In the current set of simulations, this similarity parameter was set to 0.1.

In the production and generation effect, when extra features are added to memory to account for elaborative processing, these extra features are picked up during the process of iterative retrieval. On the first iteration, only the base features of a word are presented to memory. But on subsequent iterations, the retrieved echo content (which serves as the new probe) starts to include some of the information about the additional production/generation features that were encoded during the study phase. Thus, in the same way, in the testing effect simulations, when memory is probed, only the cue is presented to memory (i.e., the first 300 dimensions, to serve as the first word in each pair). On each subsequent iteration, the retrieved echo content will then start to pick up additional features of the target trace.

Additionally, as there was an interim testing phase for testing trials in the testing effect procedure, the retrieved echo content (along with first word in each pair that served as the probe), was stored to memory after a 2nd iteration. This was done to accurately represent the additional presentation of the first word in each pair, as well as the process of retrieving an item

from memory given the probe. Namely, storing the retrieved echo content is a closer approximation to what is being encoded during a testing trial than merely restoring the probe-target pair, the latter of which would be closer to a restudy manipulation.

For the testing effect simulation, in the pure-test condition, echo content was retrieved iteratively by each block (blocks of four) until all the tested items were processed. The retrieved echo content was then encoded into memory with a learning parameter $L_{practest}$. Given that study blocks consisted of only four items, this parameter was set to 1, assuming the retrieved echo content is encoded into memory with no noise (however in other examinations with larger studied blocks, this parameter could be varied). All tested items (studied and retrieved during the practice testing phase) were then encoded into memory with probability L_{test}.

In the mixed-list condition, echo content was again retrieved iteratively by each block until all tested items were processed (now only 2 of the 4 items in each block) and encoded into memory with parameter $L_{practest}$. All tested items were encoded into memory with probability L_{test} and all read items were encoded into memory with probability L_{read}.

In the pure-read condition, echo content was not computed for any of the items and all studied words were encoded into memory with probability L_{read}. For all three conditions, following the encoding phase, calculation of the probe's familiarity and decision criterion was conducted in the same way as for Experiments 1 and 2.

Simulation of Experiment 3

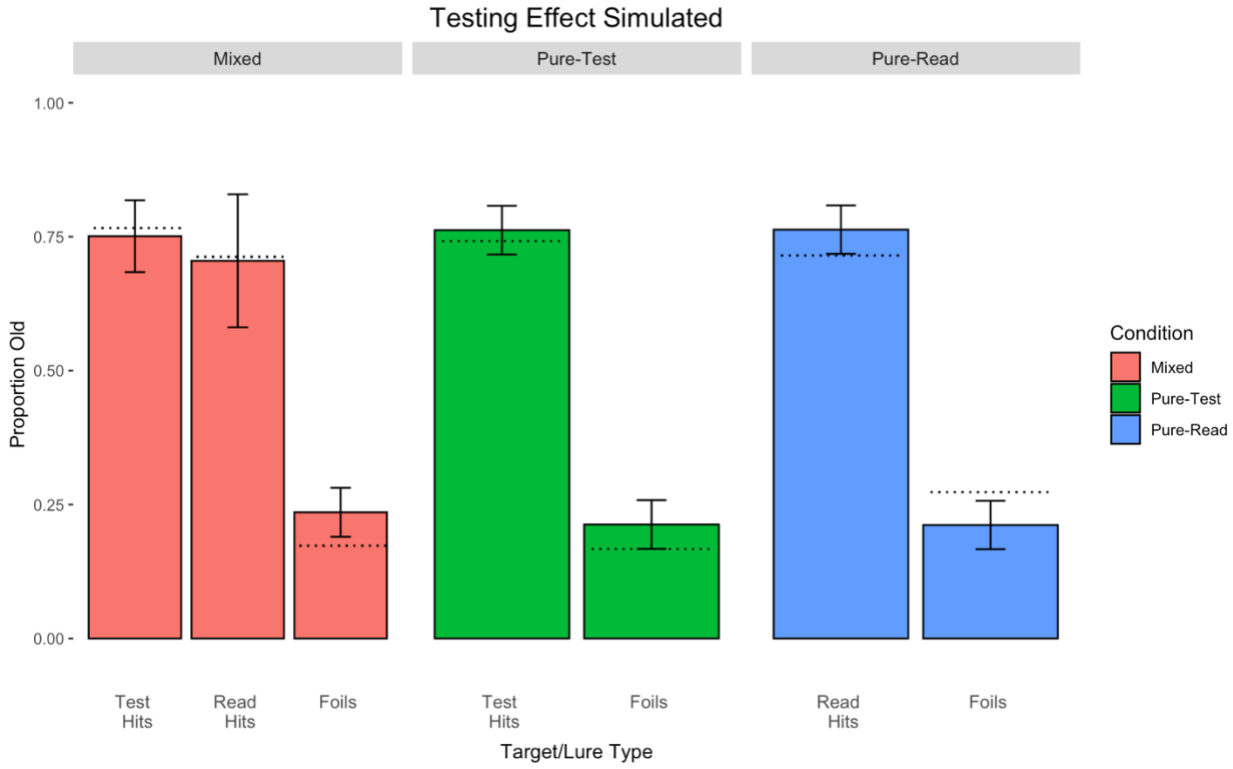


Figure 5.6. Simulation results of the testing effect in Experiment 3, assuming that the retrieved echo content with associated probe is encoded into memory after a 2nd iteration for all tested items. Parameters: $L = 0.0170$ for tested targets and $L = 0.0175$ for read targets. The decision criterion was an unbiased value of 0.48. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

Figure 5.6 shows the simulation results of the mixed-list condition on the left-hand side, the pure-test condition in the middle, and the pure-read condition on the right. By assuming that tested targets are more elaboratively encoded than read targets, we were able to closely reproduce the pattern of results that we obtained in Experiment 3 across the three conditions, RMSE = 0.0318.

The model again captured the data well but had a few misses. In the mixed-list condition, the model predicted a higher false alarm rate to unstudied foils than the empirical data, with the same miss observed in the pure-test condition. In the pure-read condition, the model predicted a slightly higher hit rate for read targets, but a lower false alarm rate for unstudied foils.

Given the different task in the testing effect procedure, strength and distinctiveness were not assessed separately as in the production and generation effects. However, despite the apparent differences between simulations of the production and generation effects and those of the testing effect, the practical impact of these differences is subtle, leading to similar outcomes. Whether additional features are encoded to represent the act of production or generation, or additional traces are added to capture the process of testing and retrieval, both mechanisms strengthen memory representations for these elaboratively encoded items. This strengthening of memory representations improves the chances that stored features will match with the probe, where correct recognition depends on evaluating the similarity between the probe and the encoded information. As in the production and generation effects, a benefit for elaboratively encoded items over read items was observed, where the model closely reproduced this pattern across the three procedures, depending on the task at hand.

Overall, a common modeling framework was able to effectively capture the core patterns observed across the production, generation, and testing effects, demonstrating that enhanced encoding—whether through production or generation or retrieval — leads to stronger memory representations. While the specific mechanisms may differ depending on the task at hand, the fundamental principle remains the same: Elaborative encoding strengthens the overlap between stored features and the retrieval probe, improving recognition accuracy. These findings highlight

the flexibility of instance theory in accounting for different encoding processes and suggest that a unified framework may help explain the benefits of active engagement for memory performance.

Next, a parallel investigation of the production effect and directed forgetting is presented, using the same empirical and modeling approach as in Experiments 1 and 2. The production and directed forgetting effects were examined in parallel investigations to determine the contributions of strength and/or distinctiveness in mixed, pure-produce/remember, and pure-read conditions.

Spear, J., Reid, N., Guitard, D., & Jamieson, R. K.,(2024). Directed forgetting and the production effect: Assessing strength and distinctiveness. *Experimental Psychology, 71* (5), 278–297.

Chapter 6: The production effect and directed forgetting: Assessing strength and distinctiveness

Abstract

The item-based directed-forgetting effect is explained as a difference in how strongly people encode remember-cued over forget-cued targets. In contrast, the production effect is typically explained as a difference in the distinctiveness of the memory of produced over unproduced targets. The procedural alignment of the two effects — directing participants to remember or forget, produce or not — coupled with their different theoretical explanations (i.e., strength versus distinctiveness) presents an opportunity to investigate common versus differential effects of elaborative encoding. This study aims to bridge the gap between these two well-established phenomena by comparing the differences in directed forgetting and the production effect in the context of recognition. Mixed- and pure-list designs were utilized to provide an index of each of these mechanisms in both procedures. Along with a standard production effect and directed forgetting effect in the mixed-list conditions, we found evidence for strength primarily driving results in both procedures. Results are explained using a global matching model of recognition memory, MINERVA 2, by assuming varying levels of encoding strength in relation to task demands.

Keywords: Production effect, directed forgetting, MINERVA 2

The field of cognitive psychology is filled with numerous demonstrations of robust memory effects that give rise to enhanced performance of one class of items over another (Oberauer et al., 2018). Yet with many of these demonstrations, much of the field has been working in silos, often lacking consideration of how other related effects could be working under a common theoretical framework. That is, there has been a shortage of attempts to make connections between different memory effects and how they may commonly or differentially influence memory performance.

Directed forgetting and the production effect are two robust cognitive phenomena that have been extensively studied (MacLeod, 1998; MacLeod et al., 2010; Hall et al., 2021; Saint-Aubin et al., 2021). Directed forgetting refers to the ability of individuals to intentionally forget information that is no longer relevant or necessary, while the production effect refers to the improved memory of information that has been actively produced (e.g., spoken aloud or typed) rather than silently read. Some have examined the connection between the two (e.g., Hourihan & MacLeod, 2008), but the effort toward that end has been limited. To build upon the work of Hourihan and MacLeod, we aim to explore the relationship between the two effects through a comprehensive and in-depth investigation.

In a typical item-method directed forgetting procedure, there are two types of items presented at study: remember-cued (R-cued) and forget-cued (F-cued) items. On both types of trials, participants are typically presented with the target word for some duration, which is then followed by the cue to remember (R) or forget (F) (although see Allen & Vokey, 1998 for an example of simultaneous item and cue presentations). The participants are (falsely) told that items cued to be forgotten will not be tested. A directed forgetting effect is observed when an

individual remembers items that they are instructed to remember better than items that they are instructed to forget.

Traditionally, it has been argued that item-method directed forgetting arises from strength, which gives rise to the better recognition of R-cued items over F-cued items. In particular, the selective rehearsal account (Basden et al., 1993) posits that while the item is presented, the participant engages in maintenance rehearsal to hold the item in working memory while they await the instructional cue. If a cue to remember is then presented, it is posited that participants engage in elaborative rehearsal of the item, whereas if an F-cue is presented, it is posited that participants terminate rehearsal. Therefore, the directed forgetting effect is not due to forgetting per se, but rather the strengthening of the memory trace for R-cued items relative to F-cued items due to the additional elaborative encoding. Although most accounts of directed forgetting agree that there is an encoding advantage for R-cued items, it should be noted that other mechanisms have been proposed as well, such as contextual change, selective search, selective rehearsal, retrieval inhibition, and attentional inhibition (Sahakyan & Kelley, 2002; Epstein, 1969a; 1969b; Bjork et al., 1968; Zacks et al., 1996; Montagliani & Hockley, 2019; Tan et al., 2020; Hourihan & Taylor, 2006; Fawcett & Taylor, 2008; Weiner, 1968; 2010, 2012). Of these accounts, three have been championed: selective rehearsal, retrieval inhibition, and contextual change.

Retrieval inhibition suggests that different mechanisms underlie the list method and item method of directed forgetting. This account proposes that forgetting occurs during the process of retrieval. After the presentation of an F-cue, the items associated with the F-cue are actively inhibited or suppressed. This suppression frees up cognitive resources, allowing more attention

and processing to be dedicated to the to-be-remembered items (Bjork, 1989; Brasden et al., 1993).

The contextual change account posits that directed forgetting results from a shift in internal context following an F-cue, as compared to an R-cue. This internal context can refer to mood, emotional state, or even the physical environment (Sahakyan & Kelley, 2002; Fawcett et al., 2024). In practice, the instruction to forget serves as a signal to enter a new internal context. If the original mental state associated with to-be-remembered items is reinstated before retrieval, this context shift can facilitate memory retrieval. By matching the retrieval context to the remembered items, recall is enhanced for R-cued items. The contextual change account is one of the dominant explanations of list-method directed forgetting, where the instructional cue follows an entire list of items rather than each individual item. Recent studies suggest that context change, or context unbinding, may play a role in item-method directed forgetting as well (Chui et al., 2021; Whitlock et al., 2022), although these studies do not discount the selective rehearsal of R-cued items.

Selective rehearsal suggests that participants intentionally focus on rehearsing the to-be-remembered items while ignoring the to-be-forgotten items (Woodward & Bjork, 1971; MacLeod, 1998; Hourihan & Taylor, 2006). This is done by actively rehearsing R-cued items. Unlike retrieval inhibition, selective rehearsal views forgetting as a more passive process, where unrehearsed F-cued items are simply neglected. Although selective rehearsal is typically associated with the item method, it has been argued that it can offer a unified theory explaining all directed forgetting effects (Sheard & MacLeod, 2005).

In a typical production effect procedure, participants are presented with a list of words, some of which they are instructed to produce (e.g., speak aloud, type, etc.), and some of which they are instructed to read silently. A production effect is observed when participants have better memory for words that they produce than for words that they read silently. The production effect has been examined using multiple modalities, including speaking aloud (Murray, 1965; Hopkins & Edwards, 1972; Conway & Gathercole, 1987; MacLeod et al., 2010), typing (Jamieson & Spear, 2014), mouthing (Forrin et al., 2012), and even imagining (Jamieson & Spear, 2014). In addition, the phenomenon has been observed across a wide range of paradigms, such as immediate recall, reconstruction of order, free recall, and recognition (Saint-Aubin et al., 2021; Cyr et al., 2022; Gionet et al., 2022; MacLeod et al., 2010). The standard account that has been used to explain the advantage for produced items is the distinctiveness account (but see Bodner et al., 2014; Taikh & Bodner, 2016; Bodner et al., 2020). That is, produced items are distinct against a backdrop of nondistinctive read items, where it is the active engagement with produced items that is responsible for the memorial benefit. Therefore, although directed forgetting and the production effect both feature one class of items being better remembered than another due to differential engagement with the items, the accounts for the two effects differ, with directed forgetting being attributed to differences in memory strength and the production effect being attributed to differences in distinctiveness.

With these standard explanations in hand, one possibility is that these two effects, directed forgetting and the production effect, indeed arise for different reasons. However, another possibility is that they arise for similar reasons. If they do arise for the same reasons, the two effects could be explained in the same framework. If they do not arise for the same reasons,

then different theoretical frameworks may be needed. Hence, an investigation of the issue is warranted.

Although investigations of directed forgetting and the production effect together have been scarce in the literature, they are not absent. Hourihan and MacLeod (2008) examined the production effect and directed forgetting together, where common versus differential effects on the two procedures were investigated. The two effects were examined using a 2 (produced vs. read) × 2 (remember vs. forget) design to examine the role of directed forgetting when words were produced or when they were read. The study revealed a directed forgetting effect for words that were read but not for those that were produced. These findings suggest that the benefit of production is robust against instructions to forget, lending support to a distinctiveness account.

The current paper aims to conduct a similar investigation to what Hourihan and MacLeod (2008) did, however, in a slightly different way. Instead of examining the production effect and directed forgetting within-subjects, we chose to examine the two effects in a between-subjects design to assess the contributions of strength and/or distinctiveness in each procedure alone. Moreover, assessing both procedures between-subjects offers a clear test of a formal model to assess similarities and differences between the two procedures, which will be illuminated in the work that follows.

To disentangle the two accounts of strength and distinctiveness, often mixed- and pure-list examinations have been utilized (Bodner et al., 2016; Zhou & MacLeod, 2021). In a mixed-list design, sometimes termed a within-subjects production effect, participants encounter a combination of produced and read items. In contrast, in a pure-list design, commonly known as a between-subjects production effect, all items are either produced or read, with no mix of

presentation methods. The size of the production effect observed in the mixed- vs. the pure-list design is used as an index of the amount of distinctiveness and/or strength that is contributing to the production effect. The signature of a strength effect is an equal benefit for produced items across mixed- and pure-list designs, whereas the signature of a distinctiveness effect is a larger benefit for produced items in a mixed-list design than in a pure-list design.

Given that a distinctiveness effect is observed in a mixed-list design vs. a pure-list design, we will additionally conduct a pure-list counterpart. Moreover, to complete the full design, we will also run the pure-list counterpart for directed forgetting. If both effects are strength-based, it will be observed that the benefit for produced or R-cued items is equal across both mixed- and pure-list designs. However, if both effects are distinctiveness-based, it will be observed that there is a larger benefit for produced/R-cued items in the mixed-list designs than in the pure-list designs. Thus, if both effects arise for the same reasons, we expect that the results across both paradigms will be consistent with either a strength- or distinctiveness-based account. If they arise for different reasons, consistent with previous work, we should observe distinctiveness-based results in the production effect (larger effect for mixed-lists than pure lists) and strength-based results in directed forgetting (similar sized effects in mixed- and pure-lists).

Furthermore, in terms of the pattern of findings with the production effect in mixed-list designs and other associated effects (e.g., the generation effect), it has been unclear whether it is a cost to unproduced items that drives the production effect or if it is a benefit to produced items (Begg & Snider, 1987; MacLeod et al., 2010). To answer the question, often the approach is again to run a pure-list counterpart and then compare the hit rates for read and produced items between the two designs. Evidence in favor of a benefit to produced items is acquired if the hit

rate for the produced items is larger in the mixed-list design than in the pure-list design. Conversely, evidence in favor of a cost to unproduced items (i.e., the lazy reading hypothesis; Bodner et al., 2014) is acquired if the hit rate for read items is lower in the mixed-list design than in the pure-list design. Bodner et al. (2014) employed these mixed- and pure-list conditions and found that the hit rate for read items was lower in the mixed-list design than in the pure-list design, suggesting that the production effect is driven by a cost to unproduced items. Additionally, the meta-analysis conducted in Bodner et al. showed that there was a cost for silent items with little benefit for produced items. Forrin, Groot, and MacLeod (2016) similarly found a larger production effect in mixed-lists than in pure-lists. However, it is unclear whether these patterns of findings hold when other modalities of production are used (e.g. typing).

To date, there have been a few approaches to modeling the production effect, including REM, the Revised Feature Model (RFM), attentional subsetting theory (AST), and MINERVA 2 (Kelly et al., 2022; Saint-Aubin et al., 2021; Caplan & Guitard, 2024; & Jamieson et al., 2016). Common to all of these models is the addition of features to a memory trace to account for the added benefit of production. Additionally, directed forgetting has been accounted for using a strength-based version of MINERVA 2 (Reid & Jamieson, 2022; Reid et al., 2023). Thus, the current paper utilizes the MINERVA 2 model, as this model has been used to successfully model the production effect and directed forgetting.

In the modeling framework of the production effect proposed by Jamieson et al. (2016), enhanced memory performance for produced items can be accounted for in one of two ways. In a strength-based account, it is assumed that produced items are better encoded into memory, with

more intact features. In a distinctiveness-based account, it is assumed that produced items are elaboratively encoded, such that these items are encoded with more unique, distinct features to distinguish themselves from unproduced or non-elaboratively encoded items. In practice, the model affords memory a global familiarity signal, whereby the system is reminded of the act of production with an iterative retrieval function.

Thus, a theoretical question is: do the parallel explanations for strength versus distinctiveness in directed forgetting and the production effect necessitate their complete independence, or do they potentially operate in tandem, with certain circumstances favoring one or the other? Moreover, can results from both the production effect and directed forgetting, two effects of elaborative processing be accounted for in a single model? Having a comprehensive model that can account for multiple memory effects can help us escape a siloed approach to memory research and instead consider in tandem different memory effects. To answer the question, we conducted experiments where both mixed- and pure-lists were used across both the production effect and directed forgetting.

Experiment 1

The goal of Experiment 1 was to examine the production effect in both mixed- and pure-lists using uncategorized words. The modality of production was a production-by-typing task (Bodner et al., 2016; Forrin et al., 2012; Jamieson & Spear, 2014; Kelly et al., 2024). We chose a production by typing procedure as this modality is both understudied and provides a convenient way to collect data online. However, more importantly, it is unclear whether the effect should be larger in mixed relative to pure-lists, as the features encoded in this task can be considered less rich than in spoken production. The standard design was adopted to provide a clean test of the

model and to subsequently compare the results from directed forgetting in Experiment 2. Moreover, the inclusion of both mixed- and the pure-lists allowed for the assessment of strength and/or distinctiveness mechanisms. Results are displayed in Figure 6.

Method

Participants. According to a two-tailed 80% power analysis conducted with G*Power 3.1 (Faul et al., 2009) a minimum of 15 participants were needed, with alpha set to 0.05. We used an effect size of $d = 0.81$, obtained from Jamieson and Spear (2014). However, as the current design involved three different conditions (a mixed-list, pure-produce, and a pure-read condition) we sought to exceed this amount by more than threefold, as we included an additional procedural change of collecting our data online. As such**,** data was collected from a total of 128 participants. From there, data from 120 participants (71 female, 49 male) were included in the analysis. Participants were recruited online from the University of Manitoba SONA psychology participant pool and received one credit towards completion of their Introduction to Psychology course. The mean age of participants was 20.75 years (range = 17–44, SD = 4.9 years). Participants were excluded based on not complying with the instruction manipulation in the production condition (e.g., not typing at least 80% of the time on produce trials, or typing on read trials) or if they reported doing something else during the experiment (e.g., some participants reported that they were in class, watching TV, etc.). Of the 8 participants that were excluded, 4 self-reported that they were distracted while doing the experiment, and 4 participants did not comply with the experimental instructions at least 80% of the time. Data was collected to ensure that there were at least 40 participants in each of 3 conditions (the mixed, pure-produce, and pure-read conditions).

Materials. Materials were 120 words taken from MacDonald & MacLeod (1998) and are listed in Appendix A. From these 120 words, 80 words were randomly selected for each participant: 20 to serve as produced targets, 20 to serve as read targets, and 40 to serve as new unstudied lures. In addition to the 40 study words, there were two buffer items at the beginning and end of the study list to mitigate contamination from primacy and recency effects on performance.

Procedure. Participants were tested online using jsPsych version 6.3.1, a JavaScript library for running behavioral experiments via a web browser (https://jspsych.org; de Leeuw, 2015). The experiment was hosted online using GitHub Pages. When the study commenced, participants were randomly assigned to a mixed, pure-produce, or pure-read condition.

All participants provided informed consent, which was followed by instructions. The instructions asked participants to turn off any background audio and to remain in full screen for the duration of the experiment. After this, they were presented with instructions for the experiment. The production task involved typing the word that was presented. Participants were instructed to type words that appeared in green and to silently read words that appeared in red. Also, embedded within these instructions was an attention check. Participants were told that when asked to "provide an answer" on the following screen, they should provide the answer to "2+2 = __". If the participant typed in anything but the number "4", the experiment looped back to the instruction screen until the correct answer was provided.

Before the start of the main experiment, participants were given a short practice phase that included "produce" and "read" trials that provided feedback on their performance. The practice study phase consisted of 16 trials, 8 of which were "produce" trials, and 8 of which were "read" trials, presented in random order. For the "produce" trials, if participants typed the word

correctly, they were presented with feedback that read, “Correct! You typed correctly!”. If participants typed anything except the exact word, they were presented with feedback that read, “Incorrect. You must type the exact word." For the “read” trials, if participants did not type anything, they were presented with feedback that read, “Correct! Thank you for reading!”. Conversely, if the participant typed anything on “read” trials, they were presented with feedback that read “Incorrect. Please do not type.”. This was followed by a practice test phase in which participants also received feedback on their recognition performance.

Once the study phase commenced, study words were first presented in black for 1,000 ms, after which each word immediately turned red or green and remained on the screen for 3,000 ms. Each study word appeared one at a time, in easy-to-read 48px font on a white background. After each study word, the screen was cleared for 500 ms, after which the next study word appeared. Other than the two buffer items at the beginning and end of each study list, all 40 study items were presented in random order.

After the study phase was over, participants were presented with test instructions, which informed them to press ‘y’ if they recognized a word and ‘n’ if they did not. During the test phase, participants were tested on their recognition of all previously studied words (other than the buffer items), along with an equal number of unstudied lures. Each test word appeared in the center of the screen, in black 48px font, on a white background, until the participant pressed ‘y’ or ‘n’. Once the participant responded, the screen was cleared for 500 ms, after which the next test word appeared. This procedure was repeated until all 80 test words were presented in random order. After the test phase was finished, participants were presented with a demographic

questionnaire, followed by a question asking if they were doing anything else while completing the experiment, and finally a debriefing screen.

Results and Discussion

All analyses were conducted in R (R Development Core Team, 2022; version 4.2.1, 2022). The results of the mixed-list (leftmost column), pure-produce (middle column), and the pure-read (rightmost column) conditions are displayed in Figure 1.

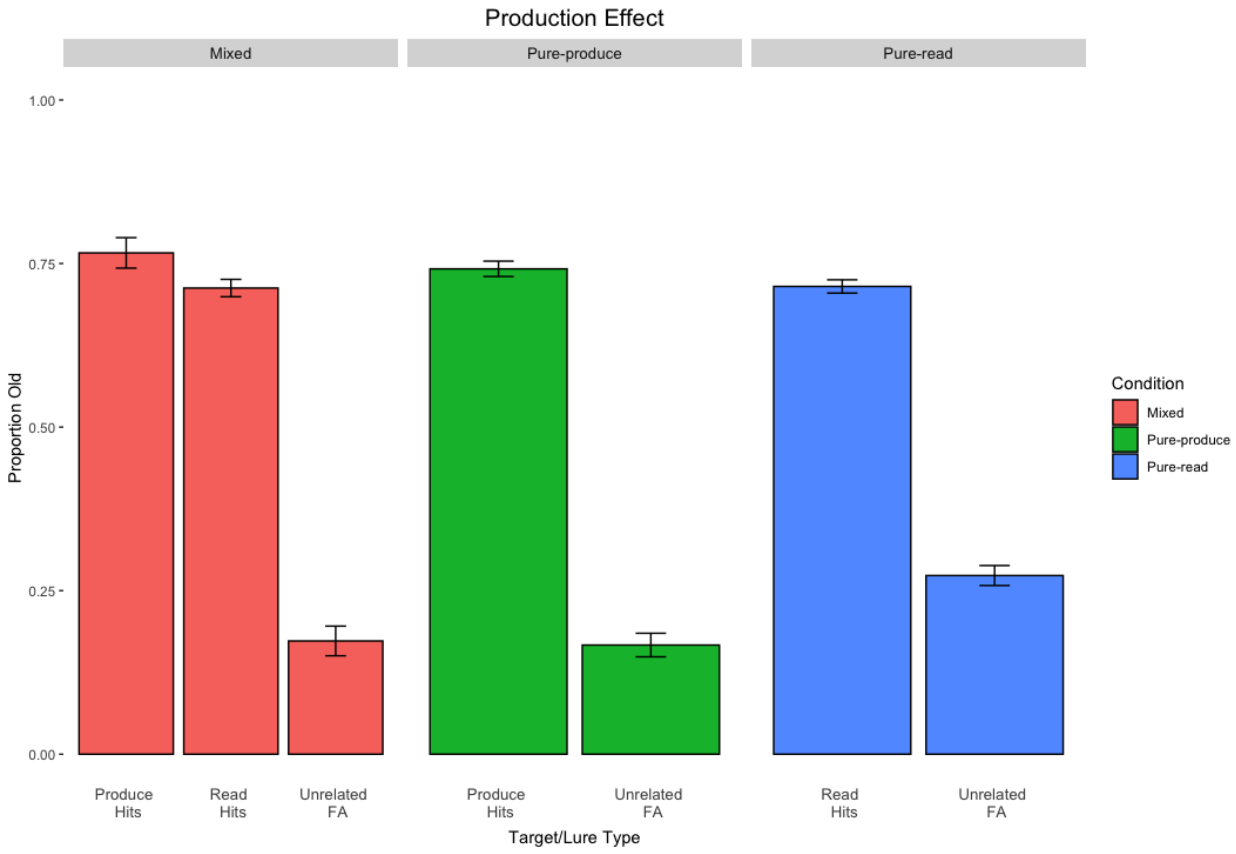


Figure 6.1. Results from Experiment 1. Error bars represent standard errors of the means.

We began with an analysis of the mixed-list results. The first comparison revealed that participants responded "yes" significantly more to old items than new ones, $t(39) = 19.26$, $p < 0.001$, $d = 0.98$. Next, the comparison between produced vs. read items revealed a significant production effect, $t(39) = 2.14$, $p = 0.04$, $d = 0.09$, with a production advantage of 6%. The pure-list production effect was not significant using the Welch correction for unequal variances, $t(74.92) = 0.86$, $p = 0.39$, $d = 0.19$, with a production advantage of only 2%.

Strength or distinctiveness? If it is strength driving the results in the production effect, then the size of the production effect in the mixed- and pure-list conditions should be of similar size.

However, if it is primarily due to distinctiveness, then the size of the production effect ought to be significantly larger in the mixed-list condition compared to the pure-list condition. To assess the role of distinctiveness in the mixed-list condition, we next applied an Erlebacher analysis.

The current examination involved three different between-subjects conditions: a mixed-list condition, a pure-produce condition, and a pure-read condition. Production (produced vs. read) in the mixed-list is a within-subjects condition, whereas it is a between-subjects manipulation across the two pure-list conditions. As such, to properly assess the difference in magnitude of the mixed-list and pure-list production effects, the Erlebacher method of analysis was used, a technique that circumvents traditional methods (Erlebacher, 1977). This method ensures an unbiased estimate of the design type and its interaction with the independent variable. R code that was developed by Merritt et al. (2014) was used for this analysis.

A 2 (item type: produced vs. read) × 2 (design: mixed vs. pure) Erlebacher ANOVA was conducted on only hits. Although there was only a 4% difference between the mixed- vs. pure-list production effects, the results revealed a main effect of production, $F(1, 78) = 4.05$, $p < 0.05$, $\eta^2 = 0.02$. In contrast, there was no main effect of design type, $F(1, 78) = 0.14$, $p = 0.71$, $\eta^2 = 0.001$, nor an interaction between design type and item type, $F(1, 78) = 0.45$, $p = 0.50$, , $\eta^2 = 0.002$. Thus, the mixed-list production effect was not significantly larger than the pure-list counterpart, failing to lend support to a distinctiveness account.

A cost or a benefit? Lastly, in terms of costs and benefits of production, it is assumed that a benefit to produced items is observed if the hit rate for produced items in the mixed-list condition is larger than in the pure-produce condition. Conversely, it is assumed that there is a cost to

unproduced items if the hit rate for read items in the mixed-list condition is lower than the hit rate for read items in the pure-read condition.

The hit rate for produced items in the mixed-list condition was 77%, with a 74% hit rate in the pure-produce counterpart (a 3% benefit for the mixed-list). Conversely, the hit rate for read items in the mixed-list vs. the pure-read list only differed by 1% (71% vs. 72%, respectively). Thus, results from Experiment 1 yielded support for a slight benefit of production, although not a strong one.

Experiment 2

The purpose of Experiment 2 was to assess the directed forgetting procedure using the same materials and design as before. This consistency allowed for a direct comparison of the production and directed forgetting effects. Both mixed- and pure-lists were utilized to evaluate the roles of strength and distinctiveness in directed forgetting. The standard design of presenting the cue to remember or forget after item presentation, was once again adopted to ensure a clear test of the model.

Method

Participants. Participants were recruited in the same way as in Experiment 1. Data was collected from a total of 131 participants. From there, data from 120 participants (71 female, 49 male) were included in the analysis. Eleven participants were excluded based on self-reporting that they were distracted while completing the experiment. The mean age of participants was 20.50 years (range = 17–43, SD = 3.8 years). Data was collected to ensure that there were 40 usable participants in each of the mixed, pure-remember, and pure-forget conditions.

Materials. The materials were the same as in Experiment 1.

Procedure. The procedure was identical to Experiment 1, with the difference that participants were instructed to remember the words presented in green and forget the words presented in red (no typing instruction was given). Moreover, no typing feedback was provided during the practice phase, as no typing instruction was given for the directed forgetting procedure. As before, participants were randomly assigned to one of the three between-subjects conditions: mixed, pure-remember, or pure-read. Participants were not explicitly informed of what condition they were assigned to. Lastly, the experimental test instructions in the directed forgetting procedure (presented after the study phase) additionally included a sentence that instructed participants to identify any items they remembered as old, regardless of the remember and forget cues.

Results and discussion

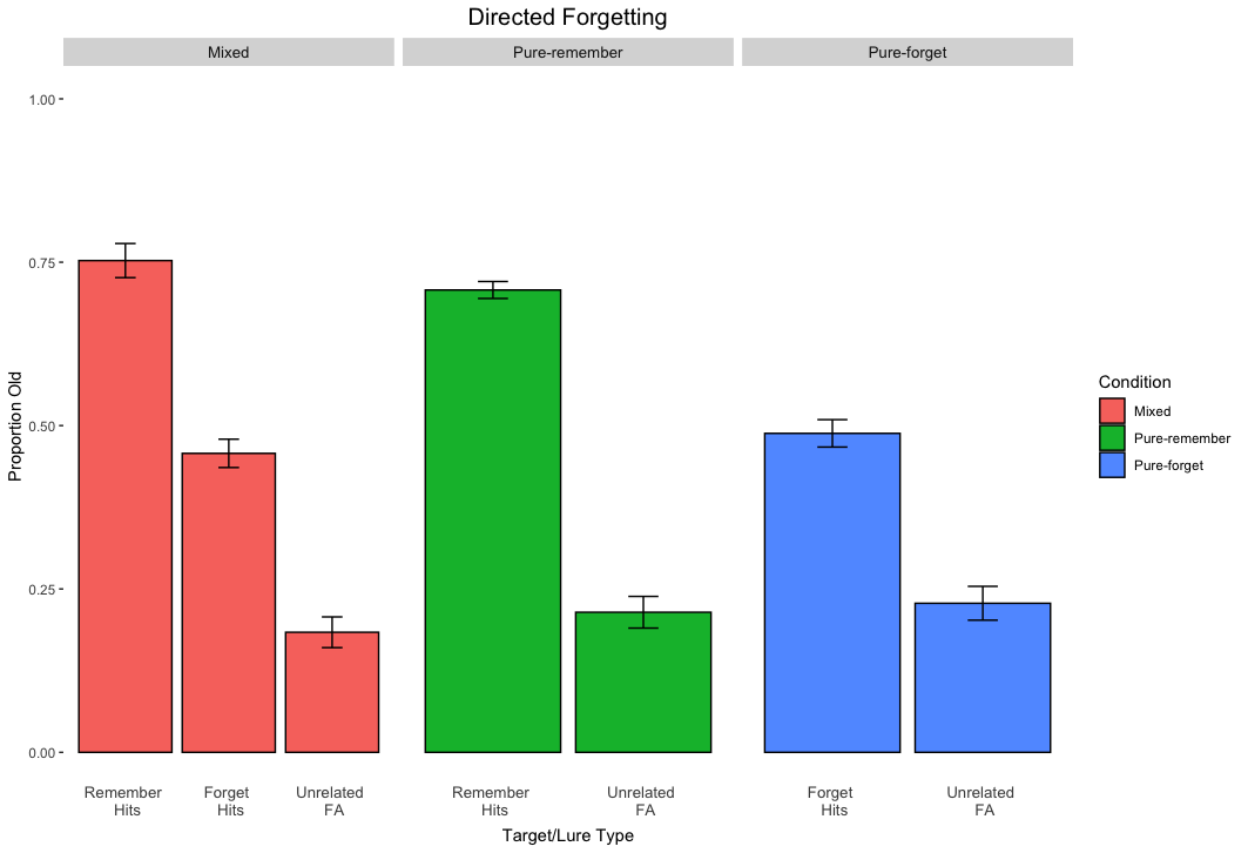


Figure 6.2. Results from Experiment 2. Error bars represent standard errors of the means.

Figure 6 displays the directed forgetting results of the mixed-list condition in the first column, the pure-remember condition in the second column, and the pure-forget condition in the third column. Proportion old is on the y-axis, and item-type is on the x-axis.

The first comparison revealed that, unsurprisingly, participants responded "yes" to items that they studied significantly more than to items that they did not (i.e., old vs. new items), $t(39) = 15.16$, $p < 0.001$, $d = 072$. More critically, a mixed-list directed forgetting effect was observed, $t(39) = 6.71$, $p < 0.001$, $d = 0.45$, with an R-cue advantage of 29%. A pure-list directed forgetting

effect was also observed using the Welch correction for unequal variances, $t(63.63) = 4.69$, $p < 0.001$, $d = 1.05$, with a 22% advantage for R-cued items.

Strength or distinctiveness? Although there was a 7% difference between the mixed- vs. pure-list directed forgetting effects, the 2 (item type: produced vs. read) × 2 (design: mixed vs. pure) Erlebacher ANOVA conducted on only hits revealed a main effect of cue type, $F(1, 78) = 64.19$, $p < 0.001$, , $\eta^2 = 0.26$, but no main effect of design type, $F(1, 78) = 0.04$, $p = 0.84$, , $\eta^2 < 0.001$, nor an interaction between design type and item type, $F(1, 78) = 1.39$, $p = 0.24$, $\eta^2 = 0.006$. As in Experiment 1, the mixed-list directed forgetting effect was not significantly larger than the pure-list counterpart, again failing to support a distinctiveness account in directed forgetting.

A cost or a benefit? As in Experiment 1, to assess whether there is a cost to F-cued items or a benefit to R-cued items, a comparison of the hit rates for each item type can be done across the mixed- and pure-list conditions.

Experiment 2 yielded mixed results. The hit rate for R-cued items in the mixed-list condition was 75%, whereas the hit rate for the pure-remember condition was 71% (a 4% benefit for the mixed-list). However, the hit rate for F-cued items in the mixed-list was 46%, whereas it was 49% in the pure-forget condition, indicating a 3% difference in favor of the pure-forget condition. Thus, it appears there was a benefit (higher hit rate for R-cued items in the mixed- than pure-remember condition) and a cost (lower hit rate for F-cued items in the mixed- than pure-forget condition) associated with R-cued and F-cued items in the directed forgetting procedure, of nearly the same magnitude.

Experiment 1 and Experiment 2. As there was a larger directed forgetting effect than a production effect, an analysis of the two types of targets across the two procedures in the mixed-

list was conducted. A mixed ANOVA with item type (produced vs. read) as the within-subjects factor and experiment (directed forgetting vs. production effect) as the between-subjects factor was used to assess this difference. The analysis confirmed that this difference was statistically significant, with a main effect of experiment $F(1, 78) = 13.35$, $p < 0.001$, $\eta^2_p = 0.15$. There was also a main effect of item type $F(1, 78) = 47.48$, $p < 0.001$, $\eta^2_p = 0.38$, and a significant interaction between Experiment and item type $F(1, 78) = 22.72$, $p < 0.001$, $\eta^2_p = 0.23$, indicating that the elaborative encoding benefit (or cost) in directed forgetting (29%) exceeded that of the production effect procedure (6%).

Experiment 3

Given that the cue to produce or read in a production effect procedure is typically presented concurrently with stimulus presentation, it could be argued that our production effect procedure in Experiment 1 deviated from the standard methodology. Therefore, we conducted Experiment 3 to examine the production effect in a conventional manner. Additionally, we sought to explore any differences between presenting the instruction to produce or read concurrently with the stimulus versus presenting it shortly thereafter, as was done in Experiment 1.

Participants. Participants were recruited in the same way as in Experiment 1. Data was collected from a total of 44 participants. From there, data from 40 participants (25 female and 15 male) were included in the analysis. Four participants were excluded based on self-reporting that they were distracted while completing the experiment. The mean age of participants was 25.33 years (range = 17-47, SD = 7.9 years).

Materials. The materials were the same as in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1, except that the instruction to "produce" or "read" was presented concurrently with the stimulus. Figure 3 illustrates both procedures, with the concurrent stimulus presentation procedure shown on the left-hand side.

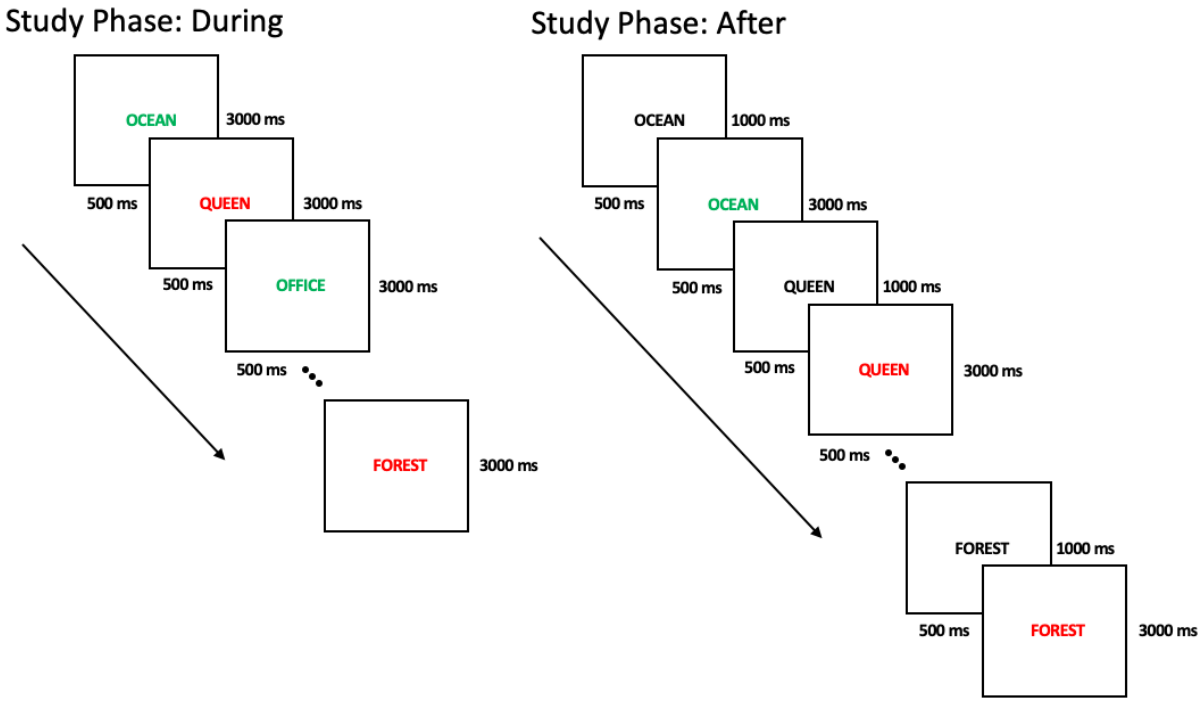


Figure 6.3. An example of the two study procedures. The "after" study procedure was used in Experiments 1 and 2, and the "during" study procedure, was used in Experiment 3. The duration of stimulus presentation is displayed on the right, whereas ISI is displayed on the left.

Results and discussion

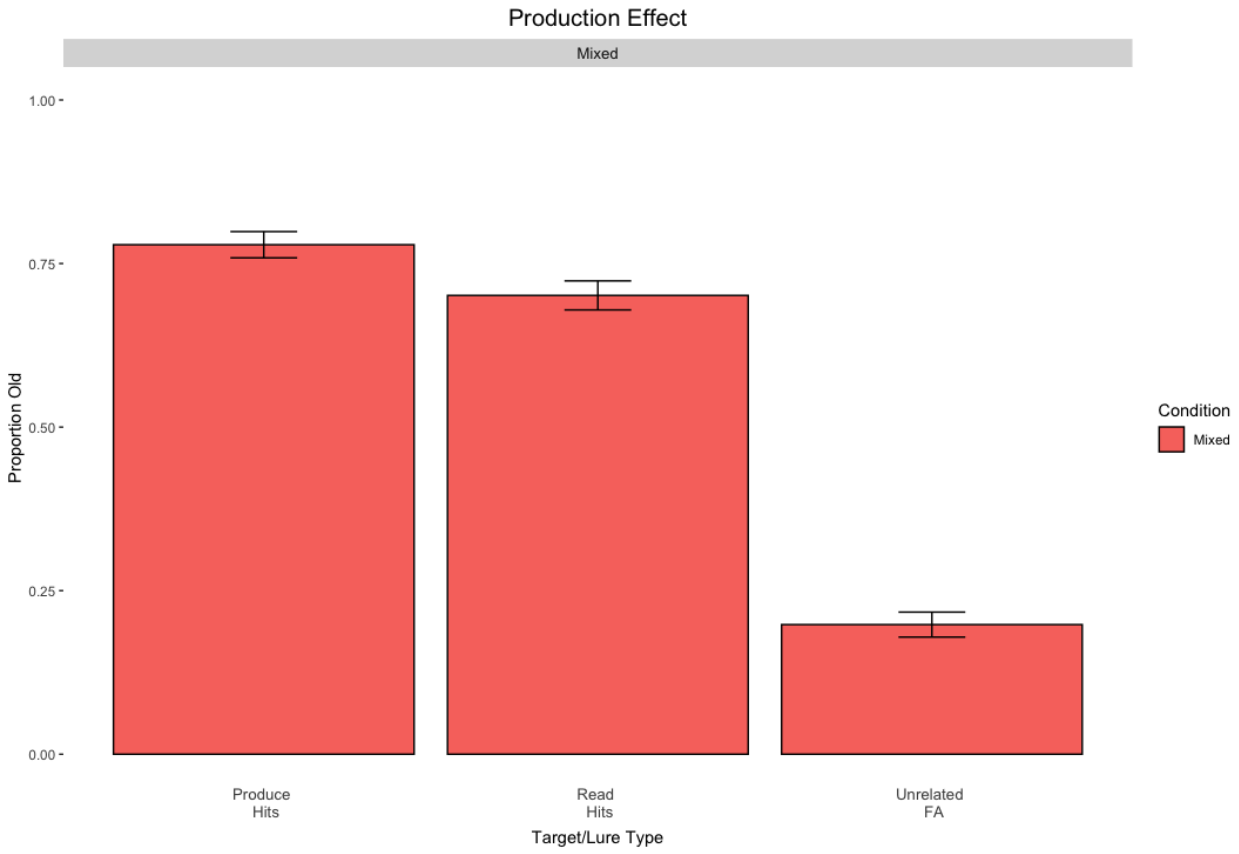


Figure 6.4. Results from Experiment 3. Error bars represent standard errors of the means.

Figure 6.4 displays the results of the mixed-list production effect procedure when the cue to "produce" or "read" was presented concurrently with the stimulus presentation.

Analyses revealed that participants responded "yes" significantly more to old items than new, $t(39) = 15.88$, $p < 0.001$, $d = 0.93$. More importantly, there was a significant production effect, $t(39) = 3.46$, $p = 0.001$, $d = 0.14$, with a production advantage of 8%.

Since the purpose of Experiment 3 was to solely assess the difference between cue presentations (during vs. after instruction presentation), the pure-list counterpart was not

conducted. However, to determine if the size of the production effect differed between the two cue presentation timings, an analysis was performed on the targets from Experiment 1 and Experiment 3. A 2 (cue presentation: during vs. after) × 2 (item type: produce vs. read) ANOVA revealed a main effect of item type, $F(1, 78) = 15.21$, $p < 0.001$, $\boldsymbol{\eta}^2_p = 0.16$, but no main effect of cue presentation timing, $F(1, 78) < 0.01$, $p = 0.99$, $\boldsymbol{\eta}^2_p < 0.001$, nor an interaction between cue presentation timing and item type, $F(1, 78) = 0.50$, $p = 0.48$, $\boldsymbol{\eta}^2_p = 0.01$. Thus, there was no significant difference between presenting the cue to "produce" or "read" either concurrently or shortly thereafter.

Summary. The production effect and directed forgetting effect are two types of elaborative processing study procedures that lend a benefit to one class of items over another. Empirical results thus far indicate that there is a sizable difference in the magnitude of this benefit in favor of directed forgetting across the two procedures. Moreover, whether the cue to "produce" or "read" is presented concurrently with stimulus presentation or shortly thereafter does not seem to affect the production benefit in the current examination.

The next major aim of this paper was to account for both the production effect and directed forgetting using a computational model of human memory, MINERVA 2. As such, we implemented the elaborative processing account in MINERVA 2 to model the results of Experiments 1–3.

The Model

To date, there have been four main approaches to modeling the production effect: REM, the Revised Feature Model (RFM), AST, and MINERVA 2 (Kelly et al., 2022; Saint-Aubin et al., 2021; Caplan & Guitard, 2024; Jamieson et al., 2016). Common to all of these models is the

addition of features to account for the added benefit of production. The RFM is an account of recall and not recognition, whereas REM and MINERVA 2 are accounts of recognition. In this paper, we use the MINERVA 2 model of recognition memory, which has also been used to model the directed forgetting effect (Reid & Jamieson, 2022; Reid et al., 2023).

MINERVA 2 (Hintzman, 1984, 1986, 1988) belongs to a class of computational models of human memory known as global matching models. MINERVA 2 accounts for memory storage, retrieval, and decision. The model has had a wide range of successes in several cognitive domains, including reaction time (Jamieson & Mewhort, 2009), false recognition (Arndt & Hirshman, 1998), associative learning (Jamieson et al., 2012), decision-making (Dougherty et al., 1999), sentence memory (Reid & Jamieson, 2023), lexical disambiguation (Jamieson et al., 2018), serial recall (Guitard et al., 2025), and implicit rule learning with semantics (Chubala et al., 2016). The model assumes that memory is a matrix, where each row represents an item and columns represent the features of the items. When words are encoded, they are encoded to a unique row in the memory matrix, with some degree of noise. This noise is introduced by the parameter L, which is the learning parameter in the model. L can also be considered the strength with which a trace is encoded into memory.

Specific to the production effect, there are a few differences from the standard instantiation of the model. Firstly, the benefit of production to memory can be accounted for in one of two ways: with a distinctiveness-based mechanism or a strength-based mechanism. In both cases, a word is first represented by a unique vector, where some number of base-features represent the word.

A distinctiveness-based mechanism works by adding some number of extra features to all items. For produced items, these extra features contain additional information, whereas for unproduced items, the extra features contain no additional information (these features are set to 0). Secondly, the model's retrieval process works in an iterative fashion, akin to the deblurring process utilized by Hintzman (1986). Functionally, the test word is used as a retrieval probe to retrieve an echo three times. On the first iteration, only the base features are included in the probe. For the second iteration, if the word was produced, the retrieved echo content (which serves as the new probe) will now include some of the information about the extra produced features that were encoded during the study phase, thus retrieving new unique features to account for distinctiveness. For the third iteration, the probe is further refined, incorporating both the base and any additional features associated with production, enhancing the retrieval accuracy. This iterative process increases the likelihood of correctly identifying the produced items due to retrieval and use of the enriched feature set. Conversely, for unproduced items, the probe is similarly submitted to memory, and the same iterative retrieval process is used, but the echoes do not strongly pick up any additional production features since the traces in memory they match most strongly to do not include these features. Therefore, the probe remains based primarily on the base features over the retrieval iterations, leading to a less distinct and weaker retrieval signal. After the third iteration, a global familiarity signal known as echo intensity is calculated, which is based on the sum of all activations in memory. Activations are calculated based on the similarity of the probe (echo retrieved after the third iteration) to items in memory. The probes for produced items elicit a stronger familiarity signal because they match their corresponding traces in memory on both base and production features, whereas the unproduced probes match only on base features.

In contrast, in a strength-based model of the production effect, there are no extra features added to memory. Instead, it is assumed that produced items are encoded more strongly in memory than read items by varying the parameter L in the model for each class of item. This is the same way that Reid and Jamieson (2022) modeled the item-method directed forgetting effect. By assuming that R-cued items are more strongly encoded with more intact features than F-cued items, Reid and Jamieson were able to demonstrate the typical directed forgetting effect found in veridical recognition, as well as a parallel directed forgetting effect that occurs in false recognition for related lures (see Montagliani & Hockley, 2019; Marche et al., 2005; Reid et al., 2023).

Given that there is a mixture of findings found in the literature, and that strength and distinctiveness likely work together in tandem given these findings, we present a model that incorporates both strength and distinctiveness mechanisms. First, to implement strength, as outlined above, we can assume that elaboratively studied items are encoded into memory with more intact features than nonelaborative items, by varying the parameter L for each class of item ($L_P > L_R$ and $L_R > L_F$). To implement distinctiveness, we assume that distinctive items (e.g., produced items) have additional non-zero features to the base features whereas for non-distinctive items, the additional features have values of zero. However, because the items are not produced at test (see Jamieson et al., 2016), it is assumed that the initial probe does not contain the distinctive features, but that these features must be retrieved from memory through an iterative retrieval process. Retrieval works in the following fashion: when a probe is presented to memory, activation is similarity-based and is calculated on a feature-to-feature basis in parallel. These activated traces are represented in an echo, where an echo is made up of two key

properties: echo content and echo intensity. Echo content, *c*, is a vector comprised of the sum of all the traces that are activated:

$$c_{j=\sum_{i=1}^{m} a_i \times M_{ij}} \qquad \{for\ each\ j = 1 \ldots n\}$$

where c_j is the j^{th} element of the echo, a_i is the activation for the i^{th} trace in memory, M_{ij} is the j^{th} element of the i^{th} trace in memory, *m* is the number of traces in memory, and *n* is the number of elements in each vector. The activation for each trace in memory is computed as the cosine similarity between the probe and that trace raised to the exponent of three (Hintzman, 1986, 1988). The retrieval process works in an iterative fashion, such that the test word is used as a retrieval probe to retrieve an echo three times. On the first iteration, only the base features are included in the probe. On the second iteration, if the word was produced, the retrieved echo content (which serves as the new probe) will now include some of the information about the extra produced features that were encoded during the study phase.

Following three iterations, the probe's familiarity is computed as an echo intensity, which is the sum of the activation elicited by the probe:

$$f = \sum_{i=1}^{i=m} \left(\frac{\sum_{j=1}^{j=d} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=d} p_j^2} \sqrt{\sum_{j=1}^{j=d} M_{ij}^2}} \right)^3$$

where a familiarity (*f*) value is calculated based on the probe's similarity to all traces in memory, *M*, where specifically, a cosine similarity calculation is used, p_j is feature *j* of the probe, M_{ij} is feature *j* of trace *i* in memory, *m* is the number of memory traces, and *d* is the dimensionality of these traces. This similarity calculation is then converted to activation by raising it to the exponent of three, enhancing the signal-to-noise ratio of all calculated familiarity values. Finally,

all these similarities are summed to yield an overall familiarity index, *f*, also called an echo intensity in other uses of the MINERVA model. The additional nonzero features encoded at study along with the iterative retrieval process to retrieve those features is how distinctiveness is represented and used in the model.

The model simulates decision-making by using all the calculated familiarity values (for both old and new words) to determine a criterion based on a chosen percentile that best fits the data. For example, if the decision criterion is set to the 55th percentile, that would mean that the top 45% of the most familiar echo intensities would be classified as "OLD" (a decision criterion of 0.45), while the remaining echo intensities would be classified as "NEW." In this example, this value represents a slightly conservative criterion, indicating that the model is marginally more inclined to classify items as "NEW", being that the criterion is just above the median.

Simulation Results

The standard design of the production effect and directed forgetting procedures examined here were used to provide an articulate basis for the model and the comparison of the two procedures. We report simulation results from the production effect findings in Experiments 1 and 3, and the directed forgetting findings in Experiment 2. In each simulation, 1000 independent simulations were conducted.

Word representations. Classically, word representations in the MINERVA 2 framework have been discrete random representations (e.g., a vector of randomly sampled +1s and -1s). However, MINERVA 2 allows for other types of representations, such as engineered representations (e.g., Arndt & Hirshman, 1998), or those derived from models of natural language processing (see Chubala et al., 2016; Reid & Jamieson, 2023; Chang & Johns, 2023; for demonstrations). Here,

we use orthogonal continuous representations drawn from a normal distribution, with $M = 0$ and a $\sigma = 1/\sqrt{(d)}$, where d is the dimensionality of each word vector (see Murdock, 1982; Jamieson & Hauri, 2012; Jones & Mewhort, 2007). In the simulations that follow, d was set to 300 for the base features.

Simulation of Experiment 1

We used a combined version of MINERVA 2, where some additional features were added to account for distinctiveness. The model also assumes that "produce" targets are encoded more strongly into memory with a higher value of L than the words that participants were instructed to "read". Familiarity was then computed for all 80 test items: 40 old and 40 new. Then, these familiarity values were converted to an old/new decision by comparing them to a chosen criterion that best fits the data. In the simulation of Experiment 1 and in the simulations that follow, we iteratively fit our data by varying values of L and number of production features.

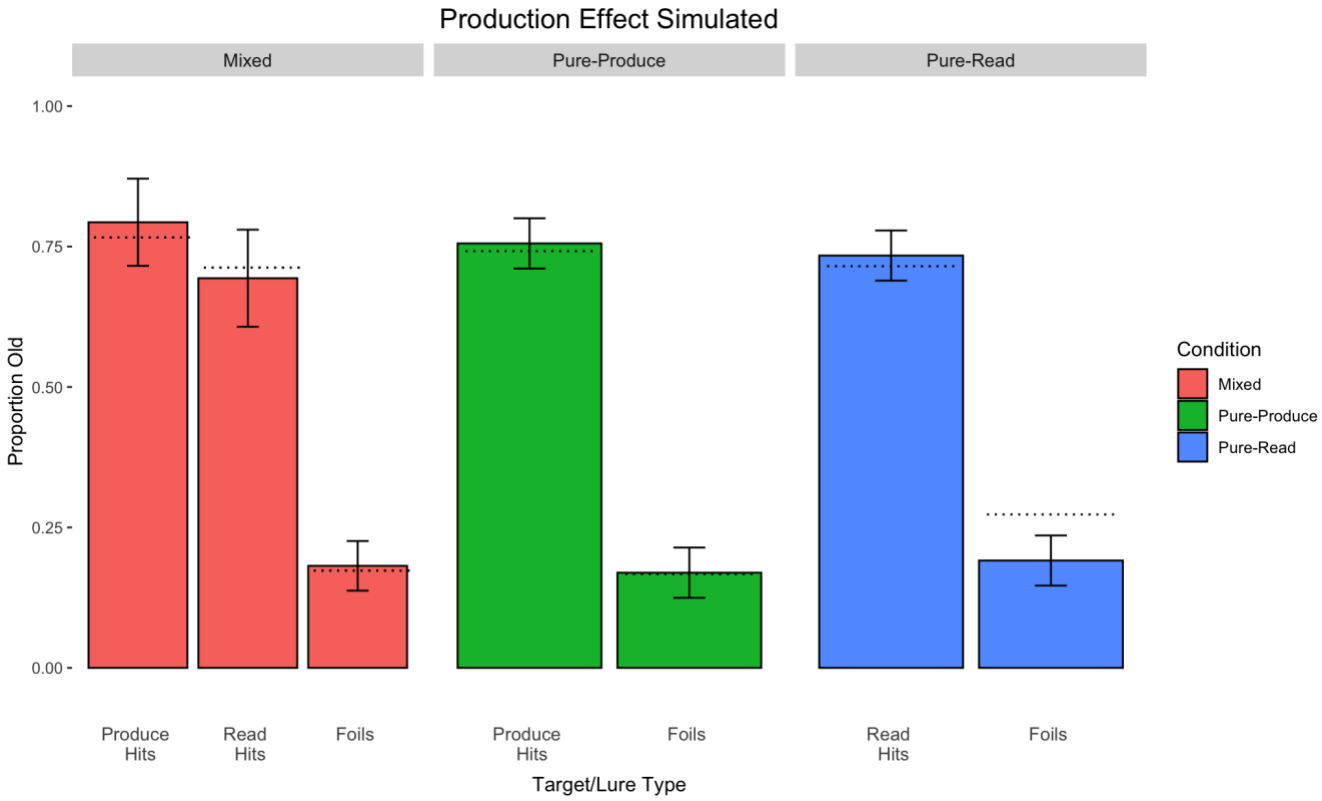


Figure 6.5. Simulation results of the production effect in Experiment 1 integrating strength and distinctiveness. Mixed-list parameters: $L = 0.057$ for produced targets, $L = 0.057$ for read targets, and the number of extra production features for produced targets was 250. The decision criterion was set to a slightly conservative value of 0.4625. Error bars represent standard deviations. Dotted lines represent corresponding empirical means.

Figure 6.5 shows the simulation results of the mixed-list condition on the left-hand side, the pure-produce condition in the middle, and the pure-read condition on the right. By assuming that produced targets are more strongly encoded than read targets and that produced targets were encoded with some extra distinctive information, we are able to generally reproduce the pattern of results that we obtained in Experiment 1 across the three conditions, *RMSE* = 0.0347.

However, the current simulation missed some key aspects of our data. First, the model in its current form predicts a larger mixed-list production effect than what we observed in our empirical data (a 10% vs. 6% production advantage). Second, the model also predicts a lower rate of false alarms to the foils in the pure-read condition. Although the model fit the data fairly well, we sought to explore whether we might get a better fit with only contributions of strength. We next present a simulation of Experiment 1 using only a strength mechanism, where no additional production features were added and where the iterative retrieval mechanism was omitted.

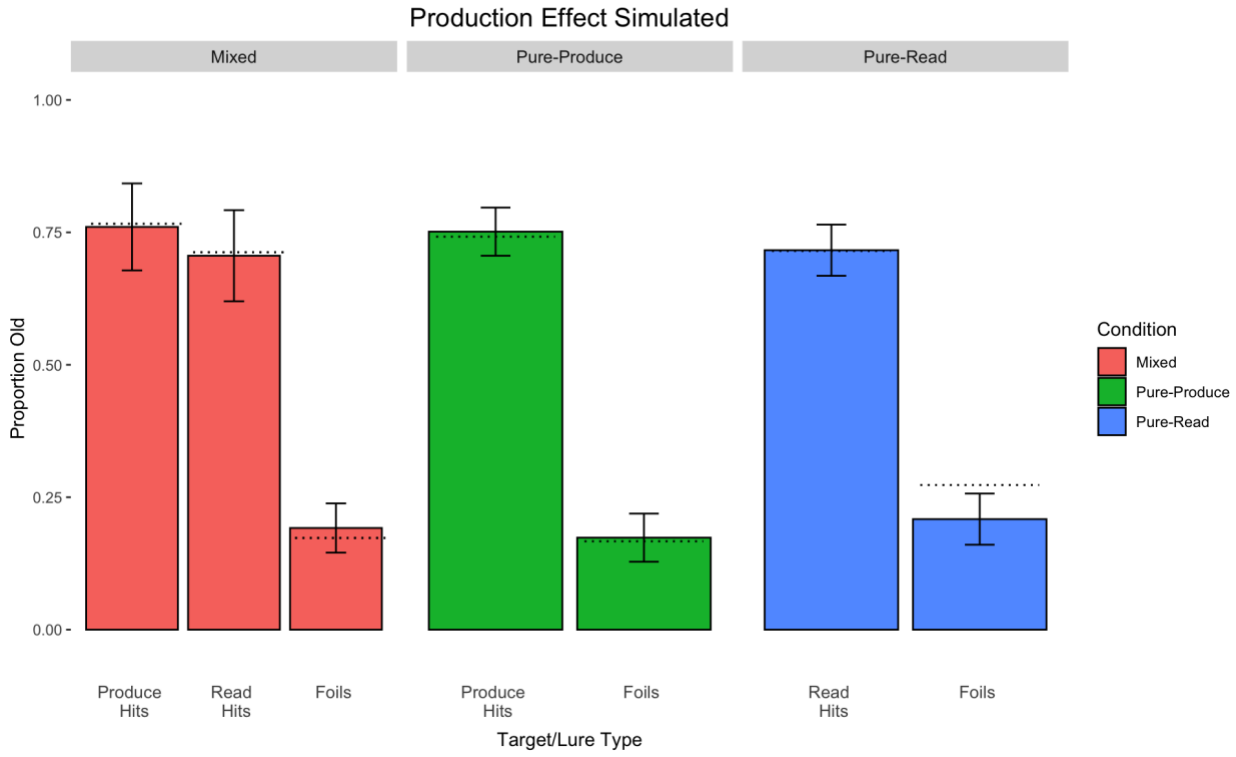


Figure 6.6. Simulation results of the production effect data in Experiment 1 assuming only strength. Mixed-list parameters: $L = 0.0425$ for produced targets and $L = 0.0375$ for read targets. There were no extra production features in the second simulation, as we did not assume any contributions of distinctiveness. The decision criterion was set to a slightly conservative value of 0.4625. Error bars represent standard deviations.

Figure 6.6 shows the simulation results of the mixed-list condition on the left-hand side, the pure-produce condition in the middle, and the pure-read condition on the right. By assuming that produced targets are more strongly encoded than read targets, we were able to closely reproduce the pattern of results that we obtained in Experiment 1 across the three conditions, *RMSE* = 0.0264.

Particularly, the model captured the production effect in the mixed-list condition and captured the trend of a pure-list production effect in the pure-list conditions. The rate of false alarms was different across the three different conditions, and the model was also able to capture this trend, although it underpredicted the heightened rate of false alarms in the pure-read condition. In the pure-read condition, participants had a tendency to say "yes" overall more often than in the pure produce condition. However, for simplicity, we kept the decision criterion in our model at a slightly more conservative value as we did in the following simulation of the directed forgetting experiment. Further, as seen in our empirical data, the model captures a muted "distinctiveness" pattern, where hits in both of the pure conditions sit between the hit rates in the mixed condition, even though there is no overt distinctiveness mechanism defined within the model. Moreover, we obtain a better model fit assuming only strength vs. assuming strength and distinctiveness (*RMSE* = 0.0264 vs. 0.0347).

Simulation of Experiment 2

Although our data again favored a strength mechanism in Experiment 2, we conducted the simulations for the directed forgetting procedure in the same way as the production effect procedure. Once again, we applied both the combined model integrating both strength and distinctiveness mechanisms, as well as the strength-based model of MINERVA 2 (assuming that R-cued items were encoded with more intact features than F-cued items).

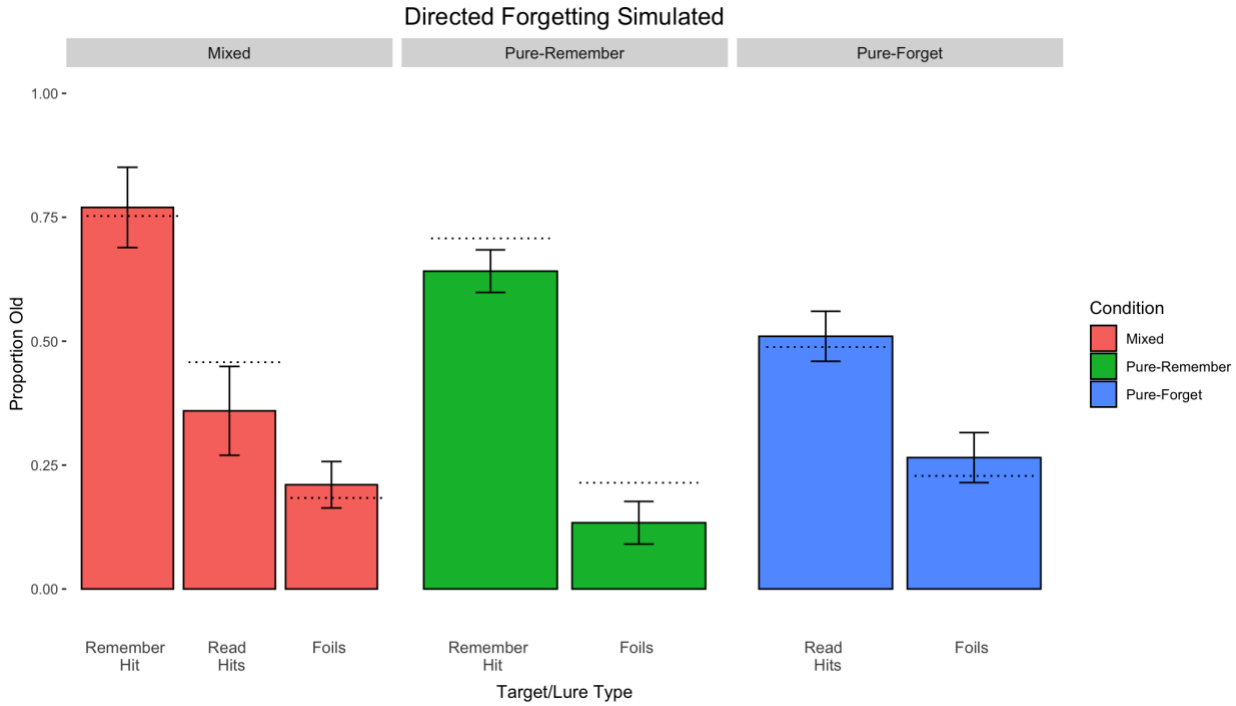


Figure 6.7. Simulation results of the directed forgetting data in Experiment 2. Parameters: L = 0.050 for R-cued targets and L = 0.026 for F-cued targets, and the number of extra elaborative features was 250. The decision criterion was set to a conservative value of 0.3875. Error bars represent standard deviations, and dotted lines represent corresponding empirical means.

Figure 6.7 shows the simulation results of the directed forgetting procedure, with the mixed-list condition on the left-hand side, the pure-remember condition in the middle, and the pure-forget condition on the right. By assuming that R-cued targets are more strongly encoded than F-cued targets, and that there are extra features to account for the elaborative processing of a "remember" instruction, we can roughly capture the pattern of results observed in Experiment 2 across the three conditions, *RMSE* = 0.0578.

However, the simulation missed some key aspects of our data. The combined model integrating strength and distinctiveness predicted a lower hit rate for F-cued items in the mixed-list condition as well as the hit rate and false alarms in pure-remember condition.

As we observed signatures of strength in our directed forgetting data, we again sought to explore if the more parsimonious strength-based model could better capture the patterns in our data.

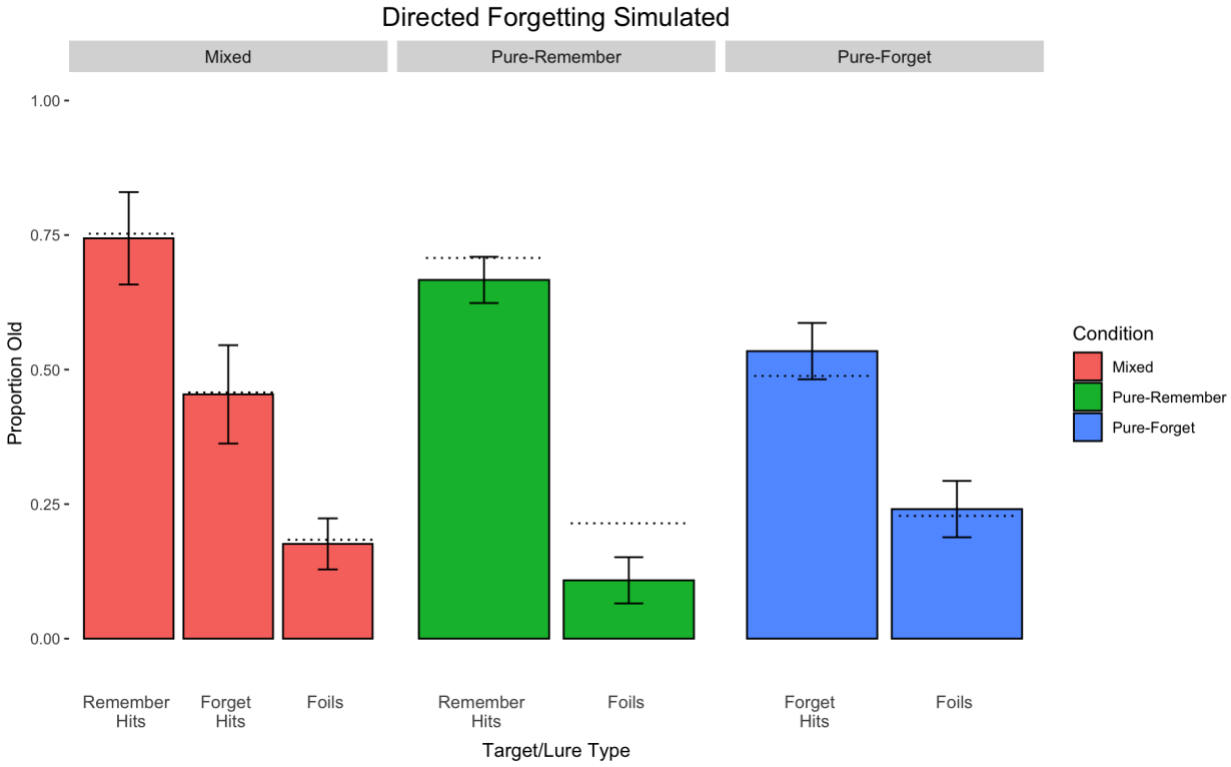


Figure 6.8. Simulation results of the directed forgetting data in Experiment 2. Parameters: L = 0.0425 for R-cued targets and L = 0.0226 for F-cued targets. The decision criterion was set to a conservative value of 0.3875. Error bars represent standard deviations, and dotted lines represent corresponding empirical means.

Figure 6.8 shows the simulation results of the directed forgetting procedure, with the

mixed-list condition on the left-hand side, the pure-remember condition in the middle, and the pure-forget condition on the right. By assuming that R-cued targets are more strongly encoded than F-cued targets, we can closely replicate the pattern of results observed in Experiment 2 across the three conditions, *RMSE* = 0.0471.

As can be seen, the model reproduces the directed forgetting effects in both the mixed- and pure-list designs. In comparison to the production effect, we had to adopt a more conservative decision criterion, driven by the fact that participants gave far fewer "yes" decisions in the pure-forget condition than in the pure-read condition. Additionally, the model once again captured a muted "distinctiveness" pattern, where hits in both of the pure conditions sat between the hit rates in the mixed condition.

Notably, when comparing the parameter values across the production effect and directed forgetting simulations, the value of *L* for produced or R-cued targets remains constant. However, the value of *L* for read or F-cued targets varies across procedures to account for the size of the effect. Our empirical data corroborate this, showing a larger cost to F-cued targets compared to read targets, which yields a greater magnitude effect of directed forgetting relative to production.

Simulation of Experiment 3

Although we found no statistical difference in the mixed-list condition in Experiment 3 compared to Experiment 1, when the cue to "produce" or "read" was presented either concurrently or shortly thereafter, we simulated the mixed-list results of Experiment 3 for a complete account of our data. As we obtained the best fit with the strength-based model for Experiment 1, we simulated Experiment 3 with this same model.

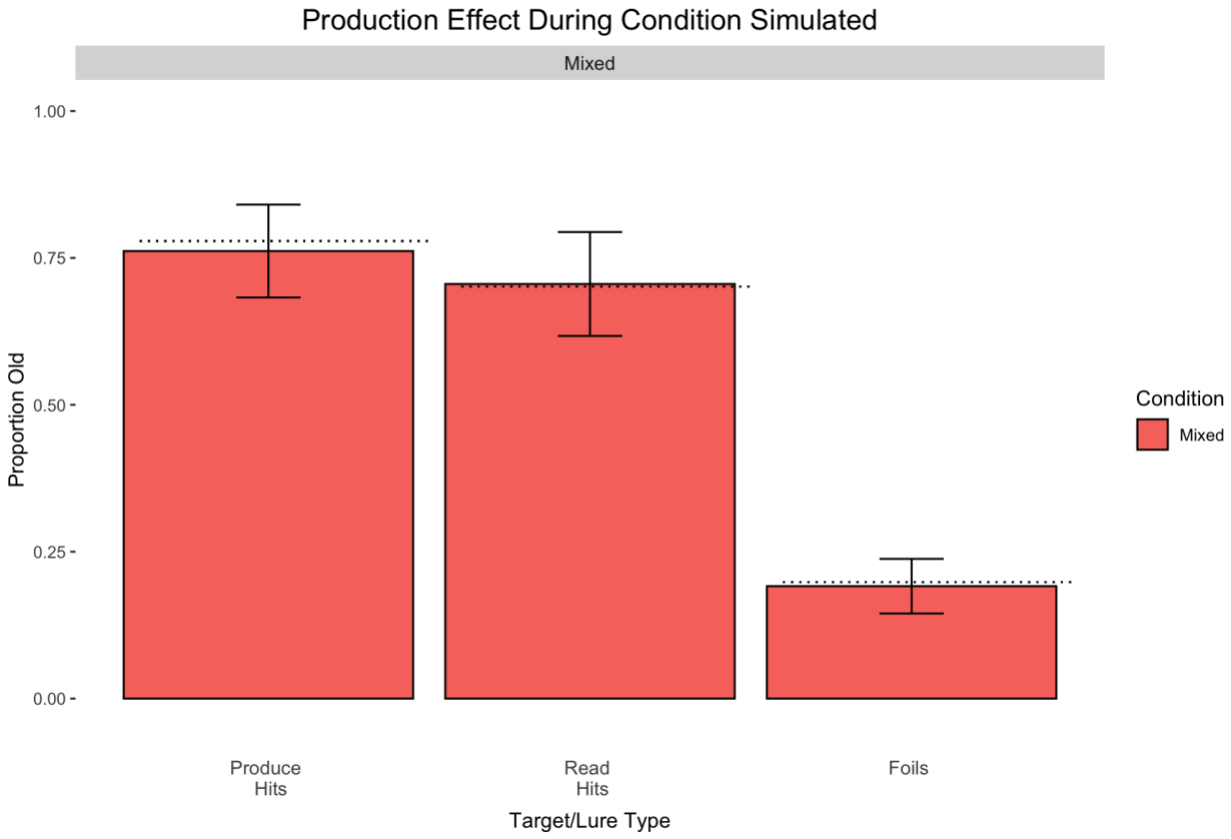


Figure 6.9. Simulation results of the mixed-list production effect data in Experiment 3. Parameters: $L = 0.0425$ for produce targets and $L = 0.0375$ for read targets. The decision criterion was set to a slightly conservative value of 0.4625, as in Experiment 1. Error bars represent standard deviations, and dotted lines represent corresponding empirical means.

Figure 6.9 shows the simulation results of the mixed-list production effect procedure when the cue to "produce" or "read" was presented concurrently with stimulus presentation. With a higher value of L for produce targets than read targets, we again capture the pattern or results as seen in Experiment 3, $RMSE = 0.0477$. Notably, we fit the Experiment 3 data well using the same parameters as used to simulate Experiment 1. This is corroborated by our empirical data, where we observed no difference when the cue to "produce" or "read" was presented concurrently with stimulus presentation or shortly thereafter.

General Discussion

When given the instruction to “produce” or “remember”, in comparison to the instruction to “read” or “forget”, a stable increase in participants’ recognition memory performance can be observed for both classes of elaboratively encoded items. However, our data show that the magnitude of this benefit differs across tasks, which is primarily due to the varying costs associated with the second class of items (i.e., the “read” and “forget” items). Specifically, the directed forgetting procedure produces a larger cost to “forget” items than the production effect incurs to “read” items.

In our empirical data, we obtained a pure-list directed forgetting effect but failed to observe the elusive pure-list production effect. However, the pure-list difference was in the predicted direction, and as evidenced by previous examinations, this difference is known to be small (Fawcett, 2013; see also Fawcett et al., 2023). Furthermore, although the differences between hits in the pure-produce and pure-read conditions were small, participants had more false alarms in the pure-read condition, suggesting that their ability to discriminate targets from lures was weaker in the pure-read condition.

Three main hypotheses have been proposed to explain the item-method directed forgetting effect: selective rehearsal (Woodward & Bjork, 1971; MacLeod, 1998; Hourihan & Taylor, 2006), retrieval inhibition (Bjork, 1989; Brasden et al., 1993), and contextual change (Sahakyan & Kelley, 2002; Fawcett et al., 2024). These suggest that the effect could be due to (1) increased memory rehearsal for elaborately encoded items, (2) active suppression of F-cued items, or (3) a shift in internal context or mental state associated with two different classes of items. In the present investigation, we provide evidence that supports a strength mechanism as

the primary driver of the directed forgetting effect, as demonstrated by our model. With the simplifying assumption that R-cued items are encoded with more intact features into memory than F-cued items, our model can closely capture the patterns we observed in our data, without the need of any additional rehearsal or inhibitory assumptions.

In this study, our goal was to further explore the role of strength and/or distinctiveness and to what extent these two principles contribute to and explain our data in memory studies. Specifically, we investigated whether these principles could reconcile two notable memory phenomena: the directed forgetting effect and the production effect. When comparing mixed- vs. pure-list designs to assess the contributions of strength and distinctiveness, we found no strong evidence in favor of distinctiveness in either procedure. However, we were able to best account for the results using a strength mechanism. Importantly, the strength-based model produced excellent fits to the empirical data from the directed forgetting and production effect tasks.

Typically, with spoken production, the interaction between production and experimental design (mixed vs. pure-list) is significant, lending support to a distinctiveness mechanism. In the current study, we do not believe the lack of a significant interaction between production-by-typing and experimental design to be due to different mechanisms underlying the two procedures (typing vs. spoken production), but rather due to the richness of the representation that is encoded into memory. If spoken production encodes both sensory feedback (auditory) and motoric features into memory, production-by-typing only encodes motoric features into memory, as there is no auditory component to this modality. As such, the differing pattern of results obtained from spoken production or production-by-typing suggests that production-by-typing is a shallower form of encoding in comparison, but robust, nonetheless. Moreover, the difference between pure vs. mixed-lists was in the right direction but did not reach the level of significance.

Rather than strength or distinctiveness, we think it is more likely that these two mechanisms exist on a continuum and can work together in tandem, where one mechanism can be favored over the other, depending on the task at hand. Therefore, we believe that future studies should examine under what circumstances a greater contribution of the distinctiveness mechanism is needed to accommodate findings compared to strength, and under what circumstances each mechanism adds predictive power to the model.

Assuming a strength mechanism, more intact features for elaboratively encoded items result in a richer representation of this class of items. This aligns with the levels of processing explanation (Craik & Lockhart, 1972; Craik & Tulving, 1975). From this perspective, varying levels of strength can be equated to varying levels of processing. In particular, the differing magnitudes of each effect due to the performance of the 'read' or 'F-cued' items can be explained by varying levels of processing in the two cases.

One limitation of the current study is the inclusion of the pure forget condition in the directed forgetting procedure. It could be argued that participants were completely disengaged from the task in this condition since all words were to be forgotten. However, participants were not informed of the specific procedure they were assigned to. They were told that, depending on their assigned condition, they might be instructed to remember all the words or forget all the words. Therefore, we are hesitant to conclude that participants disengaged entirely, as they may have been waiting for an R-cued word, at least for part of the time.

It may also be argued that our findings in the pure-forget condition (and other conditions including an F-cue instruction) reflect a demand characteristic, such that participants are trying to behave as a “good” participant, by responding that they do not recognize an F-cued item (as they

were instructed to forget it), even though they may correctly recognize it from the study phase. To address the issue in directed forgetting studies, a tactic that has been used is monetary compensation for each correctly recognized item, regardless of the cue type. However, when employing such a tactic, investigations consistently show that the forgetting effect remains, even when there is an incentive to remember the F-cued items (Woodward & Bjork, 1971; Bjork & Woodward, 1973; Geiselman et al., 1985; MacLeod, 1999; Aguirre et al., 2020). Although we did not employ such a tactic in our investigation, we believe that the evidence in these demonstrations presents a strong and compelling argument against the claim that a demand characteristic might be driving our pattern of findings. MacLeod (1999) also investigated the role of demand characteristics in directed forgetting and found that under both list and item methods, offering monetary compensation for recall or recognition of F-cued items did not result in participants having any better performance for items that were to-be-forgotten.

Similarly, a pure-read condition in the production effect might invite participants to relax or disengage, as no action is required from them, particularly when compared to mixed-list or pure-produce conditions. Thus, both pure-read and pure-forget conditions seem unnatural when compared to mixed-list or pure-remember/read conditions. However, these conditions were included to complete the full design for assessing strength and/or distinctiveness in the two cases.

In most study instruction manipulations, we rely on trusting that participants are following our instructions, especially when there are no overt measures to collect (e.g., asking participants to imagine doing a task or to remember or forget words). The evidence for participant compliance lies within the data. Our data suggest that participants treat pure-forget

and pure-read items differently compared to pure-remember or mixed-list items, as evidenced by differing hit and false alarm rates. Nevertheless, the purpose of including the pure-list conditions in this instance was to chiefly assess strength and/or distinctiveness.

Although the production effect tends to be larger with spoken production over typed production, the current results with typing show that the effect remains robust. The difference in magnitude between spoken vs. typed production we do not believe to be calling upon different processes, but rather is due to the quality of the signal that is emitted. For example, Murray (1965a; 1965b; Murray et al., 1974) found that the magnitude of recognition performance increased incrementally when the modality of production moved from silent, to whisper, to aloud. Similarly, findings from related work on the drawing effect also demonstrate that there is a larger effect when words are drawn, versus when they are viewed or written (Fernandes et al., 2018). Similar to the production effect, it is argued that the mechanism behind the drawing effect involves elaborative, motor, and pictorial components of a memory trace. However there have been discrepant findings in relation to this claim when the modality of production is by singing. Whitridge et al. (2024) found that the production effect is not always larger for singing than saying words aloud, particularly when words do not appear in the same color at study and at test.

Moreover, according to Caplan and Guitard (2024), it could be that manual typing production places an additional emphasis on motor and orthographic features, which would exist in a higher density subspace than other modalities of production where these might exist in a subspace that includes motor, orthographic, and phonological features. Finally, in terms the differing magnitudes of production effects found across papers, the nature of the stimuli might

play an important role, such that differences in frequency, word length, presentation rate, etc., could affect the results.

The current model. A careful reader may wonder why we chose to use real representations drawn from a normal distribution with $M = 0$ and $SD = 1/\sqrt{(N_{dim})}$ instead of discrete binary values as typically used in a classical MINERVA 2 approach (-1's and +1's). Going forward, we hope to adopt more structured representations, such as those derived from natural language processing models (e.g., LSA; Landauer & Dumais, 1997). By using real valued representations and, consequently, a cosine similarity calculation, going forward, the model is equipped to deal with these kinds of representations. All these changes are forward-looking.

In the same vein, a limitation of the current modeling approach is the random representations of words in memory. By representing words in this fashion, it assumes that all words in memory are orthogonal. However, in practice, words are not orthogonal and share similarities in dimensions such as semantics, orthography, and phonology. Thus, future work should aim to address this issue, where more structured representations could be used.

Although previous to this there are two separate mechanisms of the production effect in the MINERVA 2 approach, it is important to note that the difference between the two is not a large one. The "strength-based" mechanism of the production effect accounts for the typical pattern of results by assuming that there are more intact features in memory (i.e., a richer representation) for produced items vs. read items. The "distinctiveness-based" mechanism of the production effect accounts for the typical pattern of results by assuming that there are additional features added to memory for produced items vs. read items (again, a richer representation). The critical difference between the two models is the iterative retrieval mechanism that operates in

the distinctiveness-based model. Jamieson et al. (2016) accept this and note that "if distinctiveness and strength both work by adding features to a trace in memory, they are correlated concepts…" (p. 160). Thus, although one may call one implementation of the model a "strength" mechanism, and the other a "distinctiveness" mechanism, the two models are very close to being mathematically equivalent. Therefore, formal mathematical models can serve as valuable tools to overcome the limitations of vague verbal descriptions of memory effects.

In addition, this notion of creating a richer representation in memory is amenable to other recent models of the production effect (see Caplan and Guitard, 2024). In this approach, the authors vary the dimensionality and sparsity of vector subspaces, which is akin to a "strength-based" mechanism in the MINERVA 2 approach (i.e., the number of non-zero features stored in memory).

It should be noted that although the current strength-based model best accounts for the results of the experiments that are presented in this paper, it may not account for all data obtained from other production effect experiments, most of which confirm that the production effect arises from distinctiveness rather than strength. Thus, going forward, it would be beneficial to test the model where strength and distinctiveness mechanisms are combined and varied to account for the data in the full database of experimental effects. Critically, although the version of the model that combines strength and distinctiveness does not best fit our current data, we know it could serve as a valuable tool to assess most other production experiments where larger contributions of distinctiveness are observed compared to strength. As such, we find this to be the stronger model of the two presented in this paper.

Conclusion

In our study, we investigated the effects of production and directed forgetting on recognition using both mixed-list and pure-list designs. Our findings reveal that the two effects, commonly attributed to some form of elaborative processing, exhibit variations in effect size. We provide deeper insights into two theoretical mechanisms, strength and distinctiveness, that drive the effects of elaborative processing, whereby distinctiveness need not always be assumed within formal models.

Our findings also provide valuable insights for modeling efforts of directed forgetting and the production effect. Our pattern of results highlights the need for further exploration and refinement of theories regarding the production effect, underscoring different scenarios where strength and distinctiveness may work in concert rather than compete for control in a mutually exclusive arena. Additionally, our research goes beyond advancing our comprehension of directed forgetting and the production effect; it also deepens our understanding of the underlying mechanisms at play. Our data and model suggest that the dichotomy between strength and distinctiveness may oversimplify the matter, as these processes likely interact, and are correlated mechanisms. A comprehensive account must acknowledge this complexity to further delineate the contributions of strength and/or distinctiveness in effects of elaborative processing.

References

Allen, S. W., & Vokey, J. R. (1998). Directed forgetting and rehearsal on direct and indirect memory tests. In Golding, S. M., & MacLeod, C. M. (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 173-195). Psychology Press.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371-391.

Aguirre, C., Gómez-Ariza, C. J., & Bajo, M. T. (2020). Selective directed forgetting: Eliminating output order and demand characteristics explanations. *Quarterly Journal of Experimental Psychology*, *73*(9), 1514-1522.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 553.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger, III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309-330). Erlbaum.

Bjork, R. A., LaBerge, D., & Legrand, R. (1968). The modification of short-term memory through instructions to forget. *Psychonomic science, 10*(2), 55-56.

Bjork, R. A., & Woodward, A. E. (1973). Directed forgetting of individual words in free recall. *Journal of Experimental Psychology*, *99*(1), 22.

Bodner, G. E., Huff, M. J., & Taikh, A. (2020). Pure-list production improves item recognition and sometimes also improves source memory. *Memory & Cognition*, *48*, 1281-1294.

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. L., & 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*(2), 93.

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review*, *21*, 149-154.

Brasden, B. H., Basden, D. R., & Gargano, G. J. (1993). Directed forgetting in implicit and explicit memory tests: A comparison of methods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 603.

Caplan, J. B., & Guitard, D. (2024). A feature-space theory of the production effect in recognition. *Experimental Psychology*.

Chang, M., & Johns, B. T. (2023). Integrating distributed semantic models with an instance memory model to explain false recognition. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 2042-2049). Cognitive Science Society.

Chiu, Y. C., Wang, T. H., Beck, D. M., Lewis-Peacock, J. A., & Sahakyan, L. (2021). Separation of item and context in item-method directed forgetting. *NeuroImage*, *235*, 117983.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341-361.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, *104*(3), 268.

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1797.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Epstein, W. (1969a). Recall of word lists following learning of sentences and of anomalous and random strings. J*ournal of Verbal Learning and Verbal Behavior, 8*(1), 20-25.

Epstein, W. (1969b). Poststimulus output specification and differential retrieval from short-term memory. *Journal of Experimental Psychology, 82*, 168.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within-and between-subjects *manipulation* of the independent variable. *Psychological Bulletin*, *84*(2), 212.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Fawcett, J. M., Baldwin, M. M., Whitridge, J. W., Swab, M., Malayang, K., Hiscock, B., ... & Willoughby, H. V. (2023). Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *77*(1), 35.

Fawcett, J. M., & Taylor, T. L. (2008). Forgetting is effortful: Evidence from reaction time probes in an item-method directed forgetting task. *Memory & Cognition, 36*(6), 1168-1181.

Fawcett, J. M., Taylor, T. L., Megla, E., & Maxcey, A. M. (2024). Active intentional and unintentional forgetting in the laboratory and everyday life. *Nature Reviews Psychology*, 1-13.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302-308.

Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The d-Prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(7), 1090.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*, 1046-1055

Gionet, S., Guitard, D., & Saint-Aubin, J. (2022). The Production Effect Interacts With Serial Positions. *Experimental Psychology, 69*(1), 12-22.

Geiselman, R. E., Rabow, V. E., Wachtel, S. L., & Mackinnon, D. P. (1985). Strategy control in intentional forgetting. *Human Learning: Journal of Practical Research & Applications*.

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025). An embedded computational framework of memory: Accounting for the influence of semantic information in verbal short-term memory. *Journal of Memory and Language, 140*, 104-573.

Hall, K. J., Fawcett, E. J., Hourihan, K. L., & Fawcett, J. M. (2021). Emotional memories are (usually) harder to forget: A meta-analysis of the item-method directed forgetting literature. *Psychonomic Bulletin & Review, 28*, 1313–1326.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review*, *95*(4), 528.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96-101.

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*(4), 242.

Hourihan, K. L., & Taylor, T. L. (2006). Cease remembering: control processes in directed forgetting. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1354.

Jamieson, R. K., Crump, M. J. C. & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior, 40,* 61–82.

Jamieson, R. K., & Hauri, B. R. (2012). An exemplar model of performance in the artificial grammar task: Holographic representation. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale, 66*, 98–105.

Jamieson, R. K., Johns, B. T., Avery, J. E., & Jones, M. N. (2018). An Instance Theory of Distributional Semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40).

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 154.

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *68*(1), 20.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37.

Kelly, M. O., Ensor, T. M., MacLeod, C. M., & Risko, E. F. (2024). The prod eff: Partially producing items moderates the production effect. *Psychonomic Bulletin & Review*, *31*(1), 373-379.

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language*, *123*, 104299.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

MacLeod, C. (1998). Directed forgetting. In J. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 139–172). Erlbaum.

MacLeod, C. M. (1999). The item and list methods of directed forgetting: Test differences and the role of demand characteristics. *Psychonomic Bulletin & Review, 6*, 123-129.

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: Learning, memory, and cognition*, *36*(3), 671.

Marche, T. A., Brainerd, C. J., Lane, D. G., & Loehr, J. D. (2005). Item method directed forgetting diminishes false memory. *Memory*, *13*, 749-758.

Merritt, P., Cook, G., & Wang, M. (2014). Erlebacher's method for contrasting the within and between-subjects manipulation of the independent variable using R and SPSS. *Unpublished manuscript). Retrieved from https://dl. dropboxusercontent. com/u/18192026/Merritt_ Cook_Wang_CSDAv1. pdf*.

Montagliani, A., & Hockley, W. E. (2019). Item-based directed forgetting for categorized lists: Forgetting of words that were not presented. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *73*(3), 135.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

Murray, D. J. (1965b). Vocalization-at-presentation and immediate recall, with varying presentation-rates. *Quarterly Journal of Experimental Psychology*, *17*(1), 47-56.

Murray, D. J., Leung, C., & McVie, D. F. (1974). Vocalization, primary memory and secondary memory. *British Journal of Psychology*, *65*(3), 403-413.

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

Reid, J. N., & Jamieson, R. K. (2022). A computational model of item-based directed forgetting. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *76*(2), 75.

Reid, J. N., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language, 129,* 104397.

Reid, J. N., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language*, *129*, 104397.

Reid, J. N., Yang, H., & Jamieson, R. K. (2023). A computational account of item-based directed forgetting for nonwords: Incorporating orthographic representations in MINERVA 2. *Memory & Cognition*. Advance online publication.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1064.

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

Sheard, E. D., & MacLeod, C. M. (2005). List method directed forgetting: Return of the selective rehearsal account. *Dynamic cognitive processes*, 219-248. Springer, Tokyo.

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*(2), 186.

Tan, P., Ensor, T. M., Hockley, W. E., Harrison, G. W., & Wilson, D. E. (2020). In support of selective rehearsal: Double-item presentation in item-method directed forgetting. *Psychonomic Bulletin & Review*, *27*, 529-535.

Weiner, B. (1968). Motivated forgetting and the study of repression. *Journal of Personality, 36*, 213-234.

Whitlock, J., Chiu, J. Y. C., & Sahakyan, L. (2022). Directed forgetting in associative memory: Dissociating item and associative impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 29.

Whitridge, J. W., Huff, M. J., Ozubko, J. D., Bürkner, P. C., Lahey, C. D., & Fawcett, J. M. (2024). Singing does not necessarily improve memory more than reading aloud: An empirical and meta-analytic investigation. *Experimental Psychology*, *71*(1), 33.

Woodward, A. E., & Bjork, R. A. (1971). Forgetting and remembering in free recall: Intentional and unintentional. *Journal of Experimental Psychology*, *89*(1), 109.

Zacks, R. T., Radvansky, G., & Hasher, L. (1996). Studies of directed forgetting in older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 143.

Zhou, Y., & MacLeod, C. M. (2021). Production between and within: distinctiveness and the relative magnitude of the production effect. *Memory*, *29*(2), 168-179.

Appendix A

Stimuli

ACCOUNT	CENTURY	GARDEN	LANGUAGE	PLATE	TEACHER
ADDRESS	CLOTHES	GLASS	LAUGH	POCKET	THEATRE
AFTERNOON	DAUGHTER	GRAVITY	LEATHER	PORCH	THREAD
AMOUNT	DEBATE	GUARDIAN	LESSON	POWDER	TICKET
ANSWER	DEPARTMENT	HANDLE	MACHINE	QUARREL	TRAFFIC
ARROW	DINNER	HARBOUR	MARKET	QUARTER	TRAVEL
ATTENTION	DIRECTION	HISTORY	MEADOW	QUEEN	TREASURE
ATTITUDE	DISTANCE	HOLIDAY	MERCHANT	RECORD	TROUSERS
AUTHOR	EDUCATION	INDUSTRY	MESSAGE	RESORT	TURNIP
AVENUE	ELECTION	INVENTION	MINUTE	REWARD	UNCLE
BASKET	ENGINE	INVITATION	NEIGHBOUR	RIVER	UNIFORM
BATTERY	ENTRANCE	ISLAND	NEPHEW	SAILOR	VACATION
BEAUTY	ENVELOPE	JOURNEY	OCEAN	SCHOOL	VALLEY
BORDER	EVENING	JUDGE	OFFICE	SHADOW	VICTORY
BRANCH	FACTORY	JUSTICE	ORCHARD	SHOULDER	VILLAGE
BUILDING	FASHION	KETTLE	PACKAGE	SPEECH	WAGON
CAMPAIGN	FOREST	KINGDOM	PAINTING	STATION	WHEAT
CAPITAL	FOUNDATION	KITCHEN	PARTNER	STEAM	WHEEL
CAPTAIN	FRIEND	KNOCK	PEACE	STREAM	WHISPER
CASTLE	FURNITURE	LADDER	PEBBLE	SUMMER	WINTER

Chapter 7: General Discussion

When studying something that one wants to remember, elaboratively encoding that information results in a consistent benefit compared to silently reading. Over a series of three experiments using word pairs, the work in this dissertation has shown that elaborative processing is robust across a number of different procedures — the production effect, generation effect, testing effect, and directed forgetting. Thus, elaboration can take on many forms - producing a word, generating a word from a cue, testing oneself from a cue, or being instructed to remember a word.

In Chapter 2, in a production by typing task, a 7% significant mixed-list production effect was observed, along with a significant pure-list production effect, with a 10% production benefit across pure-lists. Further, the benefit bestowed to produced words was primarily driven by a strength mechanism, as evidenced by no significant difference between the size of the production effect in the mixed- versus pure-list conditions. In terms of costs and benefits of production, I observed neither a benefit of production nor a cost for read items, as evidenced by negligible difference between the produced and read items across the mixed- and pure-list conditions.

In a parallel investigation, Chapter 3 examined the generation effect using the same materials and design as in Chapter 2. There was a 25% significant generation effect in the mixed-list, and a 16% pure-list generation effect. However, the results again supported a strength account; I failed to observe a significantly larger mixed-list generation effect than a pure-list generation effect. Further, when comparing the hit rates for generated items across the mixed- and pure-list conditions, I did not observe a benefit of generation, but I did observe a cost, as the hit rate for read items was lower in the mixed-list condition than in the pure-list condition.

In Chapter 4, the testing effect was examined using the same materials and design as before, where a 5% significant testing effect was observed in the mixed-list condition, and a 7% significant effect in the pure-list condition. Results once again supported a strength mechanism, where the mixed-list testing effect was not significantly larger than the pure-list testing effect. Neither a cost for read items nor a benefit for tested items was observed in the testing effect procedure, where there were once again negligible differences between produced and read items across the mixed- and pure-list conditions.

In Chapter 5, I presented a model of elaborative processing. Across five simulations, I was able to show that a parsimonious strength mechanism was able to best account for the results obtained in Experiments 1-3. Assuming that there were more intact features encoded for elaboratively processed items in memory, the model was able to closely capture the pattern of results across the three conditions across the three procedures.

In Chapter 6, I presented a closely related examination of the production effect and directed forgetting, where a production by typing task was again used to examine the production effect. Obtained results indicated that both the production effect and directed forgetting was once again due primarily to a strength-based mechanism, where a parsimonious strength model was able to best account for the results.

Across four different elaborative encoding effects — the production, generation, testing, and directed forgetting effects — a consistent pattern emerged, supporting a strength-based mechanism as the primary driver of the benefit observed in these procedures. The production effect demonstrated that actively producing words, such as typing them out, enhanced recognition without introducing significant costs. Similarly, the generation effect showed a

robust memory advantage for self-generated words, though a cost was observed for read items in the mixed-list condition. The testing effect further reinforced this pattern, with self-testing improving recall through strengthened memory traces. Finally, the directed forgetting effect was consistent with this same strength-based account. The computational model that was presented confirmed that a parsimonious strength mechanism could accurately explain the findings across all procedures, suggesting that when using a typing task, elaborative encoding enhances memory by increasing the robustness of stored information rather than through unique mechanisms specific to each effect.

A Principled Approach

This project aimed to study three effects of elaborative processing under a common theoretical framework, in an effort to enhance our understanding of elaborative processing and memory performance. The current set of results extend the 2016 MINERVA 2 modeling framework of the production effect to other effects of elaborative processing and modalities, in an aim to build a formal theory that brings together related cognitive phenomena (Jamieson et al., 2016).

Uniting Siloed Approaches

This project further aimed to bring together three areas (and a fourth in Spear et al., 2025) of related memory effects into a common theoretical framework. In a series of three experiments, I examined the production effect, the generation effect, and the testing effect, to investigate the possibility of explaining all three with the same computational model of memory. The goal was to investigate whether a common mechanism could be responsible for each, to support a coherent account of elaborative processing across multiple phenomena.

Some researchers have claimed that the field of psychology is still dominated by naive empiricism that other disciplines have managed to escape (van Rooij et al., 2024). However, the inclusion of computational models into this line of research will enhance understanding of underlying cognitive mechanisms surrounding elaborative processing that leads to improved memory performance. Formal theories have the ability to allow for the discovery of novel insights that might not otherwise have been detected (Lewandosky, 1993). Moreover, taking verbal theories and expressing them as formal theories helps progress our science and offers a way out of relying solely on productive but imprecise mechanistic explanations. The goal of the current project was to help extend this effort, in the realm of the production effect, and in two other related effects – the generation effect and the testing effect. Formal theories help explicitly express ideas, such that:

> "The vaguenesses that have plagued the theory of higher mental processes and other parts of psychology disappear when the phenomena are described as programs. There is every reason to believe that it will prove equally fruitful in application to the theories of learning, of perception, and of concept formation." (Newell, Shaw, & Simon, 1958, p. 166).

Moreover, the approach that Jamieson et al. (2016) used is akin to the approach that Hintzman used in 1988 to model and account for the levels of processing framework. Adding more features to a trace reflects deeper encoding and allows the model to capture the production effect. Because of the nature of MINERVA regarding the production effect (and the generation and testing effects), utilization of the production record does not necessitate that this process is intentional or conscious. Rather, the relative distinctiveness conferred by production may be a byproduct of the retrieval process that activates traces that contain additional production features.

In this way, it can be considered that the effect falls out of the retrieval process and is a byproduct of how the theory operates at retrieval, as these features are subsequently picked up by the process of an iterative retrieval mechanism (i.e., on the first iteration, the probe presented to memory contains no production features). However, it is important to note that MINERVA is limited to particular contexts and cannot account for cognitive mechanisms such as attribution (e.g., misattributions of enhanced perceptual fluency to memory for a prior experience; Goldinger & Hasen, 2005). A fluency heuristic could arise due to several factors, including pleasantness, duration, and recency (Whittlesea, 1993). However, it has also been noted that determining the source of this fluency is important; the fluency heuristic is not as simple as "if fluent, then old" (Whittlesea, 1993). The MINERVA 2 model does not account for these processes in its current implementation. Instead, recognition decisions are based on a simple familiarity value where, if this value exceeds a criterion, the model responds "old", else "new".

One issue common among all these approaches is that they all rely on study instruction manipulations. Each of the previously mentioned effects is a study manipulation, where participants are instructed to do something above and beyond passively observing the study item. An inherent issue that exists with study manipulation approaches is that they rely on the assumption that participants are doing what they are being told to do (e.g., in an online experiment, if a participant is asked to produce a word by spoken production, are they actually following these instructions?). One strength of the current approach that was implemented is that production, generation, or testing were accomplished by typing. Typing is easy to record, and manipulation checks were in place to ensure that participants were complying with the experimental instructions.

In the same vein, the orienting instructions included in the experiments conducted here increased the amount of experimental control that was implemented, but conversely this implemented a tradeoff for ecological validity. When one sets out to learn or remember new material, no such orienting instructions exist; one must, and will, choose their own strategy. However, the current investigation warranted such an approach because of the nature of the goals at hand: to examine what further mediates an elaborative processing effect, to examine what is encoded into the memory trace during an elaborative encoding effect procedure, and to understand how that information is used at the moment of test.

Levels of Processing

Our findings are not the first to fail to observe a distinctiveness effect within the production effect. A similar pattern of results was observed in Bodner et al. (2016), where a cost to read items was found, instead of a benefit to produced items. With this observed pattern, one possible explanation that was put forth by the authors was the "lazy reading hypothesis", which posits that instead of produced words receiving a boost in performance via production, read words instead incur a cost, where participants fail to encode these items as strongly (Begg & Snider, 1987). This hypothesis fits well with a levels of processing explanation, where words that are not elaboratively encoded are not encoded deeply.

Moreover, the current examination yielded smaller production effects than is what is typically observed with other modalities of production (e.g. vocalization), thus this may be part of what is driving the current pattern of results. However, the results in Experiment 2 examining the generation effect yielding a much larger effect in comparison. Why the discrepancy?

We know that in other effects in cognitive psychology, such as the Stroop effect (Stroop, 1935), that reading a word is so automatic that it takes extra effort to inhibit that pattern (e.g. to say the color of the word instead of the word itself). Why then should saying a word aloud or typing it produce such a large benefit as opposed to other effects of elaborative processing (e.g. the larger observed effect of generation over production). Thus, in terms of a levels of processing effect, the differing levels or level of effort involved with the production effect versus the generation effect, it follows that the generation effect is larger than the production effect.

In terms of the testing effect, although it may be argued that this effect involves the most cognitive effort, we did not observe the largest effect in this examination. Thus, it may be that although testing enhances retrieval strength, it may not lead to deeper encoding than generation does, particularly if retrieval attempts are unsuccessful or not effortful enough. Thus, although the testing effect is known to be robust despite accuracy of retrieval, future parallel investigations could examine whether accuracy of retrieval changes the magnitude of the effect using the procedure presented in this dissertation. Moreover, testing may have a larger effect over the long term, whereas the generation effect is more beneficial over the short term, as was observed in the present investigation.

Additionally, although this dissertation examined four effects of elaborative processing, the production generation, testing, and directed forgetting effects - this is not an exhaustive list of elaborative processing effects that exist in our literature. Future work should aim to continue uniting siloed approaches, to find shared vs. differential mechanisms across different procedures.

Features – what do they represent? All of the aforementioned models of the production effect assume that there is something added to the memory trace to confer the benefit of production. However so far, little work has been conducted to investigate what these added features might

represent, and the content of those extra features remains unspecified. In each of the three out of four production effect models that were mentioned, the extra features to represent production are random. Yet it is unlikely that these added features are truly random in nature. Therefore, a theoretical question to ask is what these features should be and what they contain?

Language, and more specifically words, can be controlled on a number of different dimensions including word frequency, concreteness, and phonological regularity. However, for any particular dimension on which words are controlled, should these unique features be orthogonal or should they share some sort of similarity to one another? In other words, what is the nature or structure of these features?

There have been many proposals on what the features may be, but there has been very little formal investigation. Of those proposals, there have been accounts that suggest that features may represent the act of vocalization, where "According to the proceduralist framework, the process of vocalizing at study will be retained in a record of that processing" (Forrin et al., 2012; p. 1046; Kolers, 1973; Kolers & Roediger, 1984). Other verbal accounts have posited that saying something aloud adds additional features to memory that represent articulation and audition (Ozubko & MacLeod, 2010). More recently, Caplan and Guitard (2024) introduced the AST mathematical model, designed to further clarify the types of features that can be encoded into memory and how this encoding occurs. In their examination of a production-by-vocalization procedure, they suggested that phonological features, rather than orthographic ones, are the primary focus during encoding and are responsible for the elaborative encoding advantage observed for produced items over read items.

One set of firmer experimental evidence that the memory trace may be content-specific can be taken from Goldinger and Azuma (2004). Here, participants read words aloud over two

sessions, before and after being presented with auditory training tokens (the auditory training tokens were audio recordings of volunteers speaking words in voices that covered a wide perceptual range). Audio recordings of the participants speaking words aloud before and after exposure to the auditory training tokens were made, where perceptual judges assessed the similarity of the auditory training tokens to spoken words of the participants, thereby judging the degree of imitation exhibited by participants. Results suggested evidence of postexposure imitation, where speaking style was influenced by the auditory training tokens. This finding suggests that the memory traces of spoken words contain detailed representation of auditory information that was initially presented to them.

In addition, results from an fMRI study of the production effect showed greater activation in the sensorimotor cortex and auditory cortex, regions associated with articulation and perception, during the study phase (Bailey et al., 2021). The activation of these regions was further correlated with increased rates of recognition for produced items in the test phase. Therefore, is it motoric features that are encoded into memory?

This line of research has a lot in common with other research that is being conducted in other closely related domains. Recently, in the domain of natural language processing, work regarding the meaning of dimensions has also been a popular line of research (Hollis & Westbury, 2016). When natural language processing models (e.g., LSA) process text, once a word by document matrix is constructed, a dimension reduction technique (singular value decomposition) is used to derive new representations that represent word meaning. Up until recently, these dimensions remained unlabeled and what they represent remained a black box. However common in other areas of research, when data reduction techniques such as PCA are used, the goal is usually then to take the newly formed dimensions a step further, such that they

are interpreted and then labeled. Therefore, efforts toward investigating what unknown features are is not novel, nor theoretically unmotivated.

Moreover, thus far in the applications of the MINERVA model, and other related models (e.g., REM or the revised feature model) to the production effect, the content of these extra features has been assessed to be random, and treated equally despite production modality or material differences (although see Caplan and Guitard, 2024). However, strong empirical evidence has suggested that these added production features are not random; rather, evidence suggests that these added features conferred by production are modality and stimuli dependent.

Taken together, it is unlikely that traces are composed of random features. Therefore, a potential future direction that would take the MINERVA model of the production effect by Jamieson et al. (2016) a step further, would be to conduct a series of experiments that vary a particular dimension of the stimuli, thereby potentially changing the contents of the memory trace, where the extra production features ought to change in regards to the variable by which stimuli are being selected by (e.g., orthography or phonology). Doing so, would help elucidate the source of the production effect more formally. Put more eloquently by Mama and Icht (2016), they stated:

> "The simple reasoning explains why memory improves for words read aloud within visual presentation of the study words. Silent reading involves a single encoding process, whereas oral reading involves an extra pair of processes—(1) articulation (the execution of a motor action) and (2) audition (hearing oneself saying the word). Other learning conditions such as writing or mouthing also augment memory relative to silent reading because each entails one extra process of encoding (motor action). Since speaking aloud (vocalising) involves two

additional distinct processes, it results in the largest memory advantage relative to all other methods of production." (p.100)

From this they conclude that the magnitude of the production effect is determined by the number of *unique* encoding processes involved. However, a careful approach is warranted, as it has yet to be determined how much overlap or uniqueness particular processes might have (e.g., orthography and phonology are not easily separable dimensions). Nevertheless, from the above reasoning, this may present a path forward to investigating what the features in a memory model might represent. The production effect thus lends itself as a convenient procedure to more closely investigate what is being encoded into memory when one is learning and engages in one kind of elaborative processing versus another.

Implications

An important implication of the current set of experiments and investigation is that the model that is presented can account for four effects of elaborative processing in a very similar fashion. In addition, this work further sheds light on the mechanisms that underlie these effects of elaborative processing, and what might underlie additional effects of elaborative processing. However, further work investigating additional effects of elaborative processing is needed to confirm whether this is the case.

Moreover, the pattern of results obtained across three different experiments in this dissertation challenges prevailing wisdom that it is primarily distinctiveness that is responsible for the production effect, and other related effects of elaborative processing. Yet although the current set of results are in favor of a strength mechanism, it is unlikely that a strength signature rules out the role of distinctiveness.

The current investigation also expands the work that has been done using a typing procedure to investigate the production effect. With typing, we observed a smaller production effect than is typically observed when using other methods of production (e.g., vocalization). As the size of the effect is smaller with typing, this may contribute to the failure to observe any effects of distinctiveness across our procedures. Although the magnitude of the effects is different, it is doubtful that this in indicative of calling upon different fundamental processes. Rather, strength and/or distinctiveness could become more favored depending on the task at hand.

Strength *and* distinctiveness. This work unites two principles of cognition—strength and distinctiveness—and demonstrates that the two need not be mutually exclusive, but rather likely work together in tandem in our effects of elaborative processing. Although the current set of results presented in this dissertation primarily favor a strength-based mechanism, other empirical investigations of elaborative processing have obtained results that are indicative of a distinctiveness-based mechanism. Instead of postulating that these effects of elaborative processing are due to one mechanism over another, it is more probable that they are due to the combination of different mechanisms, where one may play a larger contribution than the other, depending on the task at hand. Thus, rather than saying a particular effect is due to strength *or* distinctiveness, it is more probable that strength *and* distinctiveness can contribute to the consistent benefit we observe in procedures of elaborative processing.

Although the parsimonious strength-based model best accounted for the results in this dissertation, the model where strength and distinctiveness mechanisms were combined will serve as a valuable tool for future investigations of elaborative processing, where more distinctiveness-based results may be observed in work to come.

Conclusion

As noted in Goldinger and Azuma (2004), it is important to recognize that although the current research proposed was conducted using words, these stimuli are not the focus, but rather a methodological convenience by which to study memory. As such, this line of research offers many future directions for other research to be conducted, as it can easily be extended to other elaborative processing effects, such as the drawing effect (Fernandes et al., 2018), picture superiority effect (Shepard, 1967), and other similar effects. Furthermore, although this research has been proposed to be conducted only with typing, depending on the task at hand, differing sizes of the production effect may be found, as it has been found to be larger when production is accomplished by vocalization rather than typing (see MacLeod et al., 2010). Moreover, there is no reason that the current modeling approach would not extend to other modalities of production (e.g., writing).

The principle of parsimony suggests that simpler explanations should be prioritized, with more complex ones considered only when necessary (see Morgan, 1903). While the current findings support a strength-based mechanism, they do not rule out the role of distinctiveness. Indeed, patterns of distinctiveness may emerge in future investigations depending on the specific task. To account for this, we have developed a model capable of handling such variations.

Building formal theories that unify common mechanisms and principles is essential for bridging different areas of research that study similar cognitive processes. The goal of our discipline should not be solely to uncover novel behavioral and cognitive findings, but also to identify overarching principles that integrate related results. Rather than developing separate theoretical explanations for each effect-driven line of research, we should strive to connect them within a broader framework that explains human behavior more comprehensively.

Moreover, advancing our understanding of memory—both in laboratory settings and real-world applications—is critical for improving educational approaches and informing the development of human-inspired artificial intelligence.

This work further clarifies the mechanisms underlying effects of elaborative processing. Specifically, we demonstrate that two fundamental principles of memory—strength and distinctiveness—can explain four well-documented effects: the production, generation, testing, and directed forgetting effects. By gaining deeper insights into these mechanisms, we can advance theoretical frameworks of memory, refine educational strategies, and contribute to the development of artificial intelligence inspired by human cognition.

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371-391.

Bacon, F. (2000). *Novum organum* (L. Jardine & M. Silverthorne, Trans.). Cambridge University Press. (Original work published 1620).

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., ... Fawcett, J. M. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition*, *152*, 105757.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 553-563.

Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 977-989.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201-210.

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. L., & 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*(2), 93-98.

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1711-1719.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220-251.

Burns, D. J. (1992). The consequences of generation. *Journal of Memory and language*, *31*(5), 615-633.

Butler, A. C. (2010). Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Caplan, J. B., & Guitard, D. (2024). A Feature-Space Theory of the Production Effect in Recognition. *Experimental Psychology*, *71*(1), 64–82.

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to an implicit rule-learning task: Implicit learning of semantic structure. *Quarterly Journal of Experimental Psychology, 69*(6)*,* 1049-1055.

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, *22*(1), 267-281.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In *Attention and Performance VI* (pp. 535-555). Routledge.

Collins, R. N., Milliken, B., & Jamieson, R. K. (2020). Minerva-de: An instance model of the deficient processing theory. *Journal of Memory and Language*, *115*, 104151.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684.

Curtis, E. T. (2019). Interactive processes in an instance model of memory: A computational analysis of Jacoby's (1983) dissociation between perception and recognition. *Canadian*

Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, *73*(4), 288-294.

Curtis, E. T., & Jamieson, R. K. (2019). Computational and empirical simulations of selective memory impairments: Converging evidence for a single-system account of memory dissociations. *Quarterly Journal of Experimental Psychology*, *72*(4), 798-817.

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The Production Effect Over the Long Term: Modeling Distinctiveness Using Serial Positions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *48*(12), 1797–1820..

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47, 1–12.

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*(3), 180–187.

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*(1), 155-161.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*(1), 180-209.

Dutton, J. M., & Starbuck, W. H. (1971). *Computer simulation of human behavior*. Wiley.

Einstein, G. O., & McDaniel, M. A. (1987). Distinctiveness and the mnemonic benefits of bizarre imagery. In: McDaniel, M.A., Pressley, M. (eds) *Imagery and related mnemonic processes* (pp. 78-102). Springer.

Engelkamp, J. (1995). Visual imagery and enactment of actions in memory. *British Journal of Psychology*, *86*(2), 227-240.

Engelkamp, J., & Dehn, D. M. (2000). Item and order information in subject-performed tasks and experimenter-performed tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 671-682.

Engelkamp, J., & Krumnacker, H. (1980). Image-and motor-processes in the retention of verbal materials. *Zeitschrift für experimentelle und angewandte Psychologie*, *27*(4), 511–533.

Engelkamp, J., & Zimmer, H. D. (1989). Memory for action events: A new field of research. *Psychological Research*, *51*(4), 153–157.

Engelkamp, J., Zimmer, H. D., & Kurbjuweit, A. (1995). Verb frequency and enactment in implicit and explicit memory. *Psychological Research*, *57*(3), 242-249.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within-and between-subjects manipulation of the independent variable. *Psychological Bulletin*, *84*(2), 212-219.

Estes, W. K. (1973). Phonemic coding and rehearsal in short-term memory for letter strings. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 360-372.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, *12*(3), 263-282.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670-2698.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329-335.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Fawcett, J. M., Baldwin, M. M., Whitridge, J. W., Swab, M., Malayang, K., Hiscock, B., ... & Willoughby, H. V. (2023). Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *77*(1), 35.

Fawcett, J. M., Bodner, G. E., Paulewicz, B., Rose, J., & Wakeham-Lewis, R. (2022). Production can enhance semantic encoding: Evidence from forced-choice recognition with homophone versus synonym lures. *Psychonomic Bulletin & Review*, *29*(6), 2256–2263.

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 99.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302-308.

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory*, *22*(5), 470-480.

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 125-138.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*(7), 1046-1055.

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110-119.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1-58.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*(4), 716-722.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21-23.

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025). An embedded computational framework of memory: Accounting for the influence of semantic information in verbal short-term memory. J*ournal of Memory and Language, 140*, 104573.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological review, 93*(4), 411-428.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96-101.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744-1756.

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in psychology*, *5*, 886.

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of verbal learning and verbal behavior*, *22*(5), 485-508.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649-667.

Jacoby, L. L., & Craik, F. I. M. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 1–21). Lawrence Erlbaum Associates.

James, W. (1890). *The principles of psychology*. Holt

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*(2), 119-136.

Jamieson, R. K., Crump, M. J., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*(1), 61-82.

Jamieson, R. K., & Mewhort, D. J. K. (2009). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, *62*(9), 1757-1783.

Jamieson, R. K., Hannah, S. D., & Crump, M. J. (2010). A memory-based account of retrospective revaluation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *64*(3), 153-164.

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010). Global similarity predicts dissociation of classification and recognition: Evidence questioning the implicit–explicit learning distinction in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1529-1535.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 154-164.

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, *1*(3), 174-183.

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *68*(1), 20-28.

Java, R. I. (1996). Effects of age on state of awareness following implicit and explicit word-association tasks. *Psychology and Aging*, *11*(1), 108-111.

Johns, B. T. (2022). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. *Memory & Cognition*, *50*(5), 1013-1032.

Johns, B. T., Jones, M. N., & Mewhort, D. J. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*(4), 486-518.

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 180-190.

Karpicke, J. D., & Roediger III, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319* (5865), 966-968.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227-239.

Kelly, M. O., Ensor, T. M., MacLeod, C. M., & Risko, E. F. (2024). The prod eff: Partially producing items moderates the production effect. *Psychonomic Bulletin & Review*, *31*(1), 373–379.

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language*, *123*, 104299.

Kelly, M. O., Lu, X., Ensor, T. M., MacLeod, C. M., & Risko, E. F. (2024). Productions Need Not Match Study Items to Confer a Production Advantage, But It Helps. *Experimental Psychology*, *71*(1), 2–13.

Kinjo, H.; Snodgrass, J.G. (2000). Does the generation effect occur for pictures?. *The American Journal of Psychology. 113*(1), 95–121.

Kornell, N. & Terrace, H. S. (2007). The generation effect in monkeys. *Psychological Science, 18*(8), 682-685.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12*(4), 703-710.

Kwantes, P. J., & Mewhort, D. J. K. (1999). Modeling lexical decision and word naming as a retrieval process. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 53*(4), 306-315.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, *4*(4), 236-243.

Lockhart, R. S., & Craik, F. I. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *44*(1), 87-112.

Loftus, G. (1985). Johannes Kepler's computer simulation of the universe: Some remarks about theory in psychology. *Behavior Research Methods, Instruments, & Computers*, *17*(2), 149-156.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Landauer, Thomas K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-40.

Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect for learning new material. *The Journal of General Psychology*, *130*(2), 171-188.

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*(2-3), 291-310.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*(4), 390-395.

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: Learning, Memory, and Cognition*, *36*(3), 671-685.

MacLeod, C. M., Ozubko, J. D., Hourihan, K. L., & Major, J. C. (2022). The production effect is consistent over material variations: support for the distinctiveness account. *Memory*, *30*, 1000-1007.

Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory*, *24*(1), 98-113.

Marchal, A., & Nicolas, S. (2000). Is the picture bizarreness effect a generation effect?.*Psychological Reports*, *87*(1), 331-340.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review, 27*, 1139-1165.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, *15*(2), 237-255.

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, memory, and Cognition*, *12*(1), 54-65.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200-206.

McDaniel, M. A., Riegler, G. L., & Waddill, P. J. (1990). Generation effects in free recall: Further support for a three-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 789-798.

McFarland Jr, C. E., Frey, T. J., & Rhodes, D. D. (1980). Retrieval of internally versus externally generated words in episodic memory. *Journal of Verbal Learning and Verbal Behavior*, *19*(2), 210-225.

McNamara, D. S., & Healy, A. F. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. *Journal of Memory and Language*, *43*(4), 652-679.

Meade, M. E., Wammes, J. D., & Fernandes, M. A. (2018). Drawing as an encoding tool: Memorial benefits in younger and older adults. *Experimental Aging Research*, *44*(5), 369-396.

Merritt, P., Cook, G., & Wang, M. (2014). Erlebacher's method for contrasting the within and between-subjects manipulation of the independent variable using R and SPSS. *Unpublished manuscript). Retrieved from https://dl. dropboxusercontent. com/u/18192026/Merritt_ Cook_Wang_CSDAv1. Pdf.*

Morgan, C. L. (1903). *An introduction to comparative psychology*. Walter Scott.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior*, *16*(5), 519-533.

Mulligan, N. W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 838.

Mulligan, N. W., Smith, S. A., & Buchin, Z. L. (2019). The generation effect and experimental design. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(8), 1422-1431.

Murray, D. J. (1965). Vocalization-at-presentation, auditory presentation and immediate recall. *Nature*, *207*(5000), 1011-1012.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251-269.

Nairne, J. S. (1991). Positional uncertainty in long-term memory. *Memory & Cognition*, *19*(4), 332-340.

Neath, I. (1999). Computer simulations of global memory models. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 74-80.

Newell, A. (1998). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. In A. Clark & J. Toribio (Eds.), *Machine Intelligence* (1st ed., pp. 121–146).

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151-166.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596-1618.

Poirier, M., Yearsley, J. M., Saint-Aubin, J., Fortin, C., Gallant, G., & Guitard, D. (2019). Dissociating visuo-spatial and verbal working memory: It's all in the features. *Memory & Cognition*, *47*(4), 603-618.

Raaijmakers, J. G., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In *Psychology of learning and motivation* (Vol. 14, pp. 207-262). Academic Press.

Ratcliff, R., & McKoon, G. (1995). Sequential effects in lexical decision: Tests of compound-cue retrieval theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1380-1388.

Reid, J. N., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language*, *129*, 104397.

Reid, J. N., Yang, H., & Jamieson, R. K. (2023). A computational account of item-based directed forgetting for nonwords: Incorporating orthographic representations in MINERVA 2. *Memory & Cognition*, *51*(8), 1785-1806.

Roberts, B. R., MacLeod, C. M., & Fernandes, M. A. (2022). The enactment effect: A systematic review and meta-analysis of behavioral, neuroimaging, and patient studies. *Psychological Bulletin*, *148*(5-6), 397.

Roberts, B. R., & Wammes, J. D. (2021). Drawing and memory: Using visual production to alleviate concreteness effects. *Psychonomic Bulletin & Review*, *28*(1), 259-267.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043-1056.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Roediger, H. L. III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

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

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125-140.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145-166.

Simon, Herbert A. 1996. *The Sciences of the Artificial* (3rd ed.). Cambridge, MA: MIT Press.

Simon, H. A. (1992). What is an "explanation" of behavior?. *Psychological Science*, *3*(3), 150-161.

Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, *6*(1), 1-10.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592-604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, *26*(6), 589-607.

Stahl, C., & Aust, F. (2018). Evaluative conditioning as memory-based judgment. *Social Psychological Bulletin*, *13*(3), 1-30.

Staniland, J., Colombo, M., & Scarf, D. (2015). The generation effect or simply generating an effect? *Journal of Comparative Psychology*, *129*(4), 329-333.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.

Surprenant, A. M., & Neath, I. (2013). *Principles of memory*. Psychology Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155-185.

Thomas, A. K., & Loftus, E. F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition*, *30*(3), 423-431.

Tulving, E. (1962). Subjective organization in free recall of unrelated words. *Psychological Review*, *69*(4), 344-354.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*(3), 127-128.

van Rooij, I., Devezer, B., Skewes, J., Varma, S., & Wareham, T. (2024). What Makes a Good Theory? Interdisciplinary Perspectives. *Computational Brain & Behavior*, *7*(4), 503–507.

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*(8), 1229-1237.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752-1776.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2017). Learning terms and definitions: Drawing and the role of elaborative encoding. *Acta Psychologica*, *179*, 104-113.

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

Worthen, J. B. (2006). Resolution of Discrepant Memory Strengths: An Explanation of the Effects of Bizarreness on Memory. In *Distinctiveness and Memory*. Oxford University Press.

Worthen, J. B., & Eller, L. S. (2002). Test of competing explanations of the bizarre response bias in recognition memory. *Journal of General Psychology, 129*(1), 36-48.

Worthen, J. B., & Roark, B. (2002). Free Recall Accuracy for Common and Bizarre Verbal Information. *The American Journal of Psychology*, *115*(3), 377–394.

Worthen, J. B., Starns, J. J., Loveland, J. M., & Eisenstein, S. A. (2006). Influence of orienting task on memory for bizarre and common stimuli: Evidence against a surprise-based explanation. *Advances in cognitive psychology*, 1-18.

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory*, *21*(8), 904-915.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Contributions of Authors

Spear, J., Reid, N., Guitard, D., & Jamieson, R. K.,(2024). Directed forgetting and the production effect: Assessing strength and distinctiveness. *Experimental Psychology, 71* (5), 278–297.

The paper titled “Directed forgetting and the production effect: Assessing strength and distinctiveness” presented in Chapter 6 was primarily programmed, conducted, analyzed, and written by Jackie Spear. The other co-authors contributed to the conceptualization and editing of the paper, however, the primary author was Jackie Spear.

Appendix A

Debates surrounding the nature of the cognitive and memorial mechanisms that underlie the production effect have yielded several current and competing modeling approaches. Moreover, given the popularity of the production effect and its empirical robustness, a number of different laboratories have made recent efforts to explain the effect computationally. Specifically, these effects have taken the form of MINERVA 2 (Jamieson et al., 2016), the revised feature model (Saint-Aubin et al., 2021), Attentional Subsetting Theory (Caplan & Guitard, 2025) and REM (Kelly et al., 2022). Common to all of these models is the addition of features to account for the added benefit of production.

REM. REM is based on the Search of Associative Memory (SAM) model, extending that framework from the domain of recall into the domain of recognition (Raaijmakers & Shiffrin, 1980). Briefly, SAM is designed with four free parameters that allow the model to account for different types of associations. Parameter *a* accounts for context associations, parameter *b* accounts for inter-item strength, parameter *c* represents the associations between an item and itself, and parameter *d* accounts for any preexisting associations. From this, the model is able to account for free recall, word frequency effects, and associative priming (Ratcliff & McKoon, 1995). Even with this success, SAM and other models are limited in the number of phenomena they can explain, with one of the shortcomings being the mirror effect. (Need to explain the mirror effect.) This therefore was one of the major motivations for the development of REM.

In a typical configuration of REM, memory is a matrix. The first encounter with a stimulus results in a new memory trace represented by a vector of features sampled from a geometric distribution, with values ranging from one to infinity, where parameter *w* determines the number of features (usually 20) and *g* determines the shape of the geometric distribution. If *g*

is a large value, features drawn from the geometric distribution are common (i.e., it is more likely that features will be overlapping); if *g* is a small value, then features that are sampled from the geometric distribution are uncommon (i.e., it is more likely that features will be nonoverlapping, or distinctive). In accordance with this geometric distribution then, the model defines the features' distinctiveness in proportion to its statistical likelihood. That is, because lower values are more common, when there is a match on a larger, more "rare" feature, this is weighted more heavily by the model, or considered more diagnostic, because of the intrinsic nature of the sampling distribution implemented in the model. As such, traces contain feature values that vary in evidentiary value.

Next, whether a feature is stored in memory is determined by parameter u^*, and if that feature is stored according to u^*, then *c* is the probability that that feature is *accurately* copied to memory (else the feature will be copied to memory with a different value sampled from the same geometric distribution). If a feature is not stored to memory according to parameter u^*, then that feature is set to zero. Therefore, features in REM consist of correctly copied features, incorrectly copied features, and missing features.

REM falls within the family of global matching models; therefore, recognition is accomplished by matching a probe to all traces that are stored in memory. This matching processing occurs on a feature-to-feature basis which operates in parallel. Specifically, retrieval in REM is operationalized by calculating the average similarity of the probe to all traces stored in memory. Similarity, λ, is calculated for each item, which activates traces in proportion to their match to the probe, where the model then returns the average of all trace activations, which serves as a global sense of familiarity. Moreover, similarity (λ) incorporates a Bayes rule, which compares expected similarities and differences according to whether the probe presented to

memory was a study item or a foil. Put differently, in comparison to other global matching models, REM compares an expected similarity to the retrieved similarity, such that the calculation is based on the expectation that an item is a target or foil. Similarity for item *j* is defined as:

$$\lambda_j = (1-c)^{n_{jq}} \prod_{i=1}^{\infty} \left[\frac{c+(1-c)g(1-g)^{i-1}}{g(1-g)^{i-1}}\right]^{n_{ijm}} \quad (1)$$

where *c* is the probability of copying a stored feature correctly, *g* is the probability of success when sampling from the geometric distribution, n_{jq} is the number of nonzero mismatching features in image *j*, n_{ijm} is the number of nonzero features in image *j* that match the probe and have value *i*, and where *m* and *q* stand for match and mismatch, respectively.

From this, the model averages all of these likelihood ratios that are computed from the comparison of the probe to all traces in memory and then converts this to an old/new decision, where a "yes" is issued when the mean odds ratio is greater than 1; else no. This average is calculated as:

$$\Phi = \frac{1}{n}\sum_{j=1}^{n} \lambda_j \quad (2)$$

where λ_j is the above likelihood ratio for each trace in memory, *j*, and where *n* is the total number of stored traces in memory in comparison to the probe.

In sum, REM operates with Bayesian style decisions, where specifically a match of a common feature is less diagnostic than a match on an uncommon feature, and where this is taken into consideration when calculating the recognition decision.

In Kelly et al. (2021), this version of REM is applied to the production effect. To accomplish this, some further modifications are made to the model. If a word is produced, then

an additional set of 10 production features were added to the base representation of that word (represented by 20 features). If a word is not produced, then those additional features contain no information (i.e. they are set to zero). All probes presented to memory at test contain these extra production features, and all of these probes, regardless of status (e.g. old/new, produced/read), have a probability of containing some production features, introducing some uncertainty and noise within the model with regard to knowing whether an item was studied or not studied. A further modification is made to the model to mimic the passage of time, as it is thought that using production features during a decision takes additional time. This is accomplished by varying the number of probe features that are available during retrieval. If there are fewer features available, there are fewer instances than there are matches between the probe and the trace. Therefore, functionally, the model's assessment that the probe is a target is reduced. When all features are available, this constitutes a full retrieval time, and when fewer features are available, this constitutes speeded retrieval conditions. Thus, in practice, the production effect is reduced under speeded retrieval conditions, which follows from previous empirical investigations (Forrin et al., 2012).

The Feature Model. The feature model was first implemented by Nairne (1990), and was developed to model immediate serial recall and modality-based effects (i.e., better recency effects for auditorily than for visually presented material). Although the feature model presented here deals with recall (and not recognition as in this project), it is discussed as another model that has been applied to an effect of elaborative processing in general and to the production effect in particular. Items in memory are represented as multicomponent vectors, which serve to represent different kinds of information and that have the ability to be overwritten by external or internal events. Retrieval is similarity based, computed based on the amount of featural overlap between

the probe and memory depending on psychological distance computed as Minkowski similarity (Shepard, 1962).

The feature model is designed to simulate an immediate serial recall task. As with other models, it is assumed that items are represented by a set of features, that typically take on values of 1 or -1. Of these features, there are two types: modality-dependent features (representing type of presentation for the stimuli) and modality-independent features (used to represent the nature of the stimuli). Importantly, memory traces that are internally generated contain only modality-independent features; the modality-dependent features contain no information. Additionally, the feature model assumes a distinction between primary and secondary memory. The function of primary memory is to maintain cues for items that were recently presented, which become partially degraded memory traces, as subsequent items can interfere with immediately preceding items – a process known as retroactive interference. Furthermore, as implemented in the feature model, retrieval confusion as a function of serial position and interference of items will not drift far from their original position, which is motivated by previous empirical findings (Poirier et al., 2019; Nairne, 1991). The function of secondary memory is to store all studied traces which are intact representations.

Recall begins with a degraded cue in primary memory, where recall candidates (traces in primary memory) are compared to intact traces in secondary memory. Next, the cue in primary memory that has the highest probability of being the cue for the first item is selected according to a similarity ratio such that:

$$P_s(SM_j|PM_i) = \frac{s(i,j)}{\sum_{k=1}^{n} s(i,k)} \quad (3)$$

where $P_s(SM_j|PM_i)$ is the conditional probability that the secondary memory trace, SM_j, will be sampled given the primary memory trace, PM_i, which is dependent on $s(i,j)$, where $s(i,j)$ is the

similarity between primary trace *i* and secondary trace *j*, and where *Σs(i,k)* is the sum of the similarity between primary trace *i* and all secondary memory items included in the search. Specifically, similarity is simply the amount of feature-to-feature overlap between the two compared traces:

$$s(i,j) = e^{-d_{ij}} \quad (4)$$

where *s(i,j)* represents the computed similarity between primary memory trace PM_i and secondary memory trace SM_j, *e* is Euler's number which helps transform distance from 0 - ∞ to 0 - 1, and where distance, *d*, is calculated by adding the number of mismatched features, *M*, between item *i* and item *j*, and dividing by the number of compared features, *N*, as in:

$$d_{ij} = \frac{a}{N}\sum\ b_k M_k \quad (5)$$

where *N* is the number of compared features, M_k is the number of times feature position x_{ik} does not match feature position x_{jk}, and where *a* is a scaling parameter that can be mapped onto available attentional resources, and that can be adjusted to raise or lower overall performance. Lastly, parameter b_k can be used to weight particular feature comparisons (e.g., to increase attention to certain modality-dependent features). This ratio works such that if the numerator is higher, then the probability of sampling a memory trace is higher, whereas if the denominator is relatively larger, then the probability of sampling a memory trace decreases. Therefore, similarity, *s(i,j)*, is related to how different two items are, d_{ij}.

When a second item is presented, retroactive interference can occur whereby features from the preceding item can overwrite some of the features of the present item. This overwriting process results in a feature being rewritten as 0. Thus, the feature model uses the process of interference and does not decay over time to account for the degradation of memory traces. In

practice, order errors occur due to a gradual loss of precision in positional coding (Poirier et al., 2019).

Recency effects emerge from the model because of internal rehearsal of those items. Because this internally generated process has no modality-dependent information, only the modality-independent features can be subjected to overwriting as a consequence of rehearsal. What consequence does this have? Within the domain of immediate serial recall, the feature model is able to account for patterns of behavior that have been labeled: recency effects, modality effects, and phonological similarity effects (Deese & Kaufman, 1957; Watkins & Watkins, 1977; Estes, 1973). However, the model in its current instantiation is limited to immediate serial recall and cannot account for free recall or recognition.

In the revised feature model proposed by Saint-Aubin et al. (2021), the production effect is modeled in the domain of immediate serial recall. The revised feature model is designed upon the following assumptions about the production effect. First, relative distinctiveness appears to be important. Second, in pure lists, modality of production is also important. Third, rehearsal is important for demonstrations of the production effect on recall performance.

As in the original feature model, items are represented by two types of features – one type representing modality (e.g., perceptual characteristics of the stimuli) and the other type representing modality independent features (e.g., conceptual meaning or category knowledge). Once the model is presented with an item, traces are generated in primary and secondary memory where the traces consist of vectors of features, and where values in these traces consist of randomly generated values ranging from 1-3. Next, the traces in primary memory are subjected to an overwriting phase, which accounts for retroactive interference from the presentation of subsequent items, and where this interference is similarity based, resulting in

more overwriting of similar items than dissimilar items. Furthermore, this overwriting process is not limited to the immediately preceding trace, as in the original feature model. This is accomplished by setting the probability that a feature of item *n-m* will be overwritten to (i.e., the probability that any particular item will be overwritten as a function of its serial position):

$$e^{-\lambda(m-1)} \tag{6}$$

where *e* is Euler's number, and when λ approaches infinity, the overwriting process of the original feature model is recovered (which only allows the overwriting of the most recent item/trace). This change allows retroactive interference to operate further back than just the recent item, which has been supported in the data from previous empirical examinations (Saint-Aubin et al., 2021).

As before, traces in secondary memory remain unchanged. Once all study items have been presented, one last overwriting phase of modality-independent features occurs to account for the cognitive preparation of recall.

Thus, two of the ways that the revised feature model differs from the original feature model is in the addition of a rehearsal process and a modified overwriting process. The added rehearsal process in the revised feature model works to restore some of the overwritten features with a probability contingent on a free parameter *r*, representing the effectiveness of rehearsal and the number of items presented. As such, the probability that an overwritten feature will be restored is:

$$p = r \times e^{-\frac{(n-1)^2}{9}} \tag{7}$$

where *r* is a parameter representing the effectiveness of rehearsal, *e* is Euler's number, *n* is the number of items, and the value 9 in the denominator of the exponent comes from previous work

suggesting a decrease of rehearsal for lists longer than four items. This is a free parameter that can be changed to model rehearsal patterns that appear in the data.

The overwriting process is further modified to be able to account for a delay of recall (or the inclusion of a filler task). To account for order, each item is tagged with order information, with a further additional parameter θ to allow for drift of that position, which is held constant at 0.05.

As for retrieval, and as in most other memory models, the presented traces serve as cues to memory, where the similarity between the cue and the stored item is calculated to serve as an activation. This is accomplished by determining a similarity value, which is calculated on a feature-to-feature basis according to a scaling constant multiplied by the number of mismatches between the cue and the probe. This similarity is then used in a probability soft-max function, which can be tuned according to a temperature parameter that controls which similarity value is chosen by the model. If the temperature parameter is high, then all items have an equal probability of being chosen, regardless of the similarity value of the cue, otherwise, when the value is lower, then the item with the highest similarity to the cue is chosen. This is accomplished by:

$$p(SM_j|PM_i) = \frac{e^{\frac{s(i,j)}{\tau}}}{\sum_k e^{\frac{s(k,j)}{\tau}}} \quad (8)$$

where the conditional probability that the secondary memory trace SM_j will be sampled, given the primary memory trace PM_i, depends on $s(i,j)$, which is the computed similarity between PM_i and SM_j, and where τ is the "temperature parameter".

As such, similarity is related to the feature-to-feature correspondence between the primary and secondary traces. This similarity is determined by calculating the psychological distance between the two traces, *i* and *j*, where:

$$s(i,j) = \text{e} \tag{4}$$

where *e* is Euler's number, and where the distance, d_{ij} is simply calculated by adding the number of mismatched features, *M*, and then dividing that by the total number of features (*N*), where:

$$d_{ij} = \frac{a}{N}\sum\ b_k M_k \tag{5}$$

where *a* is a scaling constant and b_k is an attention bias parameter (this can give more "value" to some dimensions, but is almost never used). Thus, the larger d_{ij} is, the smaller $s(i,j)$ will be.

Saint-Aubin et al. (2021) also include an extra probability function that an item is not activated enough to be recalled to account for omissions (*null*), which receives a similarity that is equal to a value that would be obtained if two independently chosen vectors of features were selected, which is approximately:

$$s(null, j) = 1.7 \text{ x } 10^{-2} \tag{9}$$

Functionally, this helps calculate a threshold of recall output. Lastly, some interference in the output is implemented in the model by assuming that there is recovery from the secondary memory set that is related to prior recall of the item, where:

$$P_r = e^{-cr} \tag{10}$$

where *e* is Eurler's number, *c* is a scaling constant and *r* is the number of times a sampled item has already been recalled. This works to implement the idea that the probability of recovering an item decreases the more times it has been previously recalled.

The production effect is accounted for in the same way that the feature model accounts for modality effects. The feature model assumes that auditory traces have a greater number of modality-dependent features than do traces generated solely through visual presentation.

The probability that any given target is selected can be thought to be akin to a measure of distinctiveness. By incorporating rehearsal and relative distinctiveness into the model, the revised feature model accounts for the production effect in an immediate serial recall task for both pure and mixed lists quite well.

AST. The Attentional Subsetting Theory (AST) is another framework proposed to explain the production effect in the domain of recognition. Briefly, AST assumes that production leads to the encoding of additional production features (Caplan & Guitard, 2025). However, in comparison to other models of the production effect, AST accomplishes this a slightly different way. AST incorporates both strength and distinctiveness, by manipulating the dimensionality of feature spaces. Caplan and Guitard (2024) assume that shallow phonological features are drawn from a sparse feature space, whereas deeper features such as strength manipulations (e.g., study time), are drawn from a denser feature space. By manipulating the sparsity of the feature space from which features are drawn, both strength and distinctiveness mechanisms are integrated into the model (i.e. a less sparse feature space results in both increased strength, and where these additional features can represent a distinctiveness heuristic, such that they are available for produced items but not unproduced items). In addition, the theory assumes that one “attends” to a small subset of features, where only attended to features are encoded into memory. Depending on the task at hand, different features can be attended to more or less. For example, if the task is a production by vocalization procedure, it is assumed that more phonological features are

attended to. Whereas if it is a production by typing task, then it would be assumed that more orthographic features are attended to.

Mechanistically, added features that represent the act of production are assumed to be accessed early in the processing of an item, thus production enhances memory by these features being activated earlier on (in contrast to other models of the production effect). In addition, AST assumes that the features that are encoded into memory represent specific properties of the word, such as phonological features, orthographic features, and semantic features.

As with the classic instantiation of MINERVA 2, recognition is accomplished by computing the similarity of the probe and memory, by computing the dot product. This similarity is then compared to a criterion, where if the computed dot product is greater than this criterion, the model responds “Old”, else “New”. By incorporating item-specific features, and by manipulating feature subspaces, AST helps make progress in the effort toward what the specific features that are encoded into memory may be, and how those features are represented.

Appendix B

Materials used in Experiments 1-3, taken from Johns and Swanson (1988).

Stem	**Words**			**Nonwords**		
ack	tack	back	pack	cack	nack	gack
act	pact	tact	fact	lact	mact	bact
age	rage	wage	page	bage	lage	tage
ail	wail	jail	nail	dail	cail	lail
ain	rain	main	gain	bain	sain	hain
air	pair	fair	hair	kair	dair	tair
ake	bake	cake	fake	nake	pake	dake
ale	gale	sale	male	nale	fale	rale
all	fall	tall	wall	dall	nall	rall
alt	malt	salt	halt	talt	lalt	balt
ame	tame	game	lame	bame	wame	rame
amp	lamp	damp	ramp	famp	mamp	pamp
and	band	sand	land	pand	cand	tand
ane	sane	pane	lane	gane	rane	tane
ang	gang	hang	rang	lang	nang	vang
ank	rank	tank	bank	mank	gank	cank
ape	cape	tape	gape	hape	lape	mape
ard	lard	hard	card	mard	sard	pard
are	bare	care	dare	sare	lare	gare
ark	lark	hark	dark	sark	tark	rark
art	part	dart	cart	lart	gart	sart
ash	hash	rash	sash	tash	fash	nash
ask	task	mask	cask	lask	rask	dask
ass	mass	lass	pass	tass	hass	rass
ast	mast	fast	last	bast	dast	tast
ath	bath	math	path	fath	gath	dath
ave	cave	save	pave	mave	bave	lave
awn	dawn	lawn	fawn	bawn	mawn	gawn

ead	bead	read	lead	nead	gead	fead
eak	beak	weak	peak	geak	seak	meak
eal	deal	veal	real	beal	geal	feal
eam	seam	team	beam	peam	neam	keam
ean	mean	lean	dean	rean	tean	cean
ear	fear	gear	dear	mear	cear	lear
eat	heat	feat	neat	leat	reat	deat
eck	deck	neck	peck	leck	meck	reck
eed	feed	deed	reed	meed	beed	leed
eek	peek	meek	week	beek	heek	feek
eel	peel	keel	reel	deel	neel	beel
een	seen	keen	teen	deen	geen	reen
eep	weep	peep	seep	teep	geep	feep
eer	beer	leer	deer	neer	teer	geer
eet	beet	feet	meet	reet	neet	leet
ell	well	sell	cell	rell	lell	kell
elt	melt	felt	pelt	nelt	helt	delt
end	tend	lend	mend	dend	pend	hend
ent	vent	sent	cent	fent	ment	kent
est	pest	nest	best	mest	dest	fest
ice	dice	mice	rice	sice	bice	gice
ick	sick	wick	nick	bick	vick	cick
ide	tide	side	wide	lide	mide	fide
ife	fife	life	wife	bife	sife	tife
ift	sift	lift	gift	nift	hift	bift
ike	bike	like	pike	sike	fike	rike
ile	file	tile	mile	lile	dile	kile
ill	pill	mill	will	lill	vill	cill
ilt	kilt	silt	tilt	milt	filt	dilt
ime	dime	lime	mime	fime	pime	bime
ind	find	wind	kind	lind	cind	dind

ine	wine	mine	vine	rine	bine	gine
ing	ring	wing	king	fing	hing	ling
ink	mink	pink	link	bink	gink	tink
int	hint	mint	lint	fint	bint	rint
ipe	wipe	pipe	ripe	sipe	tipe	mipe
ire	tire	dire	wire	bire	nire	pire
iss	hiss	miss	kiss	tiss	riss	liss
ist	list	mist	fist	rist	dist	bist
ite	mite	bite	site	tite	gite	dite
ive	five	live	dive	sive	bive	rive
oad	goad	toad	road	boad	coad	moad
oat	coat	goat	boat	loat	roat	hoat
ock	lock	mock	dock	bock	gock	nock
ode	node	code	rode	fode	sode	tode
old	gold	cold	sold	nold	rold	dold
ole	mole	role	sole	cole	fole	lole
ond	bond	fond	pond	cond	lond	mond
one	lone	tone	cone	fone	sone	mone
ong	song	long	gong	mong	pong	cong
ook	cook	book	look	mook	dook	sook
ool	pool	cool	fool	rool	mool	bool
oom	doom	loom	room	soom	toom	noom
oon	loon	moon	coon	roon	doon	poon
oop	goop	loop	hoop	doop	moop	soop
oot	boot	loot	root	goot	noot	doot
ope	hope	cope	dope	bope	gope	sope
ore	more	core	bore	dore	rore	nore
ork	fork	pork	cork	lork	gork	bork
orn	horn	corn	torn	norn	gorn	dorn
ose	hose	rose	nose	cose	mose	tose
oss	moss	toss	loss	foss	poss	goss

ost	post	most	host	rost	bost	gost
out	gout	bout	pout	mout	sout	vout
ove	rove	cove	wove	tove	nove	gove
own	town	gown	down	bown	hown	lown
uff	puff	muff	cuff	fuff	suff	duff
ull	full	bull	pull	sull	rull	tull
ump	jump	bump	pump	fump	cump	nump
ung	rung	sung	lung	tung	nung	gung
unt	hunt	punt	bunt	lunt	dunt	munt
ure	pure	cure	sure	mure	ture	bure
ust	dust	must	gust	nust	pust	cust
ute	cute	mute	lute	wute	pute	rute