HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

Episodic Retrieval: Functions and Measurement

presented by Ruben van Genugten

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Daniel Schacter

Signature _____

Typed name: Prof. Jason Mitchell

Signature _____

Typed name: Prof. Patrick Mair

Signature _____

Typed name: Prof. Talia Konkle

Signature _____

Typed name: Prof.

Date: March 31, 2022

Episodic Retrieval: Functions and Measurement

A dissertation presented by

Ruben van Genugten

to

The Department of Psychology

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in the subject of Psychology

Harvard University
Cambridge, Massachusetts

March 2022

Advisor: Daniel L. Schacter

Committee: Jason P. Mitchell, Patrick Mair, and Talia Konkle

Dissertation Advisor: Daniel L. Schacter Ruben van Genugten

Episodic Retrieval: Functions and Measurement

Abstract

A growing body of evidence indicates that people imagine specific experiences by retrieving and recombining elements of their episodic memories, a process often referred to as episodic simulation. As a result, episodic retrieval can contribute to domains of cognition not traditionally investigated in episodic memory research, so long as performance in these domains benefits from richly imagining events. In my dissertation, I will first illustrate this phenomenon in two domains. For the first and second paper, I will provide evidence that episodic retrieval contributes to mentalizing and creative writing. To further facilitate research on episodic simulation, I will use my third paper to develop and validate software that automatically scores details in episodic narratives, based on scoring guidelines from the Autobiographical Interview.

Table of Contents

Acknowledgements

This dissertation is dedicated to the wonderful people who have been with me on this journey.

To Dan: I will always be grateful for the opportunity you extended me when you invited me into your lab. I have learned an immeasurable amount from you and through the discussions here, and I want to thank you for all of your guidance as I developed my skills as a scientist and writer. I am leaving the lab immensely grateful for my time here. As Karl Szpunar said last week, it's a pretty special place.

To Jason: For the long conversations on different scientific ideas, experimental designs, and generally exciting discussions about mentalizing, thank you. It was a true joy to work with you and to be welcomed into your lab. I am very proud of the neuroimaging work we did, and I want to thank you for jumping into a rabbit hole of research with me.

To Patrick: For making me fall in love with the world of statistics, thank you. Your teaching has truly shaped my career. I taught my first class and last class with you, and I'm grateful for all our conversations along the way.

To Talia: For helping me with fMRI analyses when you had no obligation to do so, for your questions and encouragement, and for serving on this dissertation committee, thank you.

To mom, dad, Jesse, and Shirley: Thank you for your unending support and your constant belief in me. This dissertation is truly a product of your efforts as well, for I couldn't have done this without you. I love you.

Thank you to all of the Schacter and Mitchell lab members, and especially Helen, Alexis, and Sarah, for their friendships and science chats.

Last, to everyone in my cohort: I am glad you are the random group of people who came on this journey with me.

Introduction[1]

While daydreaming about a vacation to avoid the cold Boston winter, we might think about escaping to a beach in Mexico. Or, we can imagine taking advantage of the snow to go skiing. By mentally experiencing and testing out our possibilities before we invest resources in a specific option, we can potentially maximize benefits and minimize costs without engaging in the actual behavior (Ingvar, 1979). For example, we can imagine skiing down the mountain and hurting ourselves, then decide to avoid the potential costs associated with skiing and go to the beach instead. Once we have decided which option to pursue, imagining the situation further helps us plan for it. In this case, we can imagine the sun beaming down on the beach, then realize that we probably need to pack swimsuits and sunscreen. Simulations such as these can help us try out alternative possibilities and prepare to engage in the chosen option (Jing et al., 2017). Work over the past two decades has started to show how we are able to construct these vivid and helpful simulations.

A large body of work shows that imagining future experiences relies on many of the same brain regions as remembering past experiences (for review, see Schacter et al., 2007, 2012). For example, when participants are presented with a word cue or phrase (e.g. 'beach') and are asked to imagine a specific future event or remember a past event related to the cue, many of the same regions showed similarly increased activity compared with a control task that elicits semantic and visuo-spatial processing but does not involve remembering or imagining a specific event (Addis et al., 2007). Other studies, too, have documented remarkably similar activation profiles within the default network (Buckner et al., 2008) for imagining and remembering (e.g.,

[1] Some content in the introduction of this dissertation is adapted from van Genugten & Schacter (2021).

Addis et al., 2009; Okuda et al., 2003; Szpunar et al., 2007). Together, this set of regions, which includes medial temporal and frontal lobes, posterior cingulate and retrosplenial cortex, and lateral parietal and temporal regions, has been characterized as a core network that serve both remembering and imagining (Benoit & Schacter, 2015).

However, the kind of overlap observed in these studies alone does not provide conclusive evidence that the same mechanism is responsible for remembering the past and imagining the future. Many tasks elicit default network activity (e.g. creativity tasks, navigation, theory of mind, memory, mind wandering, self-referential processing, and counterfactual thinking), and it is not clear that all of them involve the same neural computations (Beaty et al., 2016; Buckner et al., 2008; Mason et al., 2007; Ochsner et al., 2004; Schacter et al., 2015). Additional evidence is needed before concluding that remembering the past and imagining the future rely on shared processes. This additional evidence comes from studies of amnesic patients with medial temporal lobe damage, who have difficulty remembering specific past events. Many of these patients are also unable to imagine future and other hypothetical events to the same degree as healthy controls (e.g., Tulving, 1985; Hassabis et al., 2007; Race et al., 2011; but see also Dede et al., 2016; Squire et al., 2010).

A variety of theoretical interpretations of these observations have been put forward (e.g., Buckner & Carroll, 2007; Hassabis & Maguire, 2007; Suddendorf & Corballis, 2007). Here we focus on an approach referred to as the *constructive episodic simulation hypothesis* (Schacter & Addis, 2007a, 2007b, 2020), which builds on earlier observations by Tulving (1985, 2002) implicating episodic memory in the ability to project into the future. According to this hypothesis, we construct future and other hypothetical events by flexibly retrieving and recombining elements of different episodic memories (that is, memories of specific occurrences).

Though such flexible recombination is adaptive for purposes of simulating novel events, the hypothesis also holds that this same process can contribute to memory distortions that result from miscombining elements of different experiences (for related experimental evidence, see Carpenter & Schacter, 2017, 2018).

The constructive episodic simulation hypothesis suggests that episodic memory retrieval plays an important role in various forms of cognition that rely on imagining specific situations. For example, episodic retrieval is hypothesized to contribute to planning steps to achieve a personal goal (autobiographical planning; e.g., Spreng et al., 2010), estimating one's response to a future event (affective forecasting; Gilbert & Wilson, 2007), and imagining alternatives to a specific past personal event (episodic counterfactual thinking; e.g., De Brigard et al., 2013). Consistent with this view, autobiographical planning engages the default network (Gerlach et al., 2011; Spreng et al., 2010; Spreng et al., 2015). Likewise, episodic counterfactual thinking elicits activity in many of the same brain regions as recalling the past does (De Brigard et al., 2013; Schacter et al., 2015). Such studies provide evidence consistent with the idea that episodic memory retrieval contributes to different forms of imagination.

While this work suggests that episodic memory contributes to these forms of imagination, we must be careful to avoid inferring this from default network activity alone. Default network activity is elicited by many different processes that likely do not all involve the same neural processes. To draw stronger conclusions about the contributions of episodic retrieval to different kinds of imagination, we can instead rely on research that manipulates the contributions of episodic retrieval. One procedure to manipulate episodic retrieval is called the Episodic Specificity Induction, which I will review in the next section, and which can provide this stronger evidence that episodic retrieval contributes to several kinds of mental simulations.

Episodic Specificity Induction: Identifying Contributions of Episodic Retrieval to Imagination

The constructive episodic simulation hypothesis states that we imagine future events in part by retrieving episodic details. To test this hypothesis, Madore et al. (2014) developed a manipulation to temporarily boost episodic retrieval. If imagining specific future events draws on episodic retrieval, the manipulation (when compared to the control manipulation) should enhance task performance. The procedure that was developed, known as the episodic specificity induction (ESI), has proven useful for identifying episodic retrieval contributions to a variety of tasks.

The ESI is adapted from the Cognitive Interview, which was designed to elicit detailed memories from eyewitnesses (Fisher & Geiselman, 1992). In the ESI procedure, participants are given a brief training in retrieving episodic details from a recent event. Participants first watch a brief video and are then asked to retrieve information about the surroundings and objects in the video, the appearance of individuals, and all the actions in chronological order. Following this procedure, participants perform the task of interest (e.g., imagining future events). The effect of this ESI on the subsequent task is then compared to the effect of a control induction, which in most experiments consists of an interview about the participant's general impressions of the video (for full interview scripts from Madore et al., 2014, see appendix).

The critical need for the ESI procedure is illustrated by earlier work on the relationship between episodic memory and imagination. For example, several experiments had indicated that older adults, who provide fewer episodic details than young adults when remembering past experiences, also provide fewer episodic details when imagining future experiences (Addis et al., 2010; Addis et al., 2008). However, a subsequent study showed that when asked to describe a picture – a task that should not involve episodic retrieval – older adults generated fewer details

that were physically present in the picture than younger adults (Gaesser et al., 2011). These findings suggest that the link between remembering past experiences and imagining future experiences could be at least partially explained by factors other than episodic retrieval, such as the manner in which people talk about their experiences in the present, past, or future. Studies that do not take account of such non-episodic influences are therefore inadequate for assessing the contributions of episodic retrieval to such cognitive tasks as future imagining because these non-episodic influences may also contribute to task performance. The ESI overcomes these limitations by manipulating episodic retrieval, thereby allowing researchers to assess the downstream impact of this manipulation on subsequent tasks.

Episodic retrieval contributes to future imagining: support for the constructive episodic simulation hypothesis

In the first study to develop and use the ESI (Madore et al., 2014), young and old adults were asked to imagine future events, remember past events, and to describe pictures. Madore et al. predicted that the two tasks hypothesized to rely on episodic retrieval – remembering the past and imagining the future – would benefit from the ESI (when compared to the control induction), while there would be no effect of the ESI on the non-episodic picture description task. Details on all three tasks were coded using procedures from the well-established Autobiographical Interview (Levine et al., 2002), which distinguishes between two types of details that people provide on autobiographical tasks: internal or episodic details (e.g., who, what, where, when) and external details (e.g., semantic details, off-task comments, and repetitive details). For the picture description task, internal details were defined as details physically present in the picture, and external details were the same as in the other tasks. Madore et al. (2014) predicted and found an interaction between induction type, detail type, and task: internal/episodic details were

selectively increased by the ESI for both young and old adults relative to the control induction when participants remembered past experiences and imagined future experiences, but not when they described pictures, and the number of external details did not differ between the two inductions on any of the three tasks. This pattern of results provides evidence that episodic retrieval contributes to remembering the past and imagining the future and is inconsistent with the hypothesis that ESI simply changes narrative style or the amount that participants talked. These findings are further bolstered by a subsequent experiment that yielded identical patterns of results using words rather than pictures to cue memory and imagination, and a non-episodic control task that required generating sentences and definitions in response to word cues (Madore & Schacter, 2016). Additional neuroimaging evidence suggests that the ESI impacts episodic retrieval processes in a future thinking task. The ESI (when compared to a control induction) increases recruitment of core network regions during a future imagination task relative to a control task (Madore, Szpunar, et al., 2016).

Taken together, these studies support the conclusion that the ESI serves as a tool to selectively manipulate the contributions of episodic retrieval to a cognitive task such as future imagining, which is not normally considered an "episodic memory task". Below, we'll discuss two additional areas that that are not traditionally thought of as memory tasks, but which may benefit from episodic retrieval.

Using the ESI to identify contributions of episodic retrieval to divergent creative thinking

Recent studies have suggested that other tasks involving mental simulation that would not ordinarily be considered "episodic memory tasks" nonetheless draw on episodic memory retrieval. For instance, divergent thinking, or the ability to combine old elements to generate creative new ideas, may be linked with episodic retrieval. Duff et al. (2013) reported that

hippocampal amnesic patients show decreased performance on a battery of divergent thinking tasks when compared to controls. In addition, individual differences in divergent thinking are correlated with differences in episodic detail generation for future events (though not past events; Addis et al., 2016). To provide even stronger evidence for a link between episodic retrieval and divergent thinking, recent studies have administered the ESI procedure prior to divergent thinking tests. In one study (Madore et al., 2015), participants were asked to generate novel alternative uses for everyday objects (AUT - Alternate Uses Test; Guilford, 1967), and in another study (Madore, Jing, & Schacter, 2016) they generated possible consequences of an unusual change in the world (e.g., living on without death) (Consequences Task; Torrance, 1962). The ESI, compared to a control induction, increased the number of appropriate alternate uses generated in the AUT and increased the number of appropriate consequences provided in the consequences task, thus providing evidence that episodic retrieval can contribute to divergent creative thinking.

A combined fMRI-ESI study further supports the conclusion that episodic retrieval contributes to divergent thinking. In this experiment, participants performed the AUT and a control task that required generating object associates in the scanner. Hippocampal activity selectively increased during the AUT after the ESI compared to the control induction (Madore, Thakral, et al., 2019). Consistent with this finding, a related study revealed common engagement of the hippocampus when participants performed the AUT, remembered past experiences, and imagined future experiences (Beaty, et al., 2018). Taken together with the previously reviewed behavioral evidence, these fMRI findings point to a role for episodic retrieval in divergent thinking.

Contributions of episodic retrieval to empathy

A growing line of work shows that episodic simulation may contribute to feelings of empathy. Gaesser and Schacter (2014) suggest that empathy benefits from episodic simulation when those we think about are not directly observable. For example, when we read a newspaper article and vividly imagine the situation of someone whose home was just flooded, we may better appreciate the strain they are under and feel more empathy towards them. To test these hypotheses, Gaesser and Schacter (2014) asked participants to imagine helping another person, after which the participants were asked to rate how likely they were to help that person. In a separate condition, participants were asked to remember an episode in which they helped another person before rating the likelihood of helping them. When compared to several control conditions, including generating comments about how the person in the situation could be helped, participants reported greater willingness to help after imagining helping or remembering an event in which they had helped others. Further supporting the link between episodic simulation and empathy, Gaesser and Schacter (2014) observed that the sensory vividness of these imagined or remembered scenarios correlated with the degree of helping intentions.

In a subsequent study, Gaesser et al. (2018) directly manipulated the vividness of scene imagery by asking participants to imagine the helping situation in either a familiar context or an unfamiliar context. They then observed helping intentions and helping behavior. As expected, familiar contexts led to more vivid imagery, which led to greater helping intentions. In addition, familiar contexts led participants to donate more money when they were given money to allocate to themselves and another person. Scene vividness in both familiar and unfamiliar context conditions was positively correlated with helping intentions and donation behavior. Again, these results suggest a role for episodic simulation in empathic behavior.

Importantly, this same experiment also examined the relationship between episodic simulation and the ability to imagine the thoughts and feelings of other people (also known as mentalizing). After each trial, participants rated the degree to which they took the perspective of the other person. In the familiar context condition, participants report greater perspective taking than in the unfamiliar context condition. Though limited by subjective ratings, these results are important because they indicate that scene imagery directly contributes to mentalizing. Gaesser (2020) suggests that this outcome further indicates that information from the imagined situation constrains what we think another person might be feeling.

Together, this body of work suggests that episodic simulation can be helpful for a variety of social cognitive tasks. While episodic simulation is not necessary for many of these tasks, as evidenced by amnesic individuals who are able to complete traditional theory of mind tasks (Rosenbaum et al., 2007), episodic retrieval may nonetheless play an important role in social cognition because it allows people to richly imagine other individuals in specific contexts.

Papers for the Dissertation

As we start to understand the contributions of episodic retrieval to domains that have not been studied in traditional episodic memory research, several challenges remain. First, more work needs to be done to identify domains of cognition that benefit from episodic retrieval. In Paper 1 and Paper 2, I will discuss research that examines the contributions of episodic retrieval to creative writing and mentalizing. These two domains are promising targets for study because existing research suggests (but does not causally test) that episodic retrieval contributes to creative writing and mentalizing. We know that episodic retrieval contributes to domain-general creativity (as measure by the AUT and the Consequences Task), but it is unclear whether episodic retrieval contributes to more naturalistic forms for creativity, such as creative writing.

And just as initial work suggests a role for episodic retrieval in creative writing, previous research suggests a role for episodic simulation in mentalizing. Gaesser and colleagues show that people rate themselves as considering the thoughts of another person more when imagining more coherent scenes. However, no published work has investigated whether episodic retrieval affects the amount of detail in the thoughts we attribute to other people.

A second challenge for the field is to determine how we can effectively conduct large studies on episodic retrieval and episodic simulation, given that the most frequently used procedure for estimating the contribution of episodic retrieval to various forms of imagination, the Autobiographical Interview (Levine et al., 2002), is highly time consuming because it involves manually counting up the details in memories and imagined events. For the third paper of my dissertation, I developed and validated software to automate components of the Autobiographical Interview scoring procedure. The aim for this third project is to allow researchers to conduct larger and more representative studies in the future.

Paper 1

Van Genugten, R.D.I., Beaty, R.E., Madore, K.P., & Schacter, D.L. (2021): Does Episodic Retrieval Contribute to Creative Writing? An Exploratory Study. *Creativity Research Journal*, DOI: 10.1080/10400419.2021.1976451

Abstract

Previous research indicates that episodic retrieval contributes to divergent creative thinking. However, this research has relied on standard laboratory tests of divergent creative thinking, such as generating creative uses for objects; it is unknown whether episodic retrieval also contributes to domain-specific forms of creativity. Here we start to explore whether episodic retrieval contributes to content generation on one such domain-specific task: creative writing. In two experiments, we used an episodic specificity induction (ESI) that selectively impacts tasks that draw on episodic retrieval. If episodic retrieval contributes to content generation during creative writing, then ESI should selectively increase the number of episodic details that people subsequently generate on a creative writing task. In our first experiment, we found evidence that ESI increased the number of episodic details participants generated. We observed a similar, though non-significant, trend in the second experiment. These findings constitute a starting point for examining the contribution of episodic retrieval to creative writing, but additional studies will be needed to more definitively characterize the nature and extent of these contributions.

Episodic memory allows individuals to recall and reconstruct their past experiences. Thinking about the past, however, is not the only function of episodic memory. A large body of research has shown that episodic retrieval also supports our ability to imagine future and other specific events. For example, many individuals with impaired episodic memory performance, including amnesic patients (e.g., Hassabis et al., 2007; Tulving, 1985; Race et al., 2011; but see Dede et al., 2016) and older adults (e.g., Addis et al., 2008; for review, see Schacter et al., 2018), have difficulty imagining specific events and novel scenes, and many brain regions involved in episodic retrieval comprise a *core brain network* (Schacter et al., 2007) that is also involved in imagining the future (e.g. Addis et al., 2007; Szpunar et al., 2007; for a meta-analysis, see Benoit & Schacter, 2015).

Several lines of evidence now suggest that participants may also rely on episodic retrieval when engaging in divergent creative thinking (for an overview, see Ditta & Storm 2018). Divergent creative thinking, or the ability to combine different types of information to generate novel ideas (Guilford, 1967), is a form of domain-general creative thinking. To respond to prompts in the Alternative Uses Task (AUT), a divergent creative thinking task in which participants provide alternative uses for everyday objects, participants sometimes report directly remembering alternative uses and invoking mental imagery to imagine uses for these objects (Gilhooly et al., 2007). Both direct retrieval and mental imagery can be supported by episodic retrieval. In addition, patients with episodic retrieval deficits as a result of hippocampal amnesia score lower on a battery of divergent creative thinking tasks when compared to controls (Duff et al., 2013). Further, scores on the AUT correlate with the number of episodic details that participants provide on a future imagination task (Addis et al., 2016). These findings contrast with other work on the contributions of memory to divergent creative thinking that has

emphasized the importance of searching for associations in semantic memory (e.g. Mednick, 1962; Kennett & Faust, 2020). According to these theories, semantic memory provides a base of general knowledge that supports creative solutions on the AUT that arise by combining multiple semantic concepts into new ideas. These theories emphasize the role of combining abstract concepts to support divergent thinking, whereas research on episodic retrieval suggests an additional role for retrieval of event-specific details. Importantly, these perspectives are not mutually exclusive, and recent work has examined the respective roles of both semantic and episodic processing during divergent creative thinking (Beaty et al, 2020).

While these studies suggest that episodic retrieval and divergent creative thinking are related, Madore et al. (2015) conducted a stronger test of the causal contributions of episodic retrieval to divergent creative thinking in a healthy population by using an Episodic Specificity Induction (ESI) to manipulate participants' reliance on episodic retrieval before they performed the AUT. ESI involves a brief training in detailed episodic memory retrieval and is based on the Cognitive Interview (Fisher & Geiselman, 1992), which was designed to improve eyewitness recall of autobiographical memories (for method, see description in the *Methods* section). In ESI experiments, researchers administer an ESI or control induction before the task of interest, then compare the performance on that task after the two inductions. If episodic retrieval contributes to the task immediately following the inductions, performance should be higher following ESI than following the control induction. If episodic retrieval does not contribute to the task, performance should be the same after an ESI and a control induction. A series of studies have demonstrated the efficacy of the ESI (for review, see Schacter & Madore, 2016). These studies have shown, for example, that the ESI impacts the generation of episodic details during episodic memory retrieval and episodic future simulation while having no impact on the number of non-episodic

details generated (Madore et al., 2014). In addition, the ESI does not have an effect on general retrieval and description tasks believed to be independent from episodic retrieval, such as retrieving semantic associates for objects (Madore et al., 2015), generating sentences with specific objects (Madore & Schacter, 2016) and describing pictures (Madore et al., 2014). Together, these studies suggest that the ESI can be used to identify tasks that rely on episodic retrieval, while having no effect on non-episodic tasks.

Madore et al. (2015) reported that participants who received an ESI (versus a control induction) subsequently generated more categories of appropriate object uses on the AUT. The ESI likewise increases the number of ideas that participants generate on a second divergent creative thinking task, the Consequences Task, which involves imagining novel implications of hypothetical scenarios (Madore, Jing, & Schacter, 2016). Neuroimaging results further indicate that episodic memory processes are involved when generating alternative uses. When participants complete the AUT, an ESI increases activity in memory-related brain regions when compared to the control induction (Madore, Thakral, et al., 2019). In addition, memory retrieval, future simulation, and the AUT all engage several regions in the aforementioned core brain network, including the hippocampus (Beaty et al., 2018). Finally, Thakral et al. (2020) recently showed that administering an inhibitory form of transcranial magnetic stimulation to the left angular gyrus, part of the core brain network, disrupted subsequent performance on the AUT and a future imagining task. Together, these studies demonstrate a strong link between episodic retrieval and the domain-general creativity that is assessed with tasks such as the AUT.

Despite this strong evidence that episodic retrieval contributes to content generation during *domain-general* creativity, the role of episodic retrieval in *domain-specific* creativity is unclear. Much of our creativity, such as musical improvisation, painting, and creative writing, is

domain-specific. Each of these activities draws on a specific skillset that is different from the others and is not fully dependent on domain-general creativity (Plucker & Beghetto, 2004). Whether episodic retrieval supports more naturalistic, domain-specific creativity—such as creative writing—remains unknown. To begin to explore this empirical gap in the literature, in two experiments we assessed whether manipulating episodic retrieval through ESI impacts performance on a subsequent creative writing task.

Some evidence already suggests that creative writing might benefit from episodic retrieval. Novels are often based on the autobiographical experiences and memories of the author; the writing of *Slaughterhouse Five*, for example, was based in part on author Kurt Vonnegut's experience as a prisoner of war detained in a slaughterhouse. Recent lab-based research has also started to explore the relationship between memory and creative writing. Van Tilburg et al. (2015) showed that retrieving a nostalgic memory before a writing task (when compared to retrieving a non-nostalgic memory) increased story creativity. While this study does not directly address whether episodic retrieval contributes to creative writing, it establishes that memory manipulations can affect performance on a creative writing task. In addition, neuroimaging evidence suggests that brain regions associated with episodic retrieval play a role in creative writing. Participants in an fMRI study showed greater hippocampal activation while writing a creative story than when copying a story (albeit with significance at a liberal statistical threshold; Shah et al., 2011). Together, these observations already suggest a link between episodic retrieval and creative writing.

In this paper, we expanded on these observations by formally testing whether episodic retrieval contributes to content generation during creative writing. We tested whether manipulating episodic retrieval via an ESI affects the number of details participants generate

when writing creatively. In both experiments, we adapted a paradigm previously used to study creative writing (Shah et al., 2011) and combined it with the ESI procedure. In this paradigm, participants were presented with excerpts of literature and were asked to continue writing the story they read. We compared performance on these stories after an ESI versus after a control induction, as assessed by the number of details participants produce.

Our specific predictions in these experiments are based on the constructive episodic simulation hypothesis. This hypothesis suggests that elements of episodic memories can be flexibly recombined into new imagined events and scenes (Schacter & Addis, 2007). For this reason, we predicted that boosting episodic retrieval via ESI would increase the number of episodic details, such as event-specific scene, person, and action details, while having no effect on the number of non-episodic details (such as factual background of the characters) in the creative writing stories. If the ESI impacts the number of episodic details in written stories, then we have evidence that episodic retrieval contributes to creative writing. If ESI also has no effect on non-episodic details, we can rule out the possibility that ESI broadly influences any type of detail in a generated story; that is, an effect selective to episodic details would suggest that the results are not attributable to participants simply trying to provide more information after ESI versus a control induction.

To further explore how episodic retrieval shapes creative writing stories, we scored these stories for originality as well. We did not expect to find a significant effect of ESI on originality, because existing research on the effect of ESI on creativity shows increases in the *amount* of original content produced, rather than increases in the originality of that content (Madore et al., 2015, 2016, 2019). For example, participants generate more appropriate categories of original uses on the Alternative Uses Task after the ESI (when compared to the control induction),

despite no significant differences in the originality of these uses (Madore et al., 2015). Based on this previous research, our primary hypothesis is that episodic retrieval contributes to the *quantity* of creative writing content produced, rather than the *originality* of that content.

Experiment 1 and Experiment 2 both test the hypotheses discussed above. Because effects of ESI on creative writing are compared to those of a control induction, we wanted to ensure that the result was not dependent on the specific control induction used. Thus, we used different control inductions in Experiments 1 and 2. We also increased our target sample size in Experiment 2 to improve our chances of finding the hypothesized effect.

Materials and Methods: Experiment 1

Procedure

Each participant in the experiment first completed an ESI or control induction. This procedure involved watching and then answering questions about a brief video. Following the induction, participants completed the creative writing task. In every trial of the task, participants were presented with the start of a story and asked to continue writing it on the computer (Shah et al., 2011). Participants were given six minutes to write each story, with five stories presented in this first segment of the experiment. Each person was then given a math filler task that involved adding and subtracting numbers for ten minutes (for similar procedures, see e.g. Madore & Schacter, 2016; Madore, Thakral, et al., 2019; Madore, Jing, & Schacter, 2019). This filler task focused participants' attention on a non-episodic task with the intent of decreasing potential carry-over effects of the induction. Following this filler task, participants underwent whichever induction they had not completed in the first portion of the experiment (ESI or impressions control induction). The induction order and video order were counterbalanced across

participants. After this second induction, participants wrote five more stories. Each session lasted approximately 2 hours. For a visual overview of the experimental procedure, see Figure 1. After the study was completed, details in the stories were counted and submitted to statistical analysis.

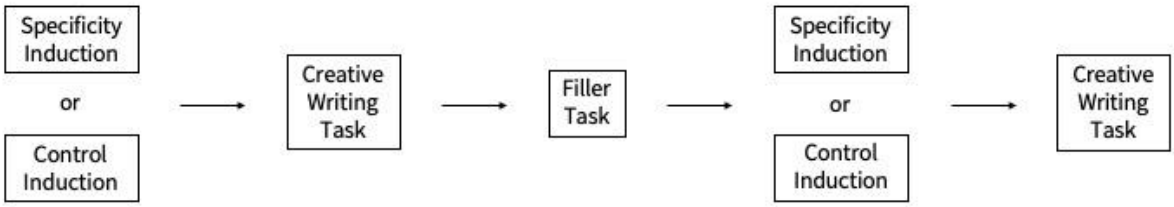

Figure 1. Experimental workflow. Each participant started with one induction (ESI or control) before they wrote creative stories. Participants then completed a ten-minute filler task, which served to decrease carryover effects of the first induction. After the delay task, each participant completed the induction they had not participated in before, then finished by writing a series of creative stories. Experiment 1 used the impressions control induction, whereas Experiment 2 used a math control task.

Participants

A sample size of 24 participants was chosen based on previous sample sizes of within-subject studies using the episodic specificity induction (e.g. Madore et al., 2014, 2015; Jing et al., 2016, 2017). Our sample was recruited from Harvard University and the community and was restricted to individuals between the ages of 18-30 with no neurological or psychiatric impairment at the time of the study. All participants provided written consent in accordance with

the ethics protocols approved by Harvard University's Institutional Review Board. Participants received course credit or payment for their participation.

25 participants were recruited, with 1 participant excluded for having already been in an ESI study. This led to our sample size of 24. One additional participant was removed during analysis for having copied sections of the original stories from the internet into their responses. Our final sample size included 23 individuals (mean age = 22.26 years old, SD = 4.07; 8 male, 15 female).

Tasks

Episodic Specificity Induction

The ESI is modeled after the Cognitive Interview, which is used to elicit detailed eyewitness memories (Fisher & Geiselman, 1992). In this procedure, participants first watch a short video, which later serves as material for memory retrieval. Immediately afterwards, participants complete three minutes of math problems. This filler task is designed to prevent participants from relying on working memory to answer subsequent questions and to prevent rehearsal. Participants are then asked questions about their memory of the video. The researcher instructs the participant to remember the video in as much detail as possible. The participant is then asked to tell the researcher everything they remember about the surroundings. After follow-up questions about the surroundings, participants are asked to describe everything they remember about the people in the video. After follow-up questions about the people in the video, participants are asked to describe the actions in the video in chronological order. This procedure has been shown to increase episodic output but not general verbosity on subsequent tasks in a series of experiments (e.g. Madore, et al., 2014, reviewed in Schacter & Madore 2016). Induction scripts can be found in Madore et al. (2014).

Impressions Control Induction

We compared the effect of ESI on creative writing to the effect of an impressions control induction. This impressions induction aims to control for participant engagement with the video and questioning while not increasing episodic retrieval. For this reason, the length of the impressions induction is approximately matched to the length of the ESI. To avoid episodic retrieval during the control induction, participants are asked to not provide specific details of what happened in the video. Instead, participants are asked to describe their general impressions of the video. Once participants provide their general thoughts and opinions of the video, a series of questions further probe their general impression of the video (e.g. "what adjectives would you use to describe the setting of the video?"). The full impressions control induction script can be found in Madore et al. (2014).

Creative Writing Task

In each trial of the creative writing task, participants read a passage from a work of literature and were instructed to continue writing the story. At the start of the experiment, participants were instructed to continue writing in the style that felt most comfortable to them, but to focus on writing as creatively as possible. Participants were further instructed to keep their stories somewhat realistic. After participants were finished with writing the stories, they were told that all prompts were based on published stories. For each prompt, they were asked if they recognized the story. If they did, they were also asked to write the story's name, author's name, or provide a sentence about the plot of the story. This allowed us to exclude any stories that participants were already familiar with before they started writing.

Story prompts were selected from the stimulus set used by Tamir et al. (2015). In their experiment, Tamir et al. presented participants with literary passages. Each of these passages was

characterized as social or non-social and vivid or abstract. For our experiment, we chose passages categorized by Tamir et al. as both social and highly vivid to promote participant engagement. We wanted to avoid story prompts that participants would recognize, as they might then complete the writing task by reciting the works of literature that the prompts came from. As a result, when we tested the task instructions for clarity in a separate online pilot sample, we additionally asked these pilot participants whether they recognized any prompts in an open-response question at the end of the study. We excluded story prompts that any pilot participant recognized. Ten stories from the remaining selection were then chosen for use in these experiments (see *Appendix* for story prompts). Stories were presented in a random order in Experiment 1 and assigned to lists that were then counterbalanced in Experiment 2.

Scoring

Scoring: Internal and External Details

To quantify the effect of the inductions on creative writing, the stories were scored for the number of episodic, or internal, details and the number of non-episodic, or external, details they contained. To do this, we used scoring procedures from the ESI studies of Madore et al. (2014) and Jing et al. (2016), which were adapted from the Autobiographical Interview (Levine et al., 2002). Internal details consisted of event details, including people, actions, objects, thoughts, emotions, locations and other similar details. External details consisted of factual, or semantic, details that do not contribute to a specific event. To illustrate internal and external detail scoring, we have included two annotated examples in the appendix. For more information about these detail categories, see Levine et al., (2002). In this study, external details consisted largely of backstory or non-perceptual descriptions of the situations or characters. While previous internal/external scoring procedures often require that all internal details belong to a single event

(e.g. Levine et al., 2002), many stories in our sample included multiple events and scenes. To avoid labeling episodic details as external, our scoring procedure only required that internal details belong to an event, rather than to the central event.

In Experiment 1, three raters obtained high interrater reliability for internal details (Cronbach's alpha = .91, assessed on 10 practice items) and external details (Cronbach's alpha = .87, assessed on 10 practice items). Items for assessing interrater reliability were taken from participants who were excluded from Experiment 1 for failing to attend the second study session (as such, their data were not included in the ESI analysis presented below). To ensure that scorers did not deviate from their training over the course of scoring, a randomly selected subset of stories were scored by two raters. Reliability remained high for internal details (Cronbach's alpha = .95, assessed on 10 stories from Experiment 1) and external details (Cronbach's alpha = .98, assessed on 10 stories from Experiment 1). All three raters were blind to the experimental condition (ESI vs. Control).

Scoring: Originality

Responses to each prompt were sorted into three equally large categories of low, medium, and high originality with high reliability (Cohen's Kappa with equal weights = .60, assessed on 62 practice items) by two raters who were blind to the experimental condition (ESI vs. Control). The stories used to determine inter-rater reliability between the two raters were taken from an online pilot. This pilot was conducted to ensure that participants understood task instructions and did not recognize the prompts. This pilot contained no induction procedures.

Guidelines for scoring originality of these stories were derived from existing subjective scoring guidelines for judging creativity of responses to the alternative uses task (Silvia & Benedek, 2019). These guidelines suggest that the creativity of a response can be assessed along

three dimensions: how common the response is, how remote a response is (or how different it is from the everyday), and how clever the response is (cf., Silvia et al., 2008).

Accordingly, stories that reiterate the writing prompt or continue the story as would be expected based on the prompt (i.e. responses that were not remote from the story prompt) and responses that were similar to many other responses (i.e. common) were assigned to the lowest originality group. Stories that stood out as highly original were then assigned to the highest originality group. These stories were often identified by being appropriate but quite different from other responses in the sample (i.e. uncommon). Stories that did not simply continue the writing as would be expected based on the prompt (i.e. remote) were also more likely to be assigned to this group. Stories that were clever, regardless of whether the topic was remote or common, were often also assigned to this group.

Remaining stories that were not as easily placed into the low or high originality bins were then ordered according to ascending subjective originality. These stories where then split among the low, medium, and high originality groups such that each group had an equal number of stories. An example story of each level of originality is presented in the *Appendix*.

This approach to scoring originality differs from typical methods used to assess creativity in AUT responses. We rated originality on a categorical 3-point scale, while AUT responses are commonly rated on a continuous 5-point scale (e.g. Silvia et al., 2008). Preliminary attempts by two raters to use a standard 5-point scale yielded low levels of agreement between raters—which is often the case with subjective creativity scoring (Forthmann et al., 2017)—resulting in our decision to use the categorical 3-point scale described above (i.e., low, medium, high originality), which can reduce ambiguity and improve rater agreement (Benedek et al., 2013), as was the case in our study.

Statistical Analyses

If episodic retrieval contributes to the generation of event details during creative writing, we would expect a significant effect of ESI on internal details when compared to a control induction. By contrast, we would expect no effect of ESI on external details.

We evaluated this hypothesis by testing whether the number of details in the stories differed based on an interaction between detail type (internal/external) and induction type (control/ESI). In addition to these fixed effects, random effects were included to account for possible individual differences in writing ability and style and differences in story prompts. Specifically, we added random intercepts for the interaction of participant number and detail type as well as random intercepts for the interaction of story prompt and detail type. The first random effect captures both variation in the amount that participants write as well as differences in their baseline use of internal and external details. The second random effect similarly accounts for differences in the lengths of stories and the number of internal and external details that particular story prompts elicit. This multi-level mixed model was implemented using the function *lmer* from the R package *lmerTest* (Kuznetsova et al., 2017).

After testing for an interaction of induction type and detail type, follow-up tests evaluated the effect of ESI (versus control) on internal details, and the effect of ESI (versus control) on external details. These regression models additionally contained random intercepts for participant identities and story prompts.

In addition to testing for the effect of ESI on internal and external details, we also tested for the effect of ESI on originality ratings. We used an ordinal regression that included participant ID as a random effect and word count as a covariate. The ordinal regression was implemented

using the function *clmm* with flexible thresholds from the R package *ordinal* (Christensen, 2015). The random effect for participant ID was included to account for differences in participants' abilities to generate creative stories. Since originality ratings were not independent from story word count, we wanted to ensure that any effect of the ESI on ratings could not be explained by a simple increase in story length. For this reason, we included word count as a covariate in our model. No random effect of story prompt is necessary because each story prompt has an equal number of low, medium, and high originality ratings as a result of the scoring procedure. Therefore, no variability in ratings can be attributed to prompt number.

Results and Discussion: Experiment 1

We found a significant interaction between detail type and induction type as hypothesized (b = 6.53, t(395.80) = 2.06, p = .040). Follow-up tests indicate that stories included more internal details after the ESI than after the control induction (b = 5.88, t(197.99) = 2.66, p = .009), but no significant difference was found for external details (b = -0.37, t(200.04) = -0.16, p = .87). The number of internal and external details (averaged across prompts and participants) are displayed as a function of induction type in Figure 2. The mean number of internal details in stories following ESI was 29.29 (SD = 20.76), whereas the mean number of internal details in stories following the control induction was 22.96 (SD = 18.15). The mean number of external details in stories following ESI was 19.23 (SD = 18.11), whereas the mean number of external details in stories following the control induction was 19.54 (SD = 18.10).

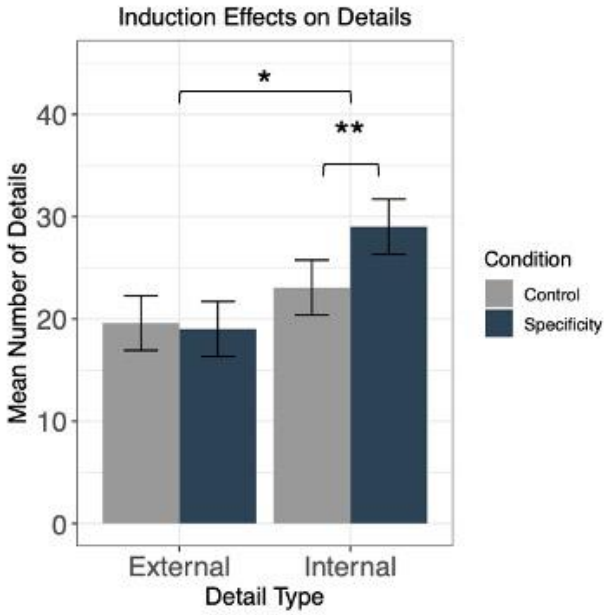

Figure 2. Experiment 1: Effects of an episodic specificity induction and impressions control induction on the mean number of details that participants include in stories. The displayed number of details result from averaging across prompts and participants. Error bars represent 1 SE. The largest grouping line indicates that there is a significant interaction effect: the difference in number of details following ESI versus control is greater for internal details than for external details. The smaller grouping line suggests that there is a significant effect of ESI on internal details, relative to the control induction. * $p < .05$, ** $p < .01$, *** $p < .001$

We found no significant effect of ESI (relative to control) on the originality of responses (proportional odds ratio = 0.91, $p = .74$), using the ordinal regression model described in *Statistical Analyses*. The number of words in each story was a significant predictor of originality ratings (proportional odds ratio: 1.02, $p < .001$), consistent with past work reporting positive

associations between word count/elaboration and creativity ratings (Beaty & Johnson, 2021; Forthmann et al., 2019).

To summarize, in Experiment 1, we found that ESI selectively increased the number of internal details in creative writing, relative to an impressions control induction. These results provide preliminary evidence that episodic retrieval contributes to the amount of content individuals generate during creative writing.

Overview: Experiment 2

To further investigate the role of episodic retrieval in creative writing, we modified our experimental design in two ways for Experiment 2. First, we replaced our impressions control induction with a different control task. Second, we increased our target sample size from 24 to 32 participant to increase power. We chose this sample size to be consistent with previous ESI studies, which typically use either 24 or 32 participants (e.g. Madore et al., 2016; Madore, Thakral et al., 2019; Madore & Schacter, 2016).

We replaced the impressions control induction with a math control task to exclude an alternative explanation for the results in Experiment 1. The results in Experiment 1 are based on the contrast of ESI to the impressions control induction. A positive result, then, could be driven by a boost in internal details caused by ESI. Alternatively, a positive result could arise from a decrease in internal details as a result of the impression induction. The former helps us understand the link between episodic memory and creative writing, as it suggests that episodic retrieval contributes to our task; the latter does not provide evidence for this link. To exclude this latter possibility, we test whether the effect of ESI persists with a different control task. Based on previous research, which has not found a difference between the impressions and math control

inductions (e.g., Madore, et al, 2014, 2015; Madore & Schacter 2016), we likewise expected no differences between control conditions, as both likely require little episodic retrieval.

Materials and Methods: Experiment 2

Procedure

Each participant attended a single session that lasted approximately 2 hours. Participants received either the math control or the ESI first, counterbalanced across participants. In the math control task, participants were given a series of addition and subtraction problems and were not asked questions about the video. The math control task was approximately matched in time to the ESI.

After the first induction, participants were given six minutes per story to write five stories. As in experiment one, participants were then asked to complete a filler task for ten minutes to prevent carry-over effects of the induction. After the filler task, participants completed the second induction (whichever one they had not completed in the first segment) before writing the remaining five stories. This procedure is depicted in Figure 1.

Participants

Thirty-two participants were recruited for this experiment with the same guidelines used for recruiting in Experiment 1. Three participants were then excluded for failure to write all of the stories. To meet our target sample size of 32 participants, three additional participants were recruited. Participants in this sample were 18-30 years old (M = 24.03 years, SD = 3.51; 21 female, 11 male).

Scoring

Internal and external details were scored with the same procedure as in Experiment 1. As before, raters were blind to the experimental condition of the stories. Two raters who had previously scored responses in Experiment 1 also scored responses for Experiment 2. These two raters obtained excellent interrater reliability for the internal (Cronbach's alpha = .94, assessed on 10 practice items) and external details (Cronbach's alpha = .92, assessed on 10 practice items). Interrater reliability was calculated from the same items used to establish reliability in Experiment 1. After scoring was completed, we checked whether scorers deviated from their training while scoring. On a random sample of ten stories from this experiment, reliability for internal (Cronbach's alpha = .93) and external details (Cronbach's alpha = .90) remained high.

Stories in Experiment 2 were additionally scored for originality by two raters with the same procedure as in Experiment 1. Both raters were blind to the experimental condition (ESI or Control) of each story. One of these raters also scored responses for originality in Experiment 1. Responses to each cue were sorted into groups of low, medium, or high originality. Interrater reliability was high, and was assessed prior to scoring on the same 62 practice items as in Experiment 1 (Cohen's Kappa with equal weights = .617).

Results

In Experiment 2, we adopted the same mixed-model approach and observed the same general trends for an effect of ESI on the key detail measures, but the trends failed to reach standard levels of statistical significance. Thus the interaction between detail type and induction type approached but did not attain statistical significance ($b = 6.6$, $t(550.7) = 1.80$, $p = .072$). Similarly, follow-up tests indicate that stories did not include significantly more internal details after the ESI than after the control induction ($b = 4.37$, $t(275.36) = 1.69$, $p = .092$), though the trend was in the same direction as Experiment 1. No significant difference as a function of

induction was found for external details (b = -2.21, t(275.35) = -0.85, p = .394), in line with Experiment 1. Results are displayed in Figure 3. The mean number of internal details in stories following ESI was 33.29 (SD = 29.17), whereas the mean number of internal details in stories following the control induction was 28.52 (SD = 28.63). The mean number of external details in stories following ESI was 30.34 (SD = 28.91), whereas the mean number of external details in stories following control induction was 32.79 (SD = 28.83).

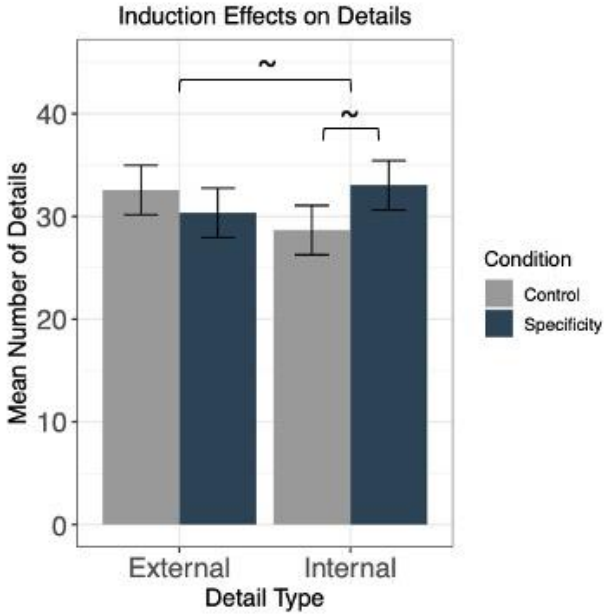

Figure 3. Experiment 2: Effects of an episodic specificity induction and math control on the mean number of details that participants include in stories. The displayed number of details result from averaging across prompts and participants. Error bars represent 1 SE. The largest grouping line indicates that there is a marginally significant interaction effect: the difference in number of details following ESI versus control is numerically greater for internal details than for external details. The smaller grouping line shows that there is a marginally significant effect of

(Continued) ESI on internal details, relative to the control induction. ~ p < .1, * p < .05, ** p < .01, *** p <.001

In addition, we found no significant effect of ESI on originality ratings (proportional odds ratio: 1.41, p =.12) when modeled with an ordinal regression that included participant ID as a random effect and word count as a covariate. Word count was a significant predictor of originality ratings in this model (proportional odds ratio: 1.01, p = .009), as it was in Experiment 1.

Finally, in an exploratory analysis, we combined data from Experiments 1 and 2 to test for the effect of ESI on creative writing relative to the two control conditions. The model for this analysis was specified in the same way as above, with the addition of a fixed effect of experiment number. We found a significant interaction between detail type and induction type (b = 6.51, t(957.7) = 2.57, p =.011). Follow-up tests indicated that stories included more internal details after the ESI than after the control induction (b = 4.94 , t(478.50) =.75, p = 0.006), but no significant difference was found for external details (b = -1.57, t(479.30) = -0.87 , p = 0.384). The mean number of internal details for the stories was 31.67 (SD = 26.11) following the ESI, and 26.19 (SD = 24.90) after the control induction. The mean number of external details was 25.82 (SD = 25.59) after the ESI, and 27.22 (SD = 25.75) after the control induction. Figure 4 displays induction effects on internal and external details with data combined across Experiments 1 and 2.

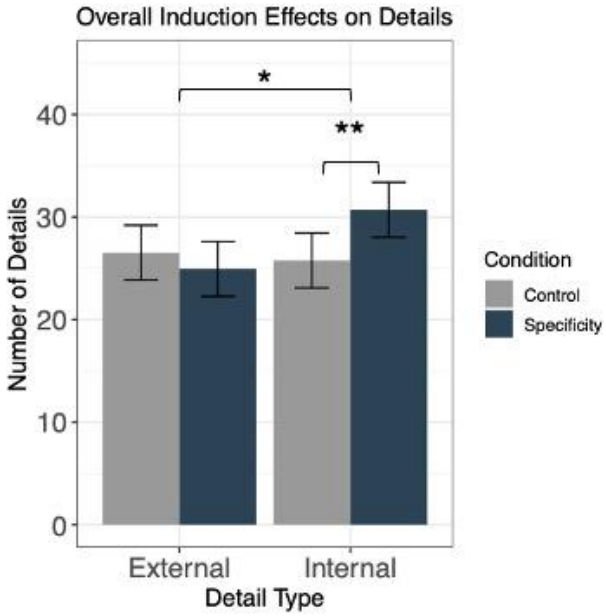

Figure 4. Experiment 1 & 2: Effects of an episodic specificity induction and control inductions on the mean number of details that participants include in stories. The displayed number of details result from averaging across prompts, participants, and Experiment 1 and 2. Error bars represent 1 SE. The largest grouping line indicates that there is a significant interaction effect: the difference in number of details following ESI versus control is greater for internal details than for external details. The smaller grouping line indicates that there is a significant effect of ESI on internal details, relative to the control induction. * $p < .05$, ** $p < .01$, *** $p < .001$

Discussion

Despite clear evidence that episodic retrieval contributes to laboratory measures of divergent creative thinking, the role of episodic retrieval in naturalistic creative tasks like creative writing is less clear. Experiment 1 suggests a role of episodic retrieval in creative

writing. Participants in this experiment included significantly more internal details in their stories following the ESI than following the control, while no change in external details was observed, as hypothesized. These findings, if considered in isolation, would suggest that episodic retrieval plays a role in the generation of event details for creative stories. These results would also suggest that episodic retrieval does not play a significant role in generating non-internal details, such as factual background of the characters, during creative writing. However, Experiment 2 does not strongly support these conclusions. While the effect of ESI on detail generation followed the same trend as in Experiment 1, the effect was not statistically significant. Nonetheless, the combined analysis of Experiments 1 and 2 did reveal a significant effect of ESI on internal details in creative writing. To summarize, the two experiments discussed here provide suggestive, but not conclusive evidence that episodic retrieval contributes to generating event and scene details of creative stories.

In addition, these results do not provide evidence that episodic retrieval impacts the originality of written stories. We found no significant effect of ESI on originality ratings in either Experiment 1 or Experiment 2. These results are consistent with similar findings in divergent creative thinking studies, which show an increase in quantity, rather than originality, of creative content following ESI. Previous studies show that ESI has no effect on originality ratings in the AUT, despite increasing the number of original items produced on this task (Madore et al., 2015, 2016, 2019).

The mixed findings that we report here may result from low statistical power due to large variability in the number of external and internal details across stories. Some responses consisted entirely of factual background and other external details, whereas other stories contained only

episodic details. This large variability could have made it difficult to detect an effect of ESI in these experiments.

To test whether low power is a possible explanation of our null result, we conducted a power analysis based on resampling data from the first experiment. We drew random samples of different sizes with replacement from the data and calculated the proportion of times that the interaction effect included in the model was significant for each sample size. This bootstrapping procedure suggested that that our second study's power was .55, and we should have included 72 participants to have an 80% chance of finding an effect if it exists. The second experiment, then, was underpowered to find an effect.

This problem could be addressed in future experiments by including larger sample sizes. Researchers could also alter the prompts and instructions used in these experiments to reduce variability in story responses. Our preliminary results indicate that episodic retrieval may aid an author in generating event-specific contextual details in a story, so instructions that focus on writing specific events may make an effect of ESI easier to detect. In other words, researchers could revise instructions so that participants are asked to write a creative story focused on a specific event. This, or similar changes, could reduce response variability and increase researchers' ability to detect contributions of episodic retrieval to creative writing. In addition, future researchers can ensure that instructions do not discourage detail generation. To do so, the instructions that were used in these two experiments, which encouraged participants to be as creative as possible, could be modified. Instructing participants to "be creative" decreases the amount of content generated in the AUT, relative to instructions that emphasize the quantity of output ("be fluent"; Nusbaum et al., 2014), so similar instructions in this paradigm might likewise have decreased output. Instead, participants could be told to generate a creative story

with as much detail as possible. Such instructions would require creative responses but would emphasize the quantity of output (for similar instructions emphasizing both quantity and creativity, see e.g. Madore et al., 2016).

To summarize, when researchers seek to investigate the role of episodic retrieval in domain-specific creativity, our results suggest that creative writing remains a promising target of study. Creative writing also remains a promising target of study because previous work suggests that episodic retrieval plays a central role in imagining specific events (e.g. Schacter & Addis, 2007), which is critical to writing new stories. According to the constructive episodic simulation hypothesis, we imagine specific scenes and events by recombining details from episodic memories (e.g. Schacter et al., 2012; Schacter & Addis, 2007, 2020). Episodic retrieval, then, may allow an author to generate relevant event and scene details as they write their story.

Other literature suggests that an interplay between episodic and semantic memory allows us to imagine future events (the semantic scaffolding hypothesis; e.g. Irish et al., 2012; Irish & Piguet, 2013). Specifically, semantic memory retrieval may be used to build a scaffold of general knowledge and schematic information that can then be filled in by episodic details. For example, to imagine a day at the beach, we may use general semantic information to frame the event (e.g. 'I usually go with a group of four friends, so I'll probably go with them for this trip as well') and then retrieve specific episodic details to develop the event (e.g. 'I can see an ice cream truck parked on the boardwalk'). These processes used to imagine future events may be involved in imagining new events for creative purposes, such as creative writing, as well. As a result, episodic retrieval may be involved in generating much of the content of creative writing stories, while semantic retrieval or other processes may be used to generate the creative idea or story arc that is expressed in the narrative.

Our experiments represent a first attempt to explore the contribution of episodic retrieval to creative writing, but future research could focus on additional forms of domain-specific creativity, such as the design of scenes in theater and film, that seem to benefit from the imagination of specific events. Episodic memory contributes to many tasks that are not traditionally thought of as memory tasks, and several forms of domain-specific creativity might benefit from its contributions.

Paper 1 Acknowledgements

We thank Andrew Rao for his help in data collection. We also thank Andrew Rao and Ethan Harris for their help with scoring. We want to thank Helen Jing for her guidance on how to score the responses. This research was funded by the National Institute on Aging Grant R01 AG008441 awarded to DLS. R.E.B. is supported by a grant from the National Science Foundation [DRL-1920653]. KPM is supported by the National Institute on Aging Grant F32 AG059341.

Paper 2

van Genugten, R.D.I., & Schacter, D.L. (in prep): Does Episodic Retrieval Contribute to Mentalizing?

Abstract

We regularly imagine how others feel in situations that we do not observe (for example, when messaging a friend about something that happened to them). Previous research suggests that episodic retrieval, which supports event simulation, impacts our empathy for individuals as we imagine their experiences. In this paper, we examined whether episodic retrieval also impacts our ability to infer the thoughts of individuals (i.e., mentalize) in imagined events. First, we manipulated episodic retrieval through a procedure known as the Episodic Specificity Induction (ESI). We compared the number of details participants provided on a mentalizing task after ESI and after a control induction. We hypothesized that ESI would affect the amount of mental state inference details, which would serve as evidence that episodic retrieval contributes to mentalizing. In Experiment 1, we found robust support for this hypothesis. In Experiment 2, we found no effect of ESI on mentalizing. We failed to find an effect of ESI on our manipulation check as well, making it difficult to interpret ESI results from this experiment. Second, we examined whether episodic retrieval contributes to mentalizing by asking participants to label which inferences were informed by the retrieval of particular memories in Experiment 2. We found that more than half of all mental state inferences were accompanied by and informed by the retrieval of specific memories. Together, these experiments suggest that episodic retrieval may contribute to mentalizing.

A growing body of research suggests that remembering the past is not the only function of episodic memory. Instead, the constructive process that allows us to retrieve and combine details from a memory during remembering can be used for constructing imagined scenes and events as well (e.g. Addis et al., 2007). For example, patients with hippocampal damage have difficulty imagining specific scenes (Hassabis et al., 2007), and the brain regions involved in remembering the past are also involved in imagining the future (Benoit & Schacter, 2015). Transcranial magnetic stimulation (TMS) applied to a brain region involved in episodic retrieval likewise impacts future imagining (Thakral et al., 2017), and experimental manipulations of episodic retrieval affect future imagining as well (Madore et al., 2014). Motivated by the strong associations found between episodic retrieval and imagining future events, Schacter and Addis (2007, 2020) proposed the constructive episodic simulation hypothesis, which states that we imagine future and other specific events by flexibly recombining details from our episodic memories.

Because episodic memory contributes to our ability to imagine specific events, this framework suggests that episodic retrieval may also contribute to other related tasks that involve imagined scenes and events. Indeed, recent work has shown that episodic retrieval contributes to various domains of cognition that involve imagined events, yet are not traditionally thought of as relying on memory. Domains such as means-end problem solving (Madore & Schacter, 2014; Sheldon et al., 2011), event reappraisal (Jing et al., 2016), and creative thinking and writing (Madore et al., 2015; van Genugten et al., 2021) can benefit from episodic retrieval when participants imagine specific situations while solving the task. Much work remains to be done to determine what other areas of cognition benefit from episodic retrieval during imagination. Because people regularly imagine specific content when imagining social situations (e.g.,

Andrews-Hanna et al., 2013), episodic retrieval may also contribute to our understanding of these situations as well.

Recent work by Gaesser and colleagues suggests that participants who engage in scene or event simulation may experience more empathy while thinking about others. For example, Gaesser & Schacter (2014) asked participants to read a series of stories about individuals in difficult situations. Participants who imagined helping that person or remembered a related event in which they helped someone were more likely to indicate that they would help the person in the story. Importantly, sensory vividness of the imagined events correlated with helping intentions, suggesting that scene construction or imagery plays an important role in increasing helping intentions. To examine whether the effect of scene construction is limited to helping intentions, Gaesser et al. (2018) tested whether engaging in scene construction affected donation behavior. On each trial, participants were asked if they wanted to donate part of their cash bonus to people in a situation similar to the scenario they imagined. Scene imagery was manipulated by asking participants to imagine the situation in a familiar or unfamiliar location. Participants were willing to donate more after imagining the helping situation in a familiar location. Together, these results suggest that scene construction and imagery contribute to empathy judgments.

To directly test whether episodic retrieval impacts empathy judgments, Vollberg et al. (2021) used the Episodic Specificity Induction (ESI) to manipulate episodic retrieval before participants made empathy judgments. The ESI is a brief training in retrieving a past memory. When compared to a control induction, ESI increases the number of central episodic details (also known as internal details) retrieved on subsequent tasks, without affecting other details (known as external details). For instance, when participants imagine future events, ESI (relative to the control induction) boosts the number of internal but not external details (e.g. Madore et al.,

2014). Importantly, the ESI does not simply increase the amount that participants talk, since it has no effect on several tasks that do not rely on episodic retrieval, such as describing the content of pictures (reviewed in Schacter & Madore, 2016).

In their experiment, Vollberg et al. asked participants to read about a series of negative situations involving another person (e.g. "Eric had a stomach ache after lunch"), then asked participants how bad they felt that the situation happened. Vollberg et al. found that ESI boosted empathy felt for both in-group and out-group members. The effect of ESI on empathy was mediated by the number of central episodic details that participants used to imagine the situations. In an online version of the experiment, Vollberg et al. used an approach similar to the ESI to manipulate episodic retrieval (Rudoy et al., 2009), which also led to increased empathy relative to a control induction. These results provide further evidence that manipulating episodic retrieval affects empathy judgments. Importantly, it highlights an approach for testing whether episodic retrieval contributes to aspects of social cognition (through the ESI) and leaves open the question of whether other forms of social cognition can benefit from episodic retrieval.

Initial evidence that episodic retrieval may contribute to mentalizing as well comes from behavioral and neuroimaging studies. For example, Krienen et al. (2010) asked participants to imagine how strangers or specific friends would respond to preference judgments (e.g. 'prefer window seats to aisle seats when flying') on a 4-point scale. For trials with similar and dissimilar strangers, participants rarely agreed with the statement "I relied on a particular memory or anecdote" (mean rating of 1.65 on a 7-point scale). However, for trials with specific friends, participants report relying on a particular memory often (mean rating of 4.37 on a 7-point scale). These results suggest that participants draw on relevant memories when they have shared experiences with the person they are making preference judgments about. These results are

consistent with informal reports from participants who stated they recalled related events from their own lives during an empathy tasks (Rameson et al., 2012).

Additional evidence for a link between episodic memory and mentalizing comes from neuroimaging studies, which suggest that regions traditionally associated with memory retrieval are often co-activated with regions implicated in mentalizing. Earlier neuroimaging work suggested that a shared network was responsible for mentalizing, future thinking, and episodic retrieval (e.g. Buckner & Carroll, 2007; Mitchell, 2009; Spreng et al., 2009; Spreng & Grady, 2010). Research focused on repeatedly scanning individual subjects has since shown that the default network fractionates into two subsystems (Braga & Buckner, 2017), with one subsystem responsible for episodic retrieval and one subsystem responsible for mentalizing (DiNicola et al., 2020). These studies suggest that previous studies found a single default network responsible for episodic retrieval and mentalizing because two interdigitated networks were regularly blurred together because of group-averaging of scans during analysis. This work, which shows separable systems for mentalizing and episodic retrieval, is consistent with research on amnesic individuals, which shows that theory of mind capabilities remain intact in individuals with severe deficits in episodic memory (Rosenbaum et al., 2007), as well as theoretical analysis arguing that the processes used during mentalizing are not shared with other domains (e.g. Mitchell, 2011). Nevertheless, the close relationship between two default subnetworks that support episodic retrieval and mentalizing (DiNicola et al., 2020), as well as the coactivation of memory and mentalizing regions during tasks such as autobiographical memory retrieval (e.g. Andrews-Hanna et al., 2014) and imagining how individuals feel in specific events (e.g. Rabin et al., 2010), suggests that these two systems regularly interact during different tasks. This in turn

leaves open the question of whether interactions between the episodic retrieval and mentalizing systems are functionally useful: can episodic retrieval contribute to mentalizing?

To explain these findings and discuss the potential role that episodic retrieval plays in mentalizing, Gaesser (2020) proposed that the scenes and events that we imagine can inform the mental state judgments we make about people in those imagined events. In other words, imagined context guides mental state inference. Mental state judgments may also be informed by the retrieval of related episodic memories (e.g., a participant may spontaneously remember their own experience losing a soccer game when they are trying to imagine how someone would feel after losing in a basketball game). Through both episodic simulation and retrieval of related memories, episodic retrieval may impact mentalizing and empathy because it allows us to better understand the situations others are in (Gaesser, 2020). Early work on the link between episodic retrieval and mentalizing has been conducted by Gaesser et al. (2018), who showed that manipulating scene imagery (imagining the situation in a familiar versus unfamiliar location) increased participant ratings on the question "When you identified media or imagined helping, did you consider the person's thoughts and feelings? 1 = not at all – 7 = strongly considered". In addition, these mentalizing ratings correlated with ratings of scene vividness. These data provide preliminary evidence that episodic retrieval can contribute to mentalizing.

In this paper, we present two experiments that test directly whether episodic retrieval contributes to mentalizing. In both experiments, we used the ESI to manipulate episodic retrieval before participants completed a mentalizing task. In this task, participants were asked to list the possible thoughts, feelings, and intentions of a hypothetical friend in a series of situations. Performance on the mentalizing task was compared after ESI and after control inductions. We predicted that reported mental states would be more detailed after the ESI (when compared to the

control induction), which would indicate that episodic retrieval contributes to mental state inference. Importantly, we predicted that the ESI would not affect the generation of other details that participants provide on this task, such as repeated information, off-task commentary, and additional information about the character in story (e.g., 'he's probably walking over there'). That is, we predicted a specific effect of ESI on mentalizing, rather than a general effect of ESI on verbosity. After the mentalizing task, participants completed an episodic simulation task, which involved imagining specific future events. In light of previous studies showing that ESI increases the number of episodic details that participants provide when imagining future experiences (e.g., Madore et al., 2014; Madore & Schacter, 2016), this task served as a manipulation check to ensure that the ESI worked as expected. We predicted that the ESI would affect the number of internal but not external details in the responses from the episodic simulation task.

In Experiment 2, we included a second approach for identifying episodic contributions to mentalizing. At the end of the experiment, we asked participants to label the thoughts they had provided. Thoughts were labeled as 'old' if the participant explicitly remembered having drawn on a memory to provide the answer. Thoughts were labeled as 'new' if the participant did not remember having relied on a specific memory to respond. These labels allowed us to examine how often a mental state inference was informed by a specific memory that was retrieved during the task. We hypothesized that a substantial number of mentalizing judgments would be informed by particular memories. These labels also allowed us to test whether ESI differently affects 'new' or 'old' thoughts. Gaesser (2020) suggests that episodic retrieval can contribute to mentalizing through retrieval of memories of oneself in similar situations (corresponding to our

'old' label), and through novel event construction (which would result in a 'new' label). As a result, these labels could provide insight into the specific ways that ESI impacts mentalizing.

We made two additional changes to Experiment 2 for practical reasons. We replaced the impressions control induction (to which the ESI is compared) with a different control task to ensure that our findings were not dependent on the specific control induction used. We also used a one-session study protocol (adapted from e.g., Madore et al., 2016; Madore, Jing, et al., 2019; Madore, Thakral, et al., 2019) rather than a two-session protocol to reduce participant dropout. The above-noted studies have revealed significant ESI effects on a future imagining task under the conditions used in Experiment 2.

To preview our results, we found some evidence in both experiments that episodic retrieval contributes to mentalizing. In Experiment 1, we found that ESI impacts the number of details in mental state inferences. In Experiment 2, we failed to replicate the ESI effects from Experiment 1, but cannot interpret the ESI results because of a failed manipulation check. However, participants in Experiment 2 labeled more than half of the thoughts they provided in the experiment as informed by the retrieval of specific memories, in line with hypotheses.

Materials and Methods: Experiment 1

Procedure

Modeled after previous studies by Madore and colleagues (e.g., Jing et al., 2016; Madore et al., 2014, 2016; Madore & Schacter, 2014; for review, see Schacter & Madore, 2016), participants participated in two laboratory sessions that were held seven to nine days apart in order to minimize possible carryover effects from one induction to another. In each session, the participant first received a specificity or control induction. Then, participants completed the

mentalizing task, in which they listed possible thoughts, feelings, and intentions of a hypothetical person in a series of situations. After completing this task, participants continued with an episodic simulation task. In it, they were asked to imagine specific future events in response to a series of prompts. The induction type and the videos used for the inductions were counterbalanced across sessions, as were the lists of stimuli used in the mentalizing and episodic simulation tasks. Induction procedures were the same as in Paper 1 of the dissertation.

Participants

In accordance with previous sample sizes reported in within-person ESI studies, we set our target sample size at 32 participants. 45 participants attended the first session, with 32 attending both sessions. One additional participant was excluded for having previously participated in an ESI experiment. Data from participants who completed only one session were not used in our analyses, except for establishing interrater reliability. Participants were recruited from Harvard University and the surrounding community and were paid for their participation or were granted course credit. Participant were aged 18-30, with no history of neurological or psychiatric impairment. Participants were on average 22.31 years old (SD = 3.48) and included 6 male and 24 female individuals. Two other individuals declined to indicate their sex. All participants provided written consent in accordance with ethics protocols approved by Harvard University's Institutional Review Board.

Tasks

Episodic Specificity Induction

The ESI is an interview procedure designed to increase participants' reliance on episodic retrieval in subsequent tasks. In this procedure, participants first watch a short video about two actors in a kitchen. Participants then complete a series of math questions for three minutes. This

filler task is designed to prevent rehearsal and ensure that participants do not answer subsequent questions by relying on working memory. After the math questions, the participant is asked to remember the surroundings of the video in as much detail as possible. After follow-up questions about the surroundings, participants are asked to remember the appearance of the individuals in the video. Participants are then asked to report all actions in the video in chronological order. The full script for induction procedures can be found in Madore et al. (2014).

Impressions Control Induction

The impressions control induction is an interview procedure intended to serve as a control condition to contrast with the ESI. Like the ESI, the impressions control interview involves asking participants about the video that they recently watched. Unlike the ESI, the impressions interview involves asking participants their general feelings about the video. The impressions control induction is roughly matched in time to the ESI. Like the ESI, scripts for this procedure can be found in Madore et al. (2014).

Mentalizing Task

In the mentalizing task, participants were asked to list all possible thoughts, feelings, and intentions of the person in the situation described in the prompt. Participants completed a series of five mentalizing trials after each induction and were given three minutes to respond to each prompt. To ensure that participants were able to engage with the material for three minutes, prompts were predominantly about atypical experiences (e.g., how someone might feel as they enter their first ever boxing match), or emotional situations (e.g., how someone might feel before their wedding). Throughout the experiment, pronouns in the prompts were matched to the participants' preferred pronouns. At the start of the experiment, participants read an example and practiced one trial. In Experiment 1, participants were asked to verbalize their thoughts during

this task. Responses were recorded and transcribed for later scoring. An example prompt and response are provided in the appendix.

Episodic Simulation Task

For each prompt in the episodic simulation task, participants were asked to imagine a personal future experience in the next few years related to the cue. Cues were adapted from prompts in the mentalizing task (e.g., "Your friend is preparing for his wedding tomorrow. What could he be thinking?" was adapted to "Preparing for a wedding"). Participants were instructed that the imagined event should not have happened yet and should be specific in time and place. Participants were instructed to report everything they imagined. These instructions were adapted from Madore et al. (2014). In each session, participants responded to a series of five episodic simulation cues. Each trial lasted three minutes. Throughout the experiment, pronouns in the prompts were matched to the participants' preferred pronouns. In Experiment 1, participants were asked to verbalize their responses. Responses were recorded and transcribed for later scoring. One practice trial was included at the start of the experiment so that participants would be able to ask the researchers questions about the task. An example prompt and response are provided in the appendix.

Scoring

Scoring: Episodic Simulation Task

For imagined future events, detail scoring followed prior scoring procedures (i.e. Madore et al., 2014, as adapted from Levine et al., 2002). Details that contain information about the central constructed scenario were labeled as internal details. Internal details concerned, for example, actions, people, thoughts, feelings, objects, and related information. External details, on the other hand, included semantic details, repeated details, details unrelated to the imagined

event, or details from any non-central events. Two raters were trained to score these responses and achieved excellent reliability for internal details (Cronbach's alpha = .98, as assessed on 10 responses from participants who were excluded for failure to attend the second session of Experiment 1) and good reliability for external details (Cronbach's alpha = .78, as assessed on the same 10 items).

Scoring: Mentalizing Task

To capture the amount of detail in the mentalizing responses, we adapted the scoring procedure used for episodic simulation responses. We defined internal details as the details related to the mental state of the person in the imagined situation. Text was divided into individual details according to segmentation rules similar to those used in Madore et al. (2014). External details were defined as details that are not specifically related to mental states. These details involved off-task commentary, repetitions, event details that are separate from attributed thoughts, and semantic information (often in the form of character backstory). We achieved high reliability between three scorers (Cronbach's alpha = .97 for internal details, Cronbach's alpha = .79 for external details), as assessed on a series of ten narratives from participants that were excluded from Experiment 1 for failing to attend the second session in the experiment.

Statistical Analyses

Because we hypothesized that episodic retrieval contributes to detail generation when mentalizing, we expected that the ESI (when compared to control) would increase the number of internal details on the mentalizing task but not external details. This hypothesis was evaluated by testing whether an interaction between induction type and detail type predicted detail count. We conducted this analysis with a linear mixed effects model to account for the nested structure of our data and control for variation that introduced by different prompts. Our model included detail

count as the outcome variable, participant ID and prompt ID as random effects and detail type (External or Internal), induction type (ESI or Control), and induction type x detail type as fixed effects. In this analysis, two trials were excluded because the participant did not understand the scenario presented. In addition to directly testing our hypothesis with the interaction of induction type and detail type, we conducted a post-hoc test to determine if the number of internal details was dependent on induction type. We conducted a second post-hoc test to determine if the number of external details was also dependent on induction type. To conduct these post-hoc tests, we created two models. The first model predicted the number of internal details based on the fixed effect of induction type, with prompt ID and participant ID as random effects. The second model predicted the number of external details based on the same independent variables. We expected that induction type would be significant in the first but not the second post-hoc model.

We analyzed data from our episodic simulation task using mixed effects models as well. We predicted that the ESI (relative to the control induction) would increase the number of internal details that participants used when imagining events. We also predicted that the ESI would have no impact on external details. To test these hypotheses, we first used a mixed effects model to assess whether detail counts were predicted by an interaction between detail type (Internal vs External) and induction type (ESI vs Control). We included random effects of prompt ID and participant ID. We then conducted two post-hoc analyses. First, we used a mixed effects model to test for the fixed effect of induction type on the number of internal details. This model included random effects of participant ID and prompt ID. Second, we tested for the fixed effect of induction type on number of external details with a mixed effects model. This model also included random effects of participant ID and prompt ID.

Results

We found a significant interaction effect of detail type and induction type ($B = -7.61\ p = 0.00372$) as hypothesized for the mentalizing task. We found that the ESI elicited a greater number of internal details (M = 72.26, SD = 20.60) related to mental state inferences than the control induction (M = 65.72, SD = 22.70) ($b = 6.21, p = 0.00134$) . Importantly, the ESI did not elicit a significantly different number of external details (M= 9.71, SD = 9.0) when compared to the control induction (M= 8.49, SD = 8.14) ($b = -1.20, p = 0.213$) . Results are displayed in Figure 5.

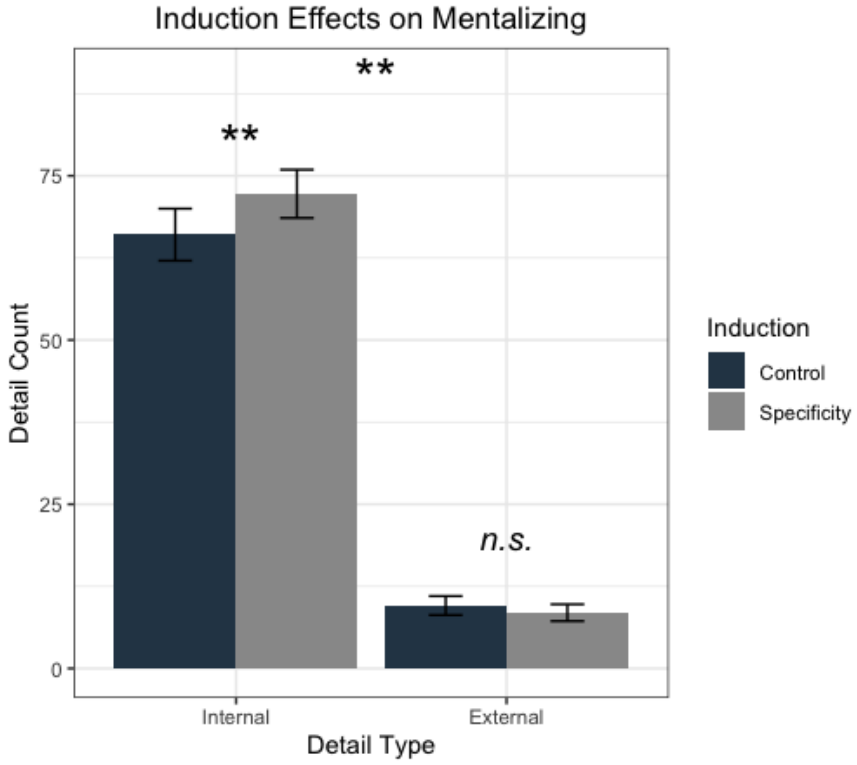

Figure 5. Experiment 1: Effects of episodic specificity and control inductions on the number of details that participants included when reporting mental state inferences. The detail counts included in this plot are averages from across cues and participants. Error bars represent 1 SE. * p < .05, ** p < .01, *** p <.001

On the episodic simulation task, we predicted a significant interaction between detail type and induction type. As predicted, the interaction of detail type (Internal vs External) and Induction Type (ESI vs Control) was significant ($b = -9.35, p = 0.00157$). Post-hoc tests revealed that participants provided more internal details (M = 79.36, SD = 22.39) after the ESI than after the control induction (M = 70.15, SD = 23.84) ($b = 8.60, p < .001$). Importantly, participants did not report a significantly different number of external details after the ESI (M = 11.12, SD = 11.88) when compared to the control induction (M = 10.56, SD = 9.61), $b = -0.49, p = 0.688$. Together, these episodic simulation results indicate that our manipulation worked. Results are displayed in Figure 6.

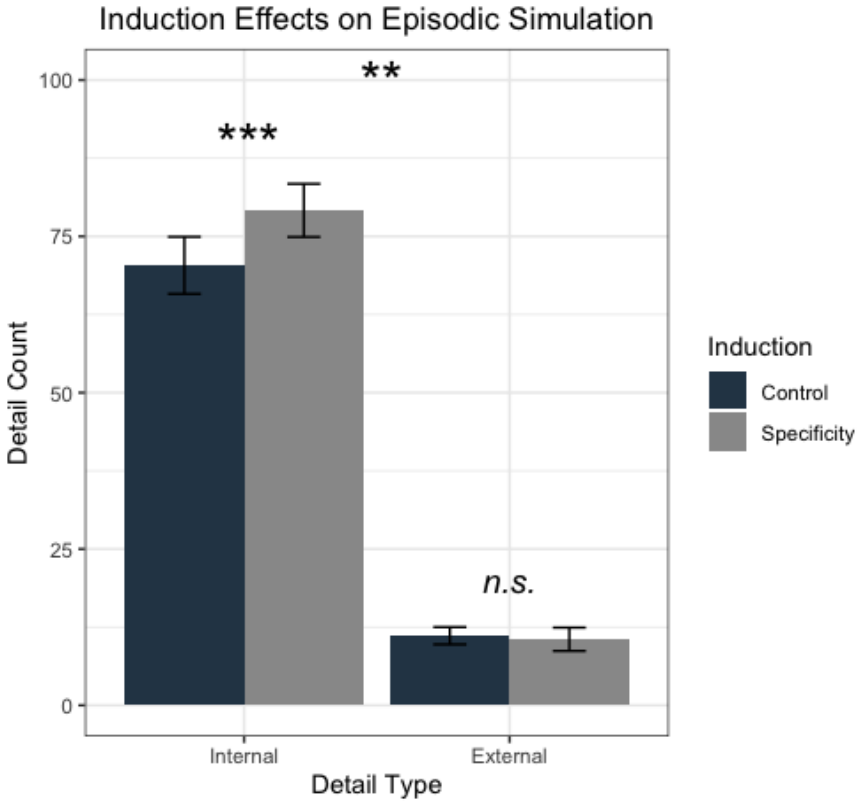

Figure 6. Experiment 1: Effects of episodic specificity and control inductions on the number of details that participants included when imagining future events. The detail counts included in this

(Continued) plot are averages from across cues and participants. Error bars represent 1 SE. * p < .05, ** p < .01, *** p <.001

To summarize, Experiment 1 provides preliminary evidence that episodic retrieval contributes to the number of details that participants provide when asked to imagine the thoughts, feelings, and intentions of others.

Overview: Experiment 2

We aimed to further explore the contributions of episodic retrieval to mentalizing in Experiment 2. After all trials were completed in the experiment, participants were asked to label each thought as 'old' (they explicitly remember having drawn on a memory during the experiment to provide the thought), or 'new' (they did not explicitly remember having drawn on a memory to provide the thought). The primary purpose of this addition was to test whether a meaningful number of mental state inferences were informed by specific memories. These labeled thoughts also allowed us to better understand how ESI affects mental state inferences. It allowed us to test whether ESI effects are specific to thoughts where participants are consciously drawing on specific memories, or whether it affects other thoughts as well. Further, it allowed us to test different explanations of the significant ESI effect in Experiment 1. Since we measured number of details in Experiment 1, the ESI effect could reflect participants providing more thoughts, providing more detailed thoughts (without changing the number of thoughts), or a combination of both. Separating text into individual thoughts allowed us to test which of these explanations best describes our data.

To allow us to isolate individual thoughts, we modified instructions for Experiment 2. In Experiment 1, participants were asked to verbally list all possible thoughts, feelings, and intentions of the person in the prompt. Identifying the boundary between distinct thoughts was

difficult for scorers, so we only obtained detail counts for the narratives. To separate thoughts in Experiment 2, we instead asked participants to type out their thoughts and to press the enter key twice after every time that they completed a thought.

In Experiment 2, we also used a different control induction to address an alternative interpretation of our results from Experiment 1. We interpreted results from Experiment 1 as providing evidence that ESI increased the number of details in mental state judgments, relative to the control induction. This interpretation suggests that episodic retrieval can contribute to mentalizing. However, an alternative explanation of these results is possible. A difference in the number of details following the ESI and control inductions could be driven by a decrease in the number of details following the control induction (rather than an increase because of ESI). Previous literature suggests that this second interpretation is unlikely: the ESI effect does not depend on using the impressions control as a contrast in other studies, since the effect exists when a math control task is used instead (e.g., Madore et al., 2014; Jing et al., 2016; for review, see Schacter & Madore, 2016). However, to address the possibility that our results from Experiment 1 were dependent on the specific control induction used, we replaced the impressions control induction with a math control induction for Experiment 2. If the ESI were to elicit more detailed thoughts in Experiment 2 as well, we could conclude that the effect is not driven by the impression control induction.

Last, a practical modification made data collection easier. In Experiment 1, participants attended two sessions. In Experiment 2, participants attended a single session, following protocols in some prior work (Madore et al., 2016; Madore, Jing, et al., 2019; Madore, Thakral, et al., 2019). A filler task was used to separate the ESI and control induction portions of the

experiment. We used a one-day protocol to avoid the participant dropout observed in two-sessions studies.

Materials and Methods: Experiment 2

Procedure

After the consent procedure, each participant first received one induction (ESI or control). Then, each participant completed five trials of the mentalizing task followed by five trials of the episodic simulation tasks. After these tasks, each participant completed a filler task (solving addition and subtraction problems), and then participated in the induction they had not yet engaged in (ESI or control). Afterwards, each participant completed five trials of the mentalizing task and five trials of the episodic simulation tasks. Presentation order of stimuli lists, induction type, and induction video type were counterbalanced across participants. Unlike in Experiment 1, the researchers did not remain in the experiment room while participants completed the future imagination and mentalizing tasks. This was done to reduce potential COVID-19 exposure.

Participants

To determine an adequate sample size, we performed a power analysis based on resampling mentalizing data from Experiment 1. This analysis indicated that we would need 22 participants to have an 80% chance of detecting an existing interaction. We would also need 22 participants to have an 80% chance of detecting an effect of the ESI on internal details. For counterbalancing purposes, we decided to recruit 32 participants.

32 participants were included in the analysis for Experiment 2. One additional participant was excluded for not completing the experiment because they felt sick, another was excluded for quitting after realizing that they were participating in the wrong study, and another was excluded

after a problem with Qualtrics. All participants were recruited from the Harvard University study pool and were granted course credit or paid for their participation. All participants were between the ages of 18 and 30 (M = 19.90 years old, SD = 1.77), with no history of neurological or psychiatric impairment. Five participants were male, and 27 participants were female. As in Experiment 1, all participants provided written consent in accordance with ethics protocols approved by Harvard University's Institutional Review Board.

Scoring

Scoring of mentalizing trials followed the scoring guidelines used in Experiment 1 and was conducted by one of the same raters. We modified scoring of episodic simulation trials to reflect recent research on autobiographical interview scoring. In Experiment 1, we scored the episodic simulation trials according to an adapted Autobiographical Interview scoring manual (i.e., Madore et al., 2014 as adapted from Levine et al., 2002). In this procedure, researchers identify segments of text as internal or external, then separate those segments into individual details. In a recent paper (van Genugten & Schacter, 2022), we found that separating these segments into individual details was unnecessary, since word count within internal segments correlated almost perfectly with internal detail counts across datasets (r = .86 to .92, mean r = .92). Likewise, the number of word count in external segments correlated almost perfectly with external detail count (r = .87 to .98, mean r = .94). As a result, we scored episodic simulation trials from Experiment 2 by annotating segments as internal and external, then extracting word counts from those segments. Using this new method, we found high interrater reliability between two raters for internal content (Cronbach's alpha = .94, as assessed on ten imagined events from participants excluded in Experiment 1) and external content (Cronbach's alpha = .84, as assessed on the same ten imagined events). Each of these two raters scored half of the episodic simulation

narratives. An example of scoring, including internal and external detail counts and internal and external word counts, is included in the appendix.

It is important to note that some experiments still require separating out content into individual details so that these details can be assigned to subcategories (e.g., event details, time details, perceptual details). However, since we were not scoring subcategories in our episodic simulation responses, we proceeded with internal and external word count.

Analysis

We predicted that ESI would increase the amount of internal content and have no significant effect on the amount of external content in episodic narratives. Likewise, we predicted that the ESI would impact the number of mental-state related details in responses, but not external details. Since these hypotheses are the same as in Experiment 1, data were analyzed with the same multilevel models used in Experiment 1. We used similar models to examine the effect of ESI on the number of ideas provided on the mentalizing task. That is, we used mixed effects models to predict the number of ideas provided from the fixed effect of induction type (ESI or Control) and random effects of participant ID and prompt ID.

Results

We found that responses on the mentalizing task after the ESI did not contain a greater number of internal details (M = 19.59, SD = 6.76) than responses after the control induction (M = 19.98, SD = 7.47) (b = -.1438, p = 0.765). We found that few external details were included in responses from Experiment 2, likely because these responses were written instead of spoken and transcribed. We found no significant difference in the number of external details that participants provided on the mentalizing task after ESI (M = .60, SD = 1.71), when compared to after the

control induction (M = .48, SD = 1.80) (b = .125, p = .395). These results are reported in Figure 7.

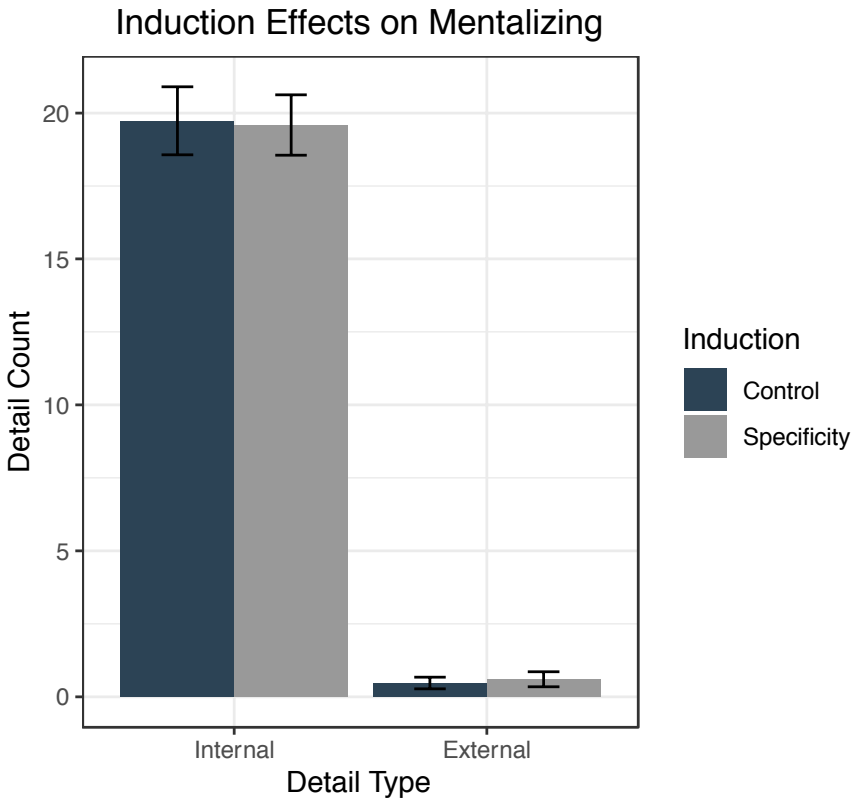

Figure 7. Experiment 2: Effects of episodic specificity and control inductions on the number of details that participants included when reporting their mental state inferences. ESI did not increase the number of internal or external details that participants provided. The detail counts included in this plot are averages from across cues and participants. Error bars represent 1 SE.

Next, we examined the number of thoughts that participants generated on the mentalizing task. We found that more than half of the thoughts were labeled by participants as having been informed by a specific memory. However, we also found that ESI had no significant effect on the number of new thoughts participant provided (b = -0.2337, p = 0.256), no significant effect on

the number of old thoughts participants provided ($b = 0.3112$, $p = 0.186$), and no significant effect on the number of total thoughts participants provided ($b = 0.07897$, $p = 0.646$). These results are reported in Figure 8.

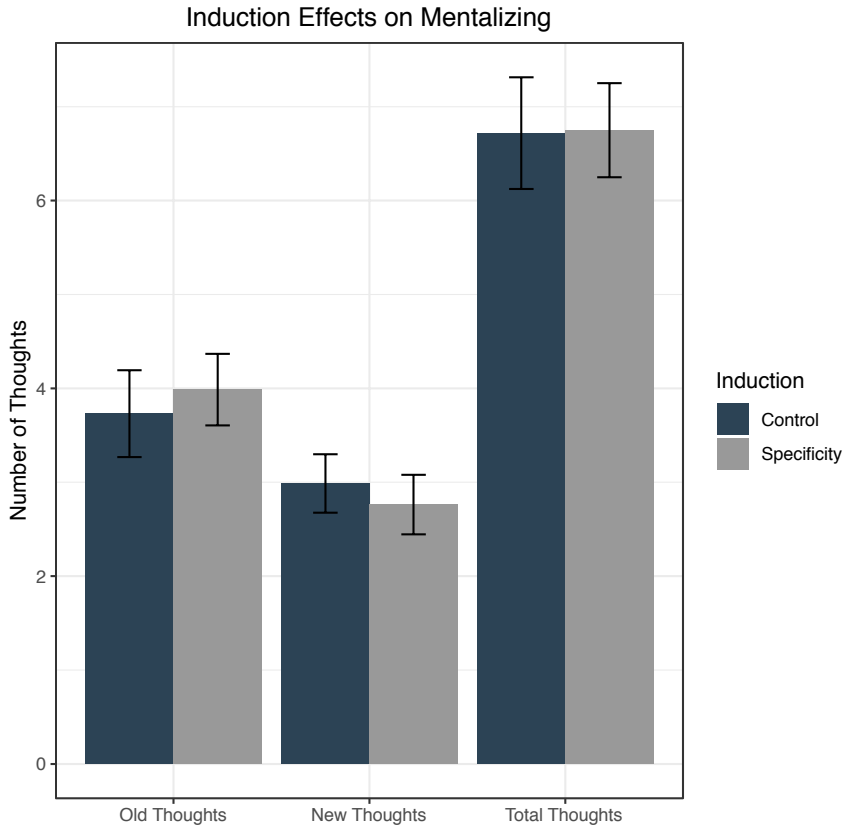

Figure 8. Experiment 2: Effects of episodic specificity and control inductions on the number of thoughts that participants attribute to others during mental state inference trials. ESI did not increase the number of old, new, or total thoughts that participants provide. The number of thoughts included in this plot are averages from across cues and participants. Error bars represent 1 SE.

Last, we examined whether ESI affected the amount of internal content, but not external content, that participants provided on an episodic simulation task. This task served as a manipulation check. We found that an interaction of detail type (Internal vs External) and

induction type (ESI vs Control) did not significantly predict the amount of content provided in the responses (b = -0.18, p = 0.972). In addition, when analyzing only internal content in stories, we found that induction type did not predict the amount of internal content (b = -1.31, p = 0.663). The amount of internal content after the ESI (M = 84.86 words, SD = 38.72) was approximately the same as after the control induction (M = 86.30 words, SD = 35.29). Likewise, induction type did not predict the number of external words, b = -1.39, p = .587. Participants provided approximately the same amount of external content after both inductions (M = 30.56, SD = 33.32 after ESI; M = 31.81, SD = 31.51 after control). Results are displayed in Figure 9.

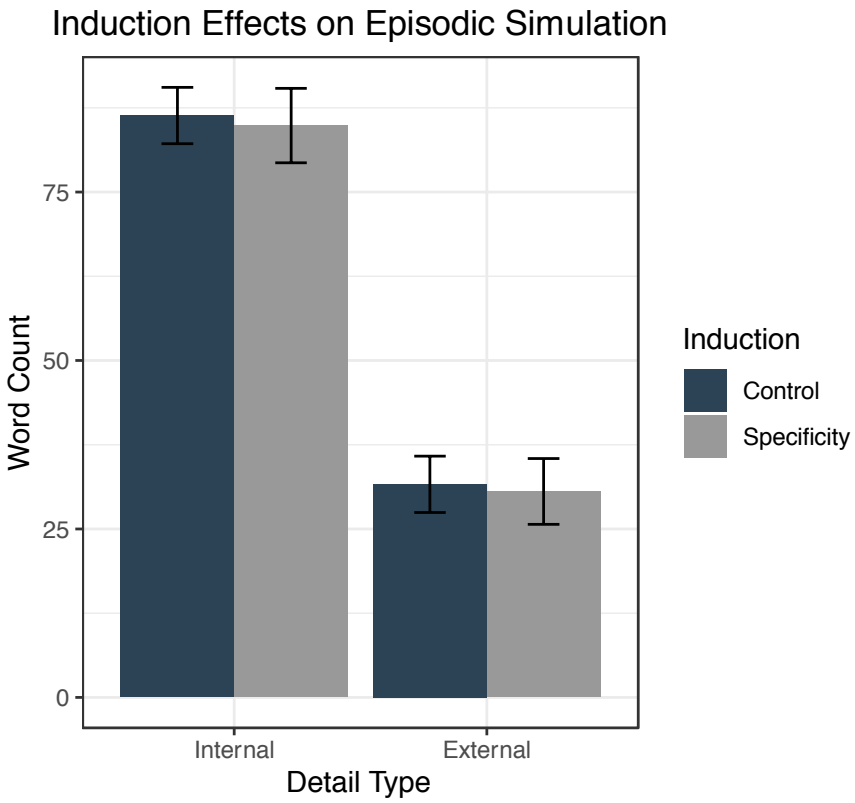

Figure 9. Experiment 2: Effects of episodic specificity and control inductions on the number of words that participants included when imagining future events. ESI did not increase the number

(Continued) of internal or external words that participants provided. The word counts included in this plot are averages from across cues and participants. Error bars represent 1 SE.

Discussion

Previous work has shown that episodic retrieval can influence empathy judgments (e.g. Vollberg et al., 2021). In this paper, we used two approaches to examine whether episodic retrieval also contributes to mentalizing. First, we manipulated episodic retrieval via ESI before participants engaged in a mentalizing task. In Experiment 1, we found that the ESI significantly increased the number of internal (but not external) details that participants included when describing the mental states of another person, which provided preliminary evidence that episodic retrieval can contribute to mentalizing. As expected, ESI also significantly increased the number of internal (but not external) details on the episodic simulation task, which was included as a manipulation check, in Experiment 1. In Experiment 2, we found that ESI had no effect on our manipulation check. Consistent with a failed manipulation, we further found that ESI had no effect on the number of details that participants provided when mentalizing. ESI also did not affect the number of old, new, and total thoughts that participants generated on this task. As a result of this failed manipulation check, and given repeated demonstrations that ESI affects this task (e.g. Madore et al., 2014; Madore & Schacter, 2016), we interpret these data as indicating procedural problems with the ESI in Experiment 2, rather than providing evidence against the hypothesis that episodic retrieval contributes to mentalizing. A second approach that examined this hypothesis in Experiment 2, however, did provide evidence that episodic retrieval contributes to mentalizing. We found that more than half of the thoughts that participants generated on the mentalizing tasks were later labeled by those participants as having been

informed by the retrieval of specific memories. Together, this evidence from Experiment 2 as well as ESI evidence from Experiment 1 suggests that episodic retrieval can contribute to mentalizing. Further work will be necessary to explore the specific conditions under which episodic retrieval plays a larger or smaller role in mentalizing.

To inform design of future research on this topic, we explore the possible causes of the failed ESI manipulation in Experiment 2 by discussing the various changes made between Experiment 1 and Experiment 2. First, we used a one-day experimental protocol in Experiment 2, rather than the two-session protocol used in Experiment 1. It is possible that (1) carryover effects or (2) participant exhaustion during this longer session introduced additional variability in our data, leading to less power to detect an ESI effect. However, we found no evidence of carryover effects: an interaction between ESI type and induction order did not significantly predict internal content for either task (mentalizing task: $b = 1.79$ details, $p = .063$; episodic simulation task: $b = 9.8$ words, $p = .103$, using mixed effects models that included participant ID and prompt as random effects). For trials in which there is no possibility of carryover effects of ESI (i.e. when the control is administered first), the number of internal details following ESI is numerically lower when compared to after control ($M_{ESI, future} = 80.53$, $M_{control, future} = 86.95$, $M_{ESI, mentalizing} = 18.72$, $M_{control, mentalizing} = 19.76$). Together, these analyses suggest that carryover effects are not responsible for the failed ESI effect observed in Experiment 2. In addition, we do not have strong evidence that participants were especially tired at the end of the experiment and responded differently as a result. Instead, it appears that they responded similarly to the first and second half of the experiment. For example, episodic simulation responses from the second half of the experiment were not shorter than those from the first half ($M_{first\ half} = 115.58$ words, $M_{second\ half} = 117.94$ words). Whether responses came from the first or second half of the experiment did

not significantly predict word count (b = 2.56, p = .266, when modeled with a mixed effects model that also contained random effects of prompt ID and subject ID). The responses on the mentalizing task were also not shorter in the second half of the experiment ($M_{\text{first half}}$ = 109.93 words, $M_{\text{second half}}$ = 114.15 words; b = 4.23, p = 0.0544). In addition, previous studies have used one-session protocols effectively (e.g. Madore et al., 2016; Madore, Jing, et al., 2019; Madore, Thakral, et al., 2019). As a result, it is unlikely that participant exhaustion or carryover effects can explain the difference in ESI results between Experiment 1 and Experiment 2.

Experiment 2 also differs from Experiment 1 because we changed our scoring approach for episodic simulation trials. For this task, we used internal word count rather than internal detail count as a measure of internal content. We also used external word count rather than external detail count as a measure of external content. However, since internal word counts and internal details have been shown to be nearly perfectly correlated (r = .92), and external details and external word counts have also been shown to be nearly perfectly correlated (r = .94) across datasets (van Genugten & Schacter, 2022), we do not think this change explains the lack of ESI effect. If the problem were only a minimal decrease in power because of scoring differences, we would expect to observe ESI related trends in our data, which we do not. Further, the change in scoring for episodic simulation trials is unable to explain the why ESI does not affect mentalizing in Experiment 2 as it does in Experiment 1.

Last, the changes we made to Experiment 2 protocols in response to COVID-19 may have affected how participants interacted with the experiment. Experiment 1 was conducted before the pandemic, so a researcher was present in the room for duration of each session. In Experiment 2, however, the researcher left as participants completed mentalizing and episodic simulation trials to decrease potential COVID-19 exposure. The participants may have interacted

with trials differently as a result. For example, participants could be less motivated to work on the task without a researcher in the room. Alternatively, participants could have become distracted during the experiment (e.g., by using their phones), which could have removed them from the retrieval mindset induced by the ESI. In addition, participants may not have adapted the retrieval strategy they engaged in during the ESI to the subsequent task without the presence of a researcher. While there may be additional explanations of differences between our experiments, we believe that changes due to COVID-19 protocols are the most likely. As a result, future studies that want to examine the effect of ESI on mentalizing should ensure that researchers stay in the room with participants as they complete the task.

To summarize, Experiment 1 showed evidence that ESI impacts mentalizing, while Experiment 2 failed to replicate this finding. The evidence from Experiment 1 is further supported by findings from Experiment 2 that participants report drawing on particular memories to infer the thoughts of other people for more than half of the thoughts they provide. Together with previous literature (e.g. Gaesser et al., 2018, Krienen et al., 2010), our results suggest that episodic retrieval contributes to mentalizing. To better understand this relationship, future work will be necessary to replicate these findings and investigate the different conditions under which episodic retrieval is (or is not) informative for mentalizing.

One particularly promising direction for future research comes from reflecting on the broader literature on mentalizing. This body of work points out that there are many ways to infer the thoughts of other individuals (discussed in e.g. Mitchell, 2006; Schaafsma et al., 2015; Waytz & Mitchell, 2011). For example, we can use our own preferences as estimates for others' preferences and subsequently adjust our predictions (e.g. Tamir & Mitchell, 2010, 2013). We can use causal logic to reason about mental states or use semantic knowledge about how people tend

to respond in common situations (reviewed in e.g. Epley & Waytz, 2010), or infer emotions from facial expressions (e.g. Ekman & Oster, 1979). As a result, when many sources of information are available to make mental state inferences (e.g., during a face-to-face interaction), it may not be necessary to draw on episodic retrieval. Instead, other information such as facial expressions and speech may better inform mental state inferences. However, when less information is available to serve as input for the mentalizing system (for example, when we cannot directly observe the situation we are mentalizing about, and when we do not have knowledge of how the person tends to respond to different situations), episodic simulation of the event may compensate and provide helpful context to the mentalizing system. As researchers test the conditions under which episodic retrieval contributes to mentalizing, then, the amount of information available to the mentalizing person may be an important variable to manipulate. To summarize, we provide initial evidence that episodic retrieval contributes to mentalizing, and much work remains to be done to determine in which situations these two cognitive processes interact.

Paper 2 Acknowledgements

We thank Ethan Harris, Jyotika Bindra, Emma Edenbaum, and Tawanda Mulalu, for their help in data collection, transcribing, and scoring. This research was funded by the National Institute on Aging Grant R01 AG008441 awarded to DLS.

Paper 3

van Genugten, R., & Schacter, D. L. (2022). Automated Scoring of the Autobiographical Interview with Natural Language Processing. PsyArXiv. https://doi.org/10.31234/osf.io/nyurm

Abstract

The Autobiographical Interview has been used in more than two hundred studies to assess the content of autobiographical memories. In a typical experiment, participants recall memories, which are then scored manually for internal details (episodic details from the central event) and external details (largely non-episodic details). Scoring these narratives requires a significant amount of time. As a result, large studies with this procedure are often impractical, and even conducting small studies is time-consuming. To reduce scoring burden and enable larger studies, we developed an approach to automatically score responses with natural language processing. We fine-tuned an existing language model (distilBERT) to identify the amount of internal and external content in each sentence. These predictions were aggregated to obtain internal and external content estimates for each narrative. We evaluated our model by comparing manual scores with automated scores in five datasets. We found that our model performed well across datasets. In four datasets, we found a strong correlation between internal detail counts and the amount of predicted internal content. In these datasets, manual and automated external scores were also strongly correlated, and we found minimal misclassification of content. In a fifth dataset, our model performed well after additional preprocessing. To make automated scoring available to other researchers, we provide a Colab notebook that is intended to be used without additional coding.

The Autobiographical Interview (AI) (Levine et al., 2002) is a widely used method to study the contents of participants' autobiographical memories. In a typical experiment using the AI, participants are asked to recall a specific event for each cue, making sure to report as much detail as they can. Several human raters then use a manual to identify and count internal details (central episodic details) and external details (mostly non-episodic details) in the narratives. Levine et al. (2002) first developed the AI and its scoring manual to study age-related differences in memory. With this procedure, Levine et al. (2002) showed that older adults provide fewer internal details than young adults when they were asked to retrieve autobiographical memories that are more than a year old, despite being able to provide the same number of external details. For memories that were less than a year old, older adult provided more external details than younger adults. These findings extended our understanding of the effect of aging on autobiographical memory and in so doing, also highlights that the AI can be used to study group differences in memory.

Studies conducted since Levine et al. (2002) have made it clear that the AI enables researchers to test specific theories about memory. For example, Addis et al. (2008) hypothesized that if episodic retrieval supports the construction of specific future events, age-related decreases in internal details during retrieval of past events should be accompanied by age-related decreases in internal details when imagining future events. Addis et al. (2008) adapted the AI to ask participants about imagined future events and remembered events. In addition to finding support for their hypothesis, they found a positive correlation between the number of internal details participants provided when remembering the past and imagining the future. These results are consistent with the constructive episodic simulation hypothesis, which suggests that we imagine future experiences by recombining elements from episodic memories

(Schacter & Addis, 2007). Other researchers have similarly adapted the AI to study future thinking. For example, Race et al. (2011) studied amnesic individuals with the AI, and showed that damage to the medial temporal lobe led to reductions in central episodic details on both episodic memory and future thinking tasks, despite intact descriptive abilities (i.e., normal performance when describing pictures).

Because the AI can be used to study features of autobiographical memory and future thinking, the procedure has seen widespread use across domains of psychology. In addition to the aforementioned studies of aging and amnesia, it has been used to study Alzheimer's disease (e.g., Irish et al., 2011), depressive disorders (e.g., Söderlund et al., 2014), and the contributions of episodic retrieval to other domains of cognition, such as means-end problem solving (e.g., Madore & Schacter, 2014). As of November 2021, over two hundred studies have used this interview (Levine, 2021), and the paper that first described the interview has over 1400 citations listed on Google Scholar.

The AI is widely used in psychology research despite the fact that scoring the AI takes a lot of time and effort (typically 10 minutes per memory, and each participant may provide ten or twenty memories). Having to manually annotate hundreds of pages of memories potentially limits the AI's usefulness and breadth of applications. Because large studies using the AI are impractical, past research typically studied only effects that could be detected with small samples (e.g., approximately thirty participants).

Here we introduce an automated scoring procedure for the AI that can reduce experimenter burden and help researchers to conduct larger experiments and study smaller effects. We believe that this new procedure will broaden the scope of research questions the field can address. Our automated scoring approach could also make online data collection more

practical. Online data collection is rarely used with the AI, likely because large numbers of participants are needed to compensate for noisy data. Online studies would allow researchers to gain access to a larger and more diverse population than the typical samples that have been used.

In the remainder of this paper, we will describe how narratives are typically scored, how researchers have attempted to streamline scoring, and our new approach for automating AI scoring.

Current approach for manually scoring memory details

Researchers follow a set of rules from the AI to score narratives. These rules explain how to classify pieces of text as internal or external, and how to identify bits of information within these segments that count as details.

Internal details refer to episodic details, and external details refer to non-episodic details (or episodic details that do not correspond to the central event being remembered or imagined). Internal details describe components of an event that are specific to time and place. The event's location and time, the people, objects, actions, thoughts, and perceptual details involved are all internal details. External details, on the other hand, are largely non-episodic details. These are any details that do not belong in the internal details category, and largely consist of factual information that does not require the participant to remember or imagine a specific event (e.g., "I've always enjoyed going to the beach for my birthday"). Participants sometimes provide information about events other than the central event they are being asked to describe. These details, while episodic in nature, are considered external details as well. Lastly, repetitive information (e.g., someone describes the same thing twice), and information unrelated to the event the participant is trying to describe (e.g., 'sorry for that cough!') are also considered external.

The manual for the AI provides clear rules on how to divide up segments of text into individual details. For internal segments, each piece of information that tells us something more about the event is generally counted as a detail. For example, “he had a hat” is considered one detail. Any additional descriptors count as additional details, e.g., “he had a brown hat” is considered two details. This description is illustrative of the general approach; for an exhaustive list of rules and exceptions, see the manual available upon request from Dr. Brian Levine (blevine@research.baycrest.org). A brief scoring example is provided in the appendix.

Existing automation approaches for memories

Several researchers have streamlined scoring of the AI, yet no group has fully automated the scoring of details in narratives. Previous work consists of two approaches: speeding up the processes involved in scoring, and predicting the number of internal and external details. For example, Wardell, Esposito, et al. (2021) automated the process of transcribing spoken narratives to text with Dragon NaturallySpeaking software. The researchers also reduced the time necessary for scoring by setting up keyboard shortcuts in Microsoft Word. Once details were manually scored, their software automatically counted the scored details in each memory. After implementing a protocol of this kind, a research group would be able to score more rapidly (see, e.g., Wardell, Madan, et al., 2021). However, much of the work remains to be done by hand: identifying internal and external content and separating the narratives into details are both still done manually.

To the best of our knowledge, only one paper reports an attempt to automatically generate AI detail scores for each narrative. Peters et al. (2017) first extracted eighty-three features from each narrative, such as the number of emotion words, the valence of these emotion words, the number of words in the story, and the number of nouns in the story. Peters et al. then

used these features in several regression models (e.g., principal component regression) to predict the number of internal and semantic details each participant provided (summed across five or twelve narratives per participant). When Peters et al. predicted the number of internal details provided by each participant, Root Mean Square Error (RMSE) was approximately .5 for internal details and .65 for semantic details. Peters et al. report one model built to predict the number of internal details in individual narratives. RMSE for this model was approximately .75 for episodic future thinking narratives and .85 for autobiographical memory narratives. No model was reported that predicted semantic details in individual narratives. To contextualize these results, a simple model that predicts the mean number of internal or semantic details for every narrative in this dataset would result in RMSE = 1, while a model with perfect predictions would result in RMSE = 0. Importantly, word count was a significant predictor for models predicting internal and semantic details counts. Since predictions were driven in part by the total amount of content, these regression models are presumably misclassifying internal content as semantic, and vice versa. So, while these researchers took an important first step by attempting to automatically score the AI, their predicted memory scores differed significantly from the actual memory scores, and additional work is needed to automate AI scoring.

Related work (Takano et al., 2017, 2018, 2019) has automated scoring of the Autobiographical Memory Test (Williams & Broadbent, 1986) with more success. When using the Autobiographical Memory Test, human raters classify memories as specific or general. Takano et al. used word frequencies and parts of speech frequencies to train a classifier to determine which memories were specific and which memories were general. Across studies, Takano et al. report good classification results, with high accuracy, frequent correct identification of specific memories, and frequent correct identification of general memories. For

example, for narratives from English-speaking adults reported in Takano et al. (2019), classification was 81.1% accurate. Specific memories were correctly identified in 81.8 % of cases, and general memories were correctly identified in 80.3% of cases. These results suggest that natural language processing provides a promising path for automated memory interview scoring.

Our automated scoring approaches

To improve automated scoring accuracy of the AI, we relied on advances in natural language processing to identify the amount of internal and external content in each sentence of an AI narrative. After classifying each sentence, we counted the amount of internal and external content in each narrative and validated these counts against detail counts obtained through manual scoring.

We trained our classifier by using data scored according to the AI. Specifically, we used these data to fine-tune weights at the end of an existing neural network, which had previously been trained on different natural language tasks. This procedure allowed us to take advantage of the language representations that the neural network had previously learned. This process, known as *transfer learning*, is a standard approach for classifying language content according to new labels, especially when few training examples are available (for introduction, see e.g. Azunre, 2021). Specifically, we fine-tuned distilBERT (Sanh et al., 2019) with the 'huggingface' library (Wolf et al., 2020).

We trained and evaluated our model with five datasets, which involved data scored according to the standard or adapted AI. We found that our code accurately identified internal and external content, with minimal misclassification of internal content as external, and minimal misclassification of external content as internal.

Methods

Model Training and Evaluation Data

To train our model to classify the amount of internal and external content in sentences, we requested data from several different researchers. All data we used were previously scored on a computer using standard or adapted AI scoring manuals. These data spanned several different tasks. Three datasets contained autobiographical memories (King et al., 2021; Sheldon et al., 2020; Strikwerda-Brown et al., 2021), and one of these contained data from both younger and older adults (Sheldon et al., 2020). Another dataset contained future simulation data from younger and older adults (Devitt & Schacter, 2018, 2019). We also included data from a study on creative writing (van Genugten et al., 2021) that was scored using an adapted AI scoring manual. These data were included to test whether the model would generalize to non-memory or future simulation paradigm that used adapted AI scoring. Last, one dataset included a picture description task and an open-ended thoughts description task (Strikwerda-Brown et al., 2021). These data were scored with guidelines that were different from the adapted or standard AI manuals. So, these data were included for exploratory analyses, without the expectation that our model would perform well on them. Because they were scored differently from all the other datasets, they were never included in the training sets. Each of these datasets is described in more details below.

Dataset 1: Autobiographical memories (King, Romero, Schacter, & St. Jacques, 2021)

King et al. (2021) examined how retrieving memories from an observer perspective (as opposed to a first-person perspective) changed the narratives. In the first session of this study, participants were asked to elaborate on a subset of memories in which they rated the event as occurring through their own eyes (at least a 5 on a 7-point scale measuring self-perspective). We

used these memories for our analyses. These data were in written form and were scored according to the AI (Levine et al., 2002).

This study generated a dataset of 40 individuals (25 female). Participants were, on average, 23.33 years old (SD = 3.17). All participants indicated that they were not previously diagnosed with a mood or cognitive disorder, nor taking any medication that could affect performance on the study. All participants were recruited from the Harvard study pool and the community.

Data Set 2: Autobiographical Memories (Sheldon et al., 2020)

Sheldon et al. (2020) collected autobiographical memories to test whether cue valence and arousal affected subsequent retrieval and elaboration of memories. In this experiment, participants listened to a series of 24 musical excerpts, which served as retrieval cues. After each retrieval cue, participants wrote down a caption to describe the memory they had retrieved. In a second session, participants were presented with the captions they had previously written down, given 30 seconds to remember the memory, and then used two minutes to describe what they remembered. Responses were audio-recorded and transcribed. Responses were then scored using the standard scoring guidelines from the AI (Levine et al., 2002).

Participants were recruited from McGill University's study pool. Each of the 42 participant was fluent in English and free of major neurological or psychiatric disorders. Participants were on average 20 years old (SD = 1.4) and had 14.6 years of education (SD = 1.1). 37 of the participants were female.

Data Set 3: Autobiographical Memories, Thoughts, and Picture Descriptions. (Strikwerda-Brown et al., 2021)

Strikwerda-Brown et al. (2021) investigated age-related changes in memory on a cued retrieval task and an open-ended task. Participants also completed a picture description task (cf., Gaesser et al., 2011). On each trial of the experiment, the participants saw a picture, were asked to retrieve a memory related to the image (memory task), to describe what was present in the image as if to someone who could not see the image (description task), or to describe the thoughts that arise when viewing the picture (thoughts task). All narratives were transcribed after being verbally reported by the participants.

Scoring of the memory task followed guidelines developed in Levine et al. (2002), with modified scoring guidelines for external details as described in Strikwerda-Brown et al. (2019). The picture description task was scored by following guidelines developed by Gaesser et al. (2011). Perceptual details in the picture were scored as internal details, and all other details (e.g., inferences about the picture, general comments about the picture) were scored as external. Details in the thoughts task were considered internal if they described any past event; all other details were considered external.

24 older adults and 25 younger adults were included in the analysis of this study. Younger adults were, on average, 21.7 years old (SD = 2.4). Participants reported no neurological or psychiatric impairments that would affect the study. Older adults were recruited from an existing database of older adults in the Montreal area. Younger adults were recruited from the McGill University study pool and surrounding areas.

Data Set 4: Future Simulation: Young Adult and Older Adult Data (Devitt & Schacter, 2018, 2019)

Devitt & Schacter (2018, 2019) examined how episodic simulation of an event before learning of its outcome affected the subsequent memory of that outcome. In their studies,

participants were presented with a series of cues for future events and were instructed to imagine the events going well or poorly for 3 minutes. Participants were instructed that each imagined event should occur within the next year. Afterwards, participants were given descriptions of how the events happened. In a second session, participants were tested for their memory of how the event happened.

Across two studies, future simulations from older and younger adults were audio-recorded, transcribed, and scored according to the AI. Data from these experiments include 27 younger adults (mean age = 22.59 years, SD = 3.18, 12 male) and 25 older adults (mean age = 72.24, SD = 6.49; 7 male). Participants indicated no history of neurological or psychiatric impairment. These participants were recruited from Harvard University and the surrounding community, using the Harvard psychology study pool.

Data set 5: Creative Writing Narratives. (van Genugten et al., 2021)

van Genugten et al. tested whether episodic retrieval contributes to creative writing performance. Specifically, van Genugten et al. used the Episodic Specificity Induction (Madore et al., 2014; for review, see Schacter & Madore, 2016) to manipulate episodic retrieval prior to a creative writing task. Detail counts after the ESI were compared to detail counts after two control inductions.

In the creative writing task, participants read a series of excerpts from literature and were asked to continue writing each story in a style that felt natural to them. Each story was scored according to scoring guidelines from the ESI studies of Madore et al. (2014) and Jing et al. (2016), which were adapted from the standard AI scoring (Levine et al., 2002). In their scoring, Van Genugten et al. also considered all event details as internal details. This procedure differs from previous guidelines, which only considered details from the central events to be internal.

This change was made to ensure that no episodic details were marked as external. Data from the first experiment were scored by hand, and as such were not used in the training and evaluation of our model. Data from the second experiment of this study were used since scoring was done on the computer.

Data used in this paper come from 32 participants, who each wrote 10 stories. Participants were young (18-30 years old, M = 24.03 years, SD = 3.51; 21 female, 11 male) and recruited from the Harvard University study pool. No participant reported neurological or psychiatric impairment at the time of the study.

Data Preparation

Data were read in from various sources, including text files and SciTos (Wickner, Englert, & Addis, 2015) html exports. Any prompting by the researcher (e.g., 'tell me more about that') was removed. Data were manipulated so that formatting was identical across datasets. Because our approach classifies individual sentences, narratives were split into sentences using pySBD (Sadvilkar & Neumann, 2020). pySBD splits text into sentences based on 48 rules that rely in part on punctuation.

Additional preprocessing was necessary after we noticed that some exceptionally long sentences contained a large majority of the narrative they came from (or even the full narrative). These narratives were transcribed with little or no punctuation, leading to few sentence splits by pySBD. To mimic narratives transcribed with full punctuation, we removed sentences that contained more than 8 details, since these sentences are likely missing punctuation. Detail counts associated with the narratives, which we used for validation, were updated to reflect the removal of this content. This preprocessing step is not included in the code we make available that other researchers can use to automatically score their own narratives. Researchers who want to use our

model in their own research should add punctuation as they are transcribing, to accommodate sentence splitting by pySBD. Alternatively, participants can be asked to type narratives, so that researchers do not have to add punctuation as they transcribe.

Each sentence was classified as belonging to one of four categories: containing 0% internal content (i.e., 100 % external content), 50% internal content, 75% internal content, or 100% internal content. We modified training datasets such that there were an equal number of sentences in all four categories. We did this by identifying the category with the greatest number of sentences, and upsampling data from all other categories. So, for example, if a training dataset were to contain 10,000 fully internal sentences, and 8,000 sentences from each of the three remaining categories, we would sample 2,000 sentences with replacement from each of those three categories, then add those sentences to the dataset so that we have 10,000 training examples in each category. We used training data with an equal number of examples for each label because this procedure prevented the model from learning to use relative frequencies of internal and external details to improve prediction accuracy. This step is necessary because if we did not upsample our training sets, and our narratives contained many more internal details than external details, the model could obtain relatively high accuracy by classifying all details as internal.

Model Training and Evaluation

We trained and evaluated the performance of our classification model with five datasets that are described in more detail in *Model Training and Evaluation Data*. We iteratively left out one dataset for evaluation, using the other four for training. For some datasets, data from multiple tasks or experiments were available. When these datasets served as the testing set, performance on each task was separately evaluated. For example, one dataset involved future

simulation for older adults and younger adults. When this dataset was left out for evaluation, the model was trained on the other four datasets and was then separately evaluated on the older adult data and the younger adult data. We report performance of all evaluation sets separately. Picture description and thoughts tasks from Strikwerda-Brown et al. (2021) were never included in training data because they were not scored with the adapted or standard AI.

As we trained and evaluated our model, we aimed to mimic the AI scoring, which involves two primary steps: annotating information as internal or external, and separating text segments into individual details. Below, we outline our approach to both steps.

Separating text segments into details

The AI scoring manual provides guidelines on how to split internal and external text segments into individual details, which allows researchers to quantify how much content is present in these segments. We chose not to train a model to split segments into individual details, since we have near-perfect proxies for internal detail counts and external detail counts: the number of words in internal segments, and the number of words in external segments.

To determine if internal word count adequately captures the number of internal details in narratives, we examined the correlation between these two variables across our datasets. The correlation between the number of internal details and internal word count ranged from .86 to .98 (mean = .92). We repeated this process for external details. The correlation between the number of external details and external word count ranged from .87 to .98 (mean = .94). Together, these extremely high correlations suggest that we can use internal and external word counts to quantify the amount of internal and external information. That is, we do not need to split sentences into individual details for the purposes of this project. An example of scoring with internal and external details and internal and external word counts is provided in the appendix.

We should note that internal and external word counts would not be adequate for all circumstances. For example, splitting content into individual details may be helpful for researchers that want to assign each detail to a subcategory (e.g., place, time, perceptual etc.). However, because the purpose of this project is simply to quantify the amount of internal and external content in narratives, we used internal and external word count as excellent approximations of internal and external details.

Classifying Information as Internal or External: Overview

To identify internal and external information, we adapted a common approach for classifying text. We fine-tuned weights at the end of an existing neural network with new data. We trained our model to classify sentences as containing only external content, 50% internal content, 75% internal content, or 100% internal content. To select classification labels, we calculated the percent of internal content in each sentence in the first dataset we obtained (Devitt & Schacter, 2018, 2019). A histogram of these percentages showed clusters at approximately 0, 50, 75 and 100%; hence, our labels.

Classifying Information as Internal/External: Model Specifics

To identify internal and external information, we used a model designed to be fine-tuned on sentences for classification. Specifically, we used distilBERT (Sanh et al., 2019), which is a language representation model that can be fine-tuned to new tasks by adding a classification head. The classification head contains a single linear layer for classification at the end of the network's pooled output. Fine-tuning this model involves changing the weights of this last layer to improve predictions on the fine-tuning data. distilBERT provides state-of-the-art performance on a range of natural language processing benchmark tests, while using fewer parameters than its ancestor BERT (Devlin et al., 2019). distilBERT has been trained to mimic BERT's performance

on two tasks: masked word prediction and next sentence prediction. Training with next sentence prediction involves providing the model with pairs of sentences. For each pair of sentences, the model must determine whether the second sentence followed the first sentence in the source text, or whether that pair of sentences is randomly paired. Training with masked word prediction involves randomly masking a subset of words in each sentence (e.g., 'the [MASK] gave the soccer player a yellow card'), then training the model to predict what the masked words are ('referee' in this case). Both types of learning require no human annotation but allow the network to acquire language knowledge that can then be taken advantage of in subsequent fine-tuning. Training data for these prediction tasks come from English Wikipedia text and the BookCorpus (a dataset of 11,038 unpublished books). We chose to use distilBERT instead of BERT or RoBERTa (Liu et al., 2019) because of its rapid training, as our model had to be trained six times: five times for our leave-one-dataset-out cross validation, and once for training on all datasets together.

We used Huggingface Transformers (Wolf et al., 2020) to fine-tune distilBERT on our classification task. We used accuracy as our evaluation criterion when training. We used Huggingface's default training arguments for fine-tuning. We used three training epochs, a batch size of 16 per device during training, a batch size of 64 for evaluation, 500 warmup steps, and a .01 weight decay.

Evaluating Model Performance

After classifying each of our sentences in the evaluation sets, we aggregated all sentences for each narrative and obtained an estimate of the amount of internal and external content in each narrative. We correlated these estimates with internal and external detail counts for validation. If our model were successful, we would expect the predictions of internal and external content to

match actual internal and external detail counts. Our expectations, then, were that 1) the number of internal details in the narratives would correlate with the amount of predicted internal content; 2) the number of external details in the narratives would correlate with the amount of predicted external content. Because the purpose of the AI is to correctly label content as internal or external, we also expected the model to not misclassify internal content as external. In other words, we also expected that 3) the number of internal details would be unrelated to the amount of predicted external content. Likewise, we should not misclassify external content as internal, so we further expected that 4) the number of external details would be unrelated to the amount of predicted internal content.

These four predictions are displayed graphically below in Figure 11. For each evaluation dataset, we report results in a similar format.

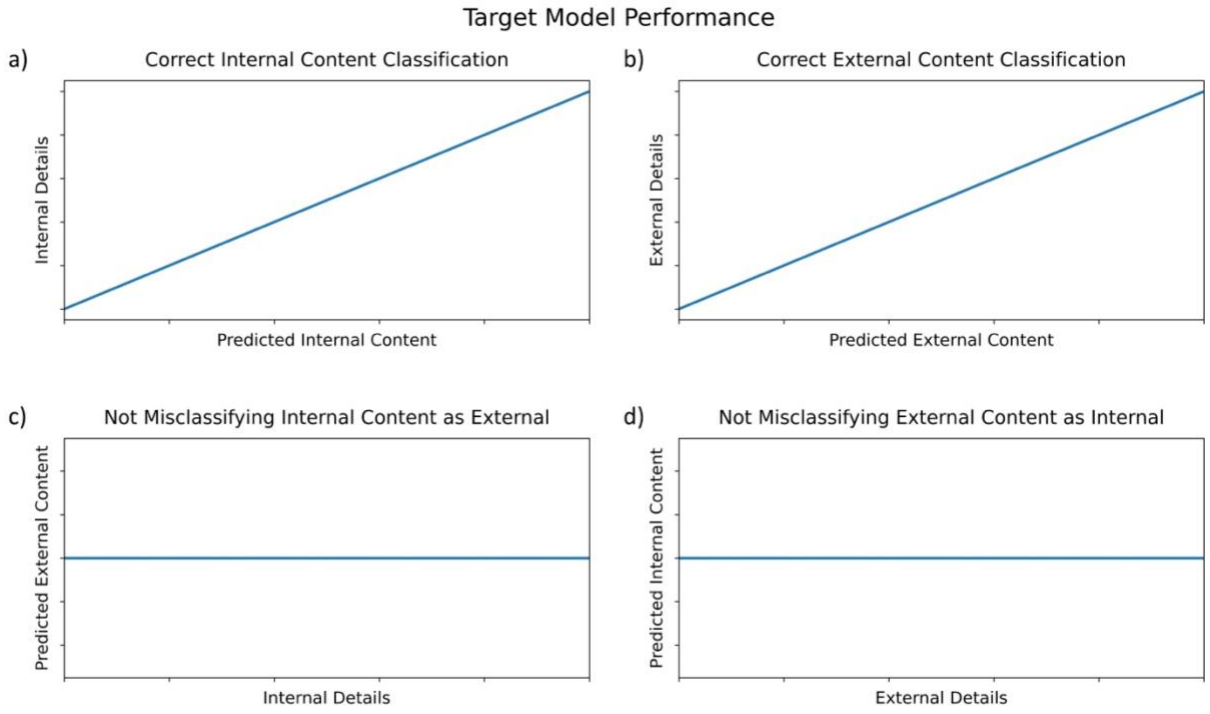

Figure 11. Target model performance. Internal (panel A) and external (panel B) content are accurately identified, with no misclassification of internal details as external (panel C) and external details as internal (panel D).

Results

To evaluate how well our model scored narratives, we examined internal and external detail scores as a function of predicted internal and external content. We evaluated performance on each left-out dataset separately.

Narratives Scored with the Standard or Adapted AI

Results: Future Simulation: Older Adult Data (Devitt & Schacter, 2018, 2019)

We examined whether our model correctly identified internal content. We found a strong relationship between predicted internal content and the number of internal details in future simulation narratives (fig 12a, $r = .81$, $p < .001$). We also examined the extent to which our model correctly identified external content. We found a strong relationship between predicted external content and the number of external details in future simulation narratives (fig 12b, $r = .65$, $p < .001$). We expected to find lower correlations when we examined the extent to which our model misclassified data. We examined how much internal content was misclassified as external content by our model. We did not find a significant relationship between internal details and predicted external content (fig 12c, $r = .12$, $p = .151$). We also examined how much external content was misclassified as internal content. We found a weak negative relationship between external details and predicted internal content (fig 12d, $r = -.22$, $p = .012$). To summarize, we found greater correct classification of internal content than misclassification ($R^2 = 0.65$ vs $R^2 = 0.02$). We also found greater correct classification of external content than misclassification ($R^2 = 0.42$ vs $R^2 = 0.05$). These results are summarized in figure 2 below.

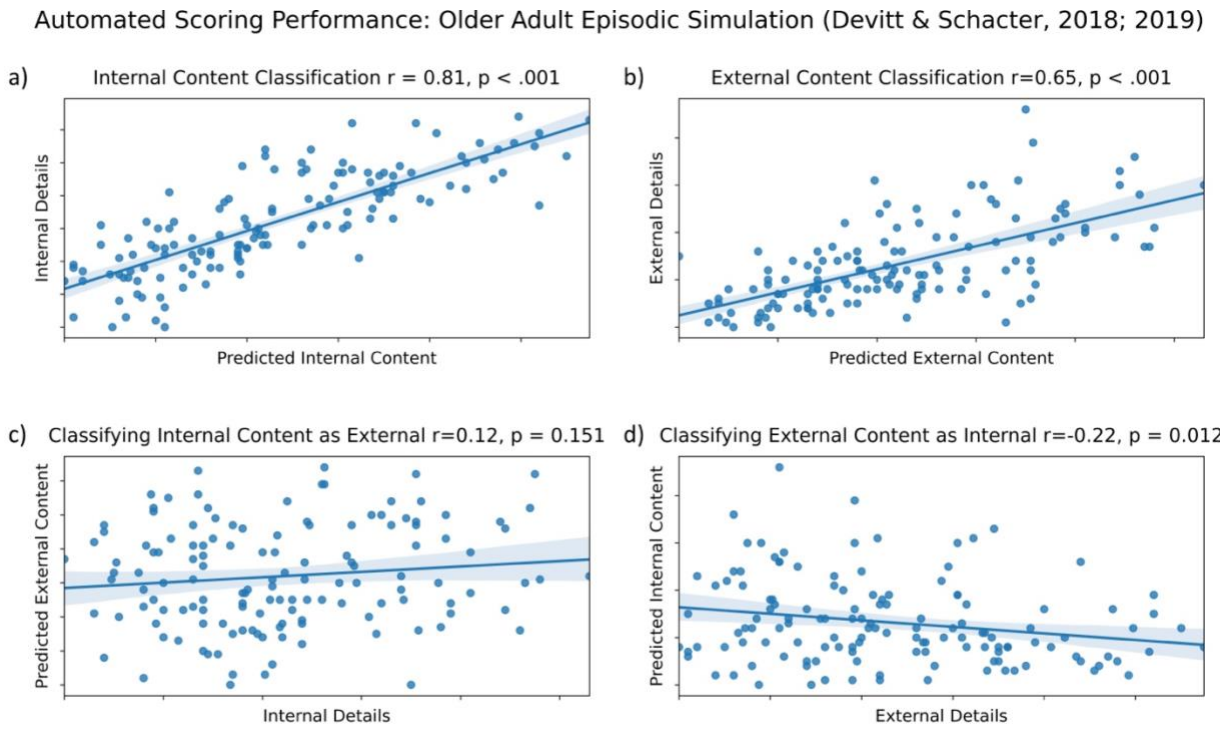

Figure 12. Model performance on older adult episodic simulation data from Devitt & Schacter (2018, 2019). Internal (panel A) and external (panel B) content are accurately identified, with minimal misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Future Simulation: Young Adult Data (Devitt & Schacter, 2018, 2019)

Our model correctly identified much of the internal content (fig 13a; r = .71, p <.001) and also correctly identified much of the external content (fig 13b, r = .40, p <.001). As expected, we observed less misclassification than correct classification. Internal content was not significantly misclassified as external content (fig 13c, r = .09, p = .27), and external content was not significantly misclassified as internal content (fig 13d, r = .08, p = .33). To summarize, we found greater correct classification of internal content than misclassification ($R^2 = 0.47$ vs $R^2 = 0.01$).

We also found greater correct classification of external content than misclassification ($R^2 = 0.16$ vs $R^2 = 0.01$). These results are summarized in the figure below.

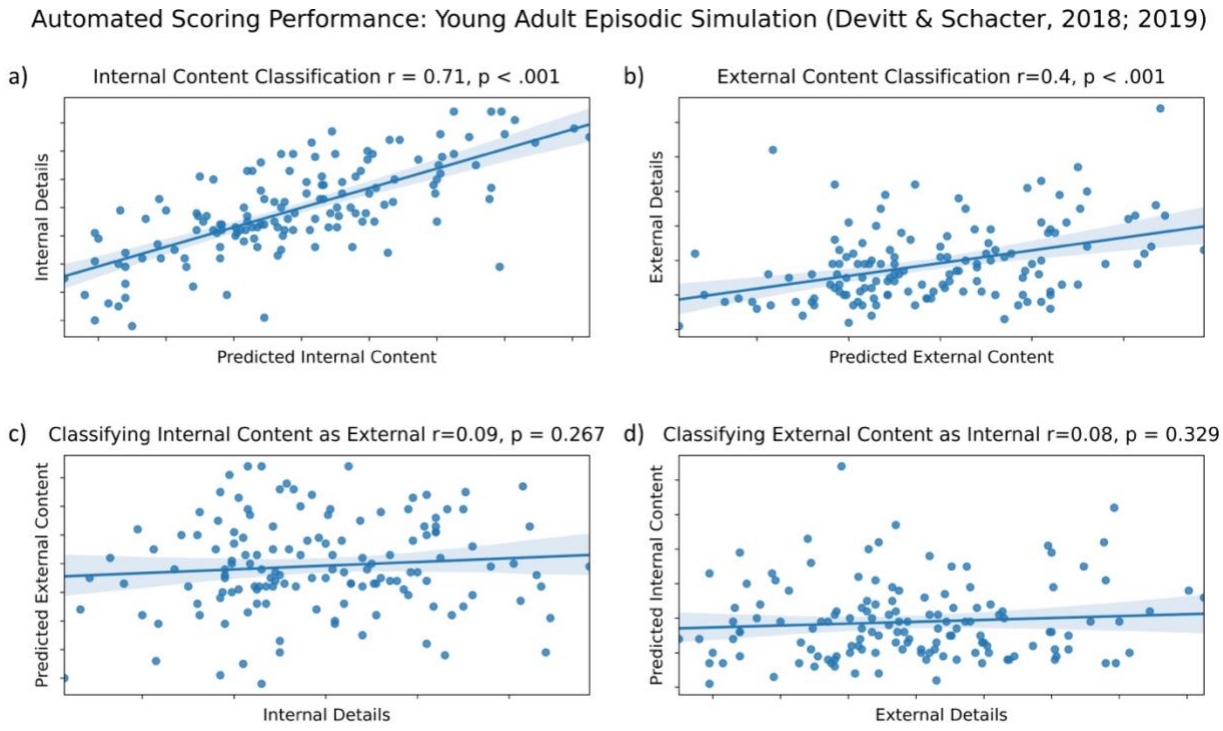

Figure 13. Model performance on young adult episodic simulation data from Devitt & Schacter (2018, 2019). Internal (panel A) and external (panel B) content are accurately identified, with minimal misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Autobiographical Memory (King et al., 2021)

Once again, our model correctly identified much of the internal content (fig 14a; r = .97, p <.001) and also correctly identified much of the external content (fig 14b, r = .54, p <.001). As expected, we observed less misclassification than correct classification. Internal content was not often misclassified as external content (fig 14c, r = .18, p < .001), and external content was not

significantly misclassified as internal content (fig 14d, r = -.01, p = .791). To summarize, we found greater correct classification of internal content than misclassification ($R^2 = 0.94$ vs $R^2 = 0.03$). We also found greater correct classification of external content than misclassification ($R^2 = 0.29$ vs $R^2 = 0.00$). These results are summarized in the figure below.

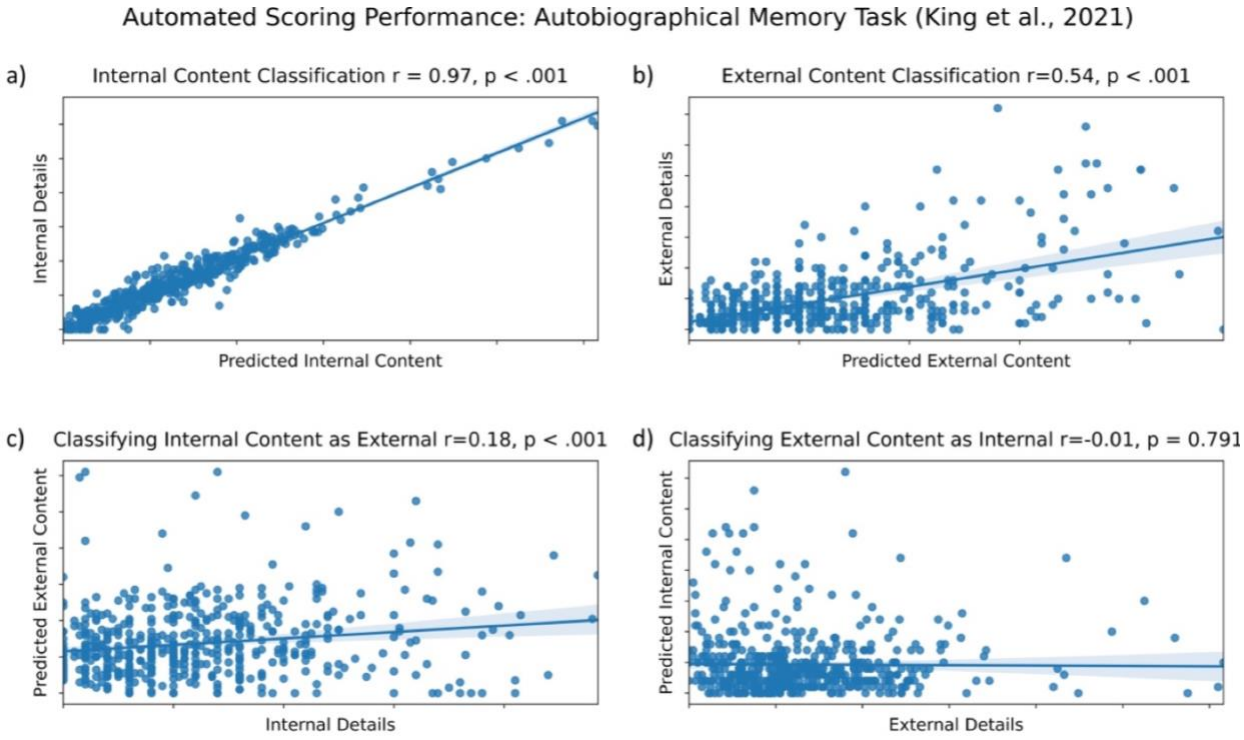

Figure 14. Model performance on autobiographical memory data from King et al. (2021). Internal (panel A) and external (panel B) content are accurately identified, with minimal misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Autobiographical Memories (Strikwerda-Brown et al., 2021)

Our model correctly identified much of the internal content (fig 15a; r = .90, p <.001) and external content in autobiographical memories (fig 15b, r = .79, p <.001). Internal content was not significantly misclassified as external content (fig 15c, r = .08, p = .366), and external

content was also not significantly misclassified as internal content (fig 15d, r = .08, p = .35). To summarize, we found greater correct classification of internal content than misclassification (R^2 = 0.80 vs R^2 = 0.01). We also found greater correct classification of external content than misclassification (R^2 = 0.68.vs R^2 = 0.01). These results are summarized in the figure below.

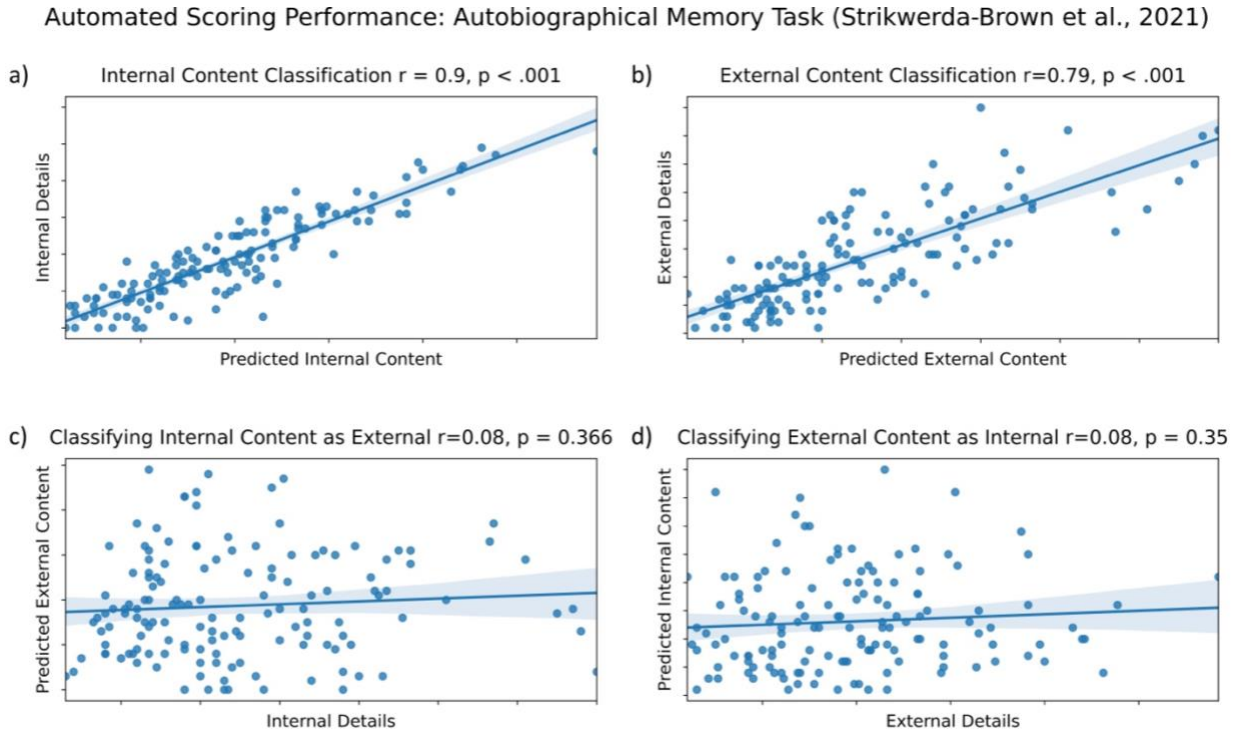

Figure 15. Model performance on autobiographical memory data from Strikwerda-Brown et al. (2021). Internal (panel A) and external (panel B) content are accurately identified, with minimal misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Creative Writing Narratives. (van Genugten et al., 2021)

Our model correctly identified much of the internal content (fig 16a; r = .80, p <.001) and external content (fig 16b, r = .77, p <.001) in creative writing narratives. As expected, we observed less misclassification than correct classification. Internal content was rarely

misclassified as external content (fig 16c, r = -.17, p = .003), and external content was not significantly misclassified as internal content (fig 16d, r = .01, p = .848). To summarize, we found greater correct classification of internal content than misclassification (R^2 = 0.64 vs R^2 = 0.03). We also found greater correct classification of external content than misclassification (R^2 = 0.59.vs R^2 = 0.00). These results are summarized in the figure below.

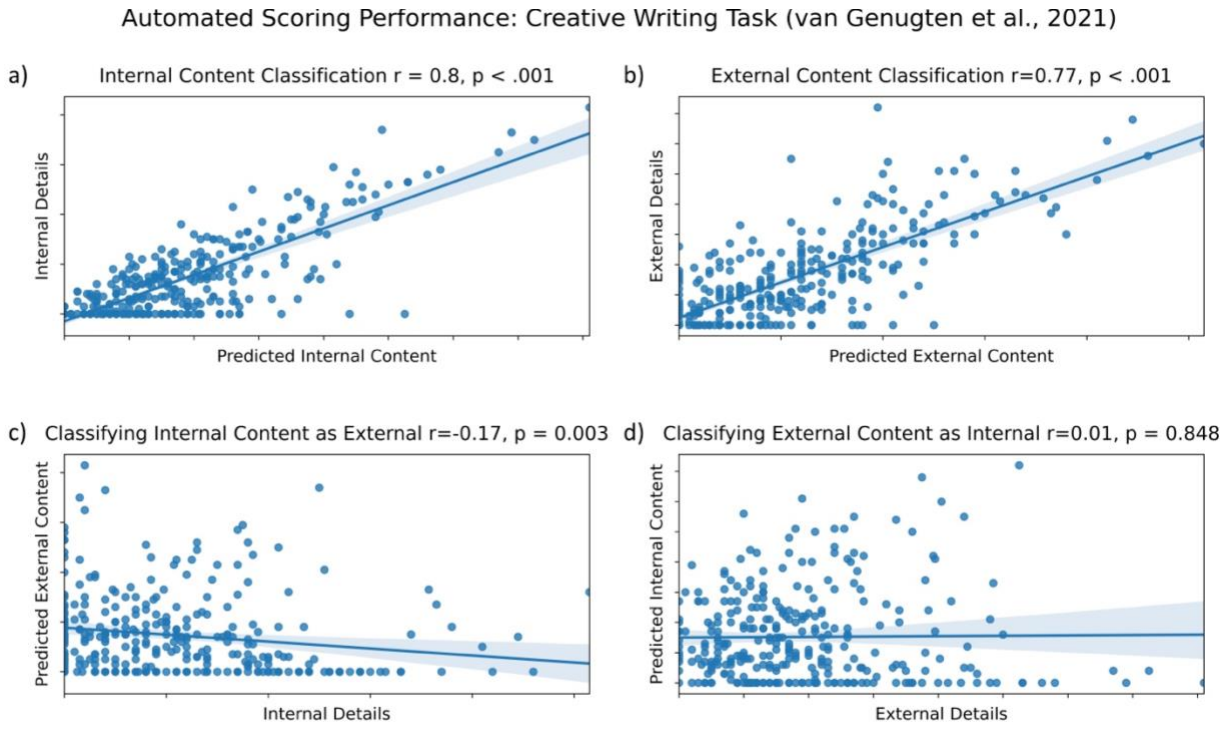

Figure 16. Model performance on autobiographical memory data from van Genugten et al. (2021). Internal (panel A) and external (panel B) content are accurately identified, with minimal misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Autobiographical Memories (Sheldon et al., 2020)

Consistent with the previous analyses, our model correctly identified much of the internal content (fig 17a; r = .86, p <.001) and also correctly identified much of the external content (fig

17b, r = .67, p <.001). In contrast to previous results, we observed significant misclassification. Internal content was often misclassified as external content (fig 17c, r = .40, p < .001) and external content was often misclassified as internal content (fig 17d, r = .33, p < .001). Even though we found greater correct classification of internal content than misclassification (R^2 = 0.76 vs R^2 = 0.16), misclassification rates were high. We also found greater correct classification of external content than misclassification (R^2 = 0.46.vs R^2 = 0.11), but misclassification of external content is frequent. These results are summarized in the figure below.

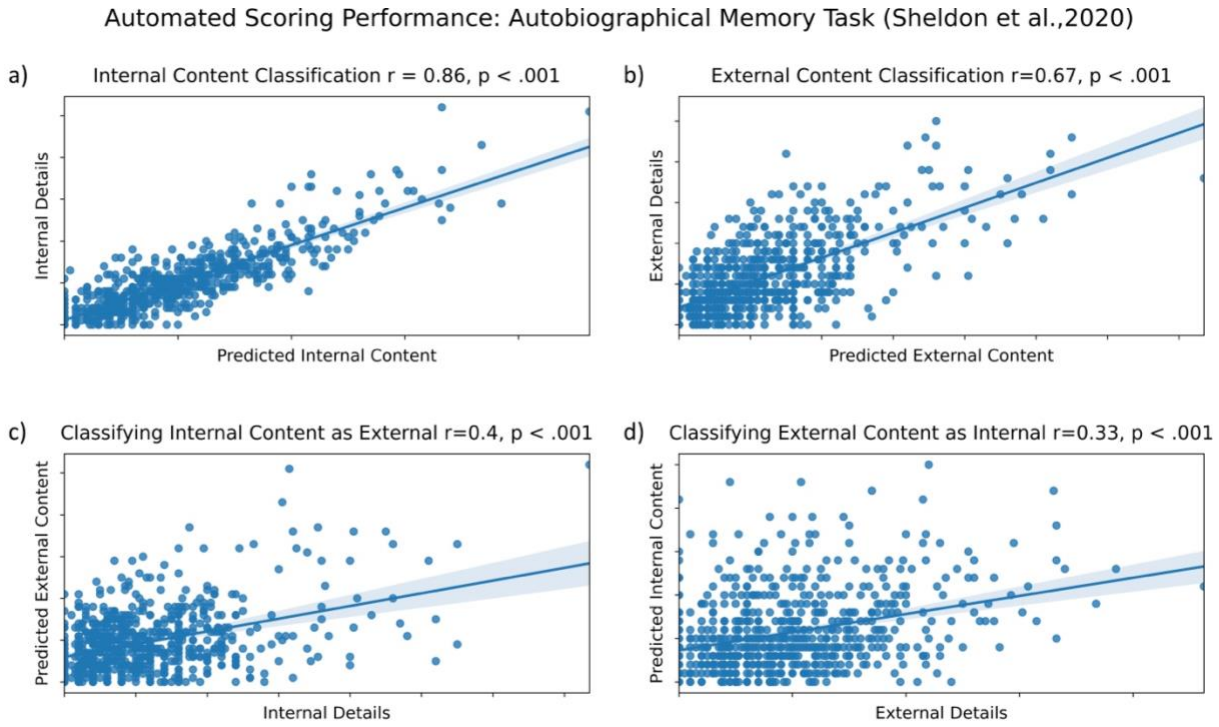

Figure 17. Model performance on autobiographical memory data from Sheldon et al. (2020). Internal (panel A) and external (panel B) content are accurately identified, with significant misclassification of internal details as external (panel C) and external details as internal (panel D).

Narratives Scored with Alternative Scoring Procedures

Results: Picture Description Task (Strikwerda-Brown et al., 2021).

Our model correctly identified much of the internal content (fig 18a; r = .97, p <.001) and external content (fig 18b, r = .67, p <.001) in picture descriptions. However, internal content was often misclassified as external content (fig 18c, r = .41, p = .035) and external content was often misclassified as internal content (fig 18d, r = .34, p = .161). While we found greater correct classification of internal content than misclassification ($R^2 = 0.95$ vs $R^2 = 0.16$), and we found greater correct classification of external content than misclassification ($R^2 = 0.45$.vs $R^2 = 0.11$), misclassification is frequent. These results are summarized in the figure below.

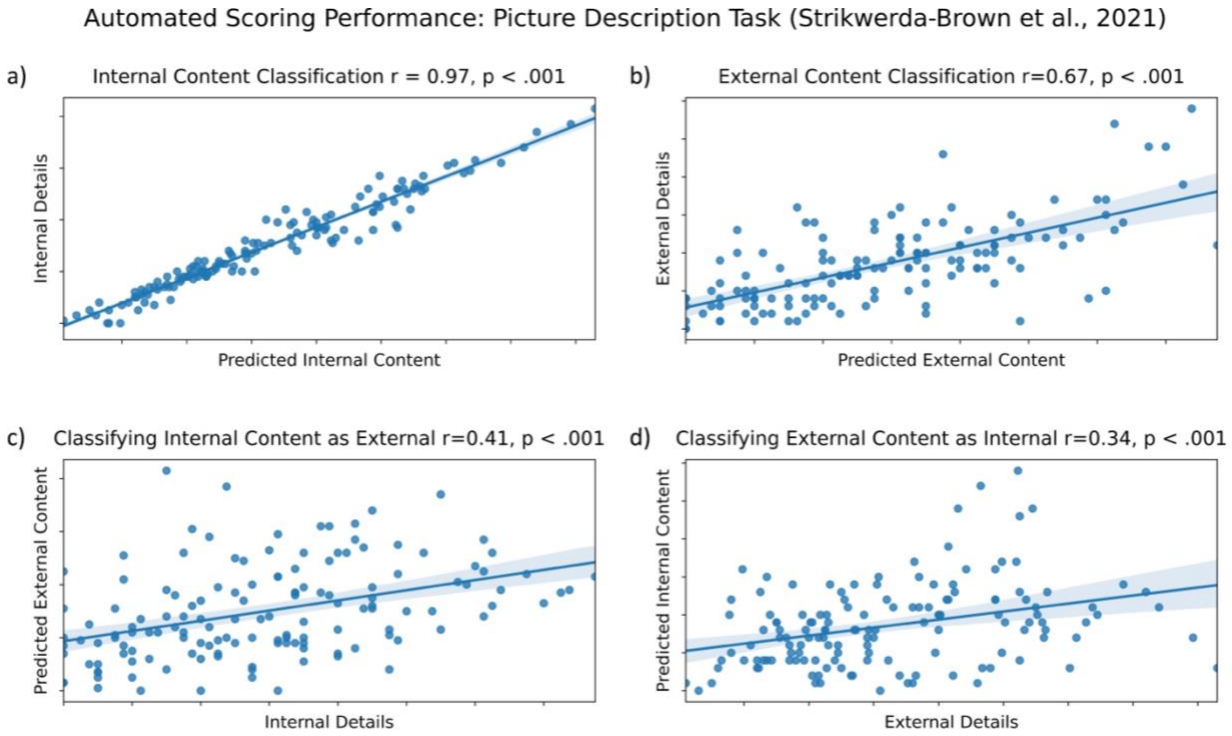

Figure 18. Model performance on picture description data from Strikwerda-Brown et al. (2021). Internal (panel A) and external (panel B) content are accurately identified, with significant misclassification of internal details as external (panel C) and external details as internal (panel D).

Results: Thoughts Task (Strikwerda-Brown et al., 2021)

When participants describe their unconstrained thoughts, our model correctly identified much of the internal content (fig 19a; r = .89, p <.001) and external content (fig 19b, r = .83, p <.001) in the resulting transcribed narratives. Significant misclassification was present. Internal content was often misclassified as external content (fig 19c, r = .29, p < .001), and external content was often misclassified as internal content (fig 19d, r = .50, p <.001). To summarize, even though we found greater correct classification of internal content than misclassification (R^2 = 0.80 vs R^2 = 0.09) and greater correct classification of external content than misclassification (R^2 = 0.68 vs R^2 = 0.25), we observed significant misclassification. These results are summarized in the figure below.

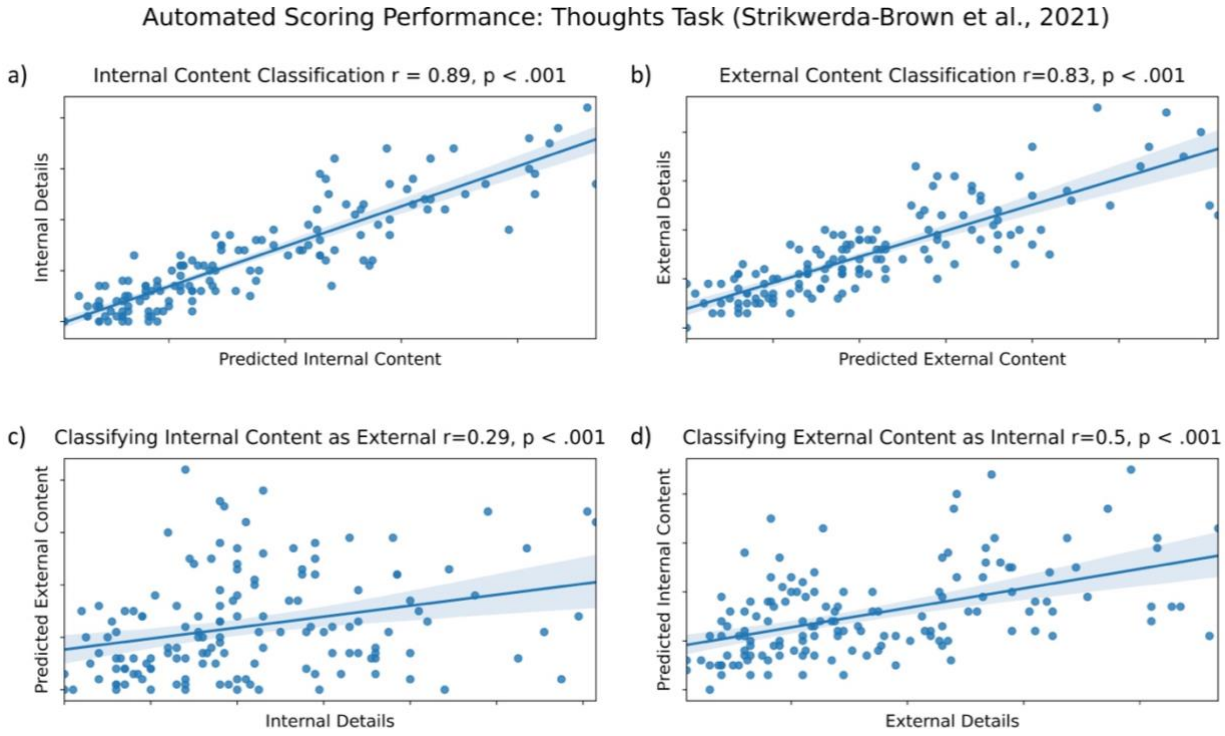

Figure 19. Model performance on thoughts task data from Strikwerda-Brown et al. (2021). Internal (panel A) and external (panel B) content are accurately identified, with significant

(Continued) misclassification of internal details as external (panel C) and external details as internal (panel D).

Discussion

We have described and tested a model-based approach that can automatically score memories, imagined events, and related narrative output for internal and external content. In general, we found that our model performs well across datasets with a variety of tasks in both older and younger populations. The amount of predicted internal content is highly correlated with actual internal detail counts in narratives. Likewise, the amount of predicted external content is highly correlated with actual external detail counts. Importantly, in most of the datasets content misclassification is relatively low: the number of internal details has little relationship to the amount of predicted external content. Likewise, there is no strong relationship between the number of external details and predicted internal content.

However, we found that model performance differed across datasets. Model performance was very good for future simulation narratives from younger adults, future simulation narratives from older adults (Devitt & Schacter, 2018, 2019), creative writing narratives (van Genugten et al., 2021), and memory narratives from Strikwerda-Brown et al (2021) and King et al. (2021). For these datasets, there was little misclassification of content. For musically cued memories (Sheldon et al., 2020), however, we found rates of misclassification that were higher than in the other datasets.

We believe that model performance on data from Sheldon et al. (2020) differed from performance on other datasets because many narratives in this dataset, in contrast to the others, were transcribed 1) without much punctuation and 2) without removing uninformative speech (e.g. ‘I don’t know, we went to, I guess, well,’). Uninformative speech is not a problem for

manual scoring, but it is problematic for our model because this text will still be classified as internal or external content. Narratives with little punctuation are also problematic for our approach because it leads pySBD to split our narratives into sentences incorrectly. Resulting segments of text contained multiple sentences, which makes accurate prediction difficult for the model. While our preprocessing removed very long pieces of text without punctuation, some pieces of text still contained multiple sentences.

To explore whether our model would work on this dataset if it had been transcribed with more punctuation, we manually added punctuation to a random subset of 100 narratives. We did not remove any sentences from these newly punctuated narratives before classification. We found that adding punctuation to the transcriptions led to correct identification of internal content with no detectable misclassification of internal content as external (R^2 = .55 vs R^2 = .00). Our model also correctly identified external content with less misclassification of external content as internal (R^2 = .49 vs R^2 = .07). We would expect even less content misclassification for this dataset if meaningless text were also removed. Together, these results suggest that performance of our model is very good across datasets when punctuation is present in narratives.

We expected performance to be worse on the two tasks scored with different guidelines, because the model was not trained to mimic those guidelines. Indeed, performance was comparatively poor on the thoughts task, in which participants provided the thoughts that came to mind as they looked at a picture, and also on the related picture description task. These results suggest that our automated scoring procedure should not be used, or used with caution, on tasks that are not optimized for scoring with the AI (Levine et al., 2002) or the adapted AI (e.g., Addis, et al., 2008).

Optimal Setting & Potential Limitations

Our model will likely perform best when used with data that are similar to our training data, i.e., when scoring internal and external data from future simulation and autobiographical memory tasks. We believe that the model can also be used for scoring other narrative data, as evidenced by strong performance on the creative writing dataset. However, we do not know how well this code will perform under new circumstances. For example, while the model seems to work well with data from both healthy young and older adults with relatively intact speech, its use with patient populations and populations with more rambling speech is untested. Researchers who want to use this automated scoring approach for new populations should manually score a subset of narratives to verify reliability. We are also unsure how well this model will perform for different dialects and for different language usage more generally. The datasets that we used for fine-tuning and evaluation were collected in the United States and Canada and thus were presumably from WEIRD (Henrich et al., 2010) populations. The data used for training distilBERT (the model we fine-tuned) comes from English Wikipedia text and the bookCorpus (a set of unpublished English books). Accordingly, narratives that use language that is significantly different from the English text found in our fine-tuning datasets or in the pretraining data may or may not be scored as accurately. Researchers who want to use this procedure in new populations can manually score a subset of narratives to confirm accuracy.

There are several situations in which we expect the model to score text differently from the standard AI procedure. First, we expect our code to improperly score narratives that do not have punctuation to mark sentence boundaries, as discussed earlier. Second, we expect narratives that contain several events to be scored differently from the standard AI procedure. With the typical AI scoring, researchers identify a central event, and mark all details in non-central events as external. The current code is not able to identify which details belong to central versus

peripheral events. As a result, the model is likely to identify event-related details as internal, regardless of which event the details came from. Depending on the research question, this feature may or may not matter. Researchers interested in the total amount of episodic and non-episodic content in narratives can use the code as is, while researchers interested in only the central event may have to manually read the narratives and score a subset of them by hand.

Recommendations for Study Design

We provide several design recommendations to maximize the power to detect an effect when using our automated scoring approach. First, we recommend using written narratives to maximize scoring accuracy. Splitting the narrative into sentences will be easiest with written text, so scoring is more likely to be accurate. In addition, memories written by a participant will have less meaningless content (e.g. 'I don't know, we went to, I guess, well, ') than transcribed memories. If transcribed memories include meaningless text, this text will get scored, which adds noise to our internal and external content measures. Further, distilBERT was pretrained on written text, so we expect performance to be best when the evaluation data is also written. While we provide this recommendation, we do not expect performance to drop a great deal for transcribed narratives with enough punctuation. Second, to achieve the same power as a manually scored study, studies using our procedure should recruit larger sample sizes. Third, prompting by the researcher (e.g., 'Is there anything more you can tell me?') should not be transcribed, as this text will be automatically scored as details.

Future Directions

Model Modifications

In our work, we retrained distilBERT for classification. We opted to use distilBERT because of its training speed. Other neural network architectures that perform slightly better on

transfer learning tasks, such as BERT and RoBERTa could be used in future work. Future work could also systematically search for different hyperparameters to improve classification accuracy. We used default hyperparameters for fine-tuning distilBERT.

Scoring Internal and External Detail Subcategories

The approach that we have presented here is useful for automatically calculating the amount of internal and external content in narratives. A second function of the AI is to sort internal and external content into further subcategories. Internal details can be further classified as perceptual, event, time, place, and thought/emotion related. External details can be classified as event (event details from non-central events), semantic (general knowledge or facts), repetition, and other content such as metacognitive statements and editorializing (for more detailed external subcategories, see Strikwerda-Brown et al., 2019).

The same approach that we used for classifying internal and external content could be extended for classifying detail subcategories. That is, researchers could train models to determine what percentage of content in each sentence belongs to each detail subcategory. In an alternative approach, researchers could first split sentences into individual clauses and train a model to classify the resulting clauses into subcategories. Classifying at the clause levels may lead to fewer misclassifications. If researchers are unable to find adequate training data, since only a subset of AI studies score for subcategories, an alternative approach may be to use zero-shot classifiers. These classifiers can attempt to predict the topic, from a list of given topics (e.g. action, time, perceptual, etc.), that a piece of text belongs to (e.g. Yin et al., 2019; for demo, see https://huggingface.co/zero-shot/). These models do not require training data to fine-tune predictions.

Conclusion and Application

We believe that the tool presented here will enable researchers to conduct studies with considerably larger sample size than typically used, and thus perhaps capture smaller effect sizes, using the AI. We believe this tool will also facilitate internet-based research with the AI. This type of research has often been impractical because of the scoring burden that comes from collecting more participants to offset data quality. Internet-based research would allow researchers to study more diverse populations and would allow memory researchers to take advantage of strategies used in other areas of psychology, such as rapidly piloting multiple experiments online. Importantly, the automated scoring procedure will enable research groups that have fewer resources for scoring narratives to also conduct large AI studies.

To accompany this paper, we provide a Colab notebook that researchers can open in their web browser. Researchers can use this notebook to automatically score memories by providing a spreadsheet with narratives. The notebook is intended to be useable out-of-the-box without any additional coding required. This notebook and instructions for using it can be found at https://github.com/rubenvangenugten/. The final model used in this notebook has been retrained on all adapted and standard AI-scored datasets.

Paper 3 Acknowledgments

We thank Aleea Devitt, Signy Sheldon, and Peggy St. Jacques for sharing data with us. We thank Patrick Mair for discussions related to this manuscript. Preparation of this manuscript was supported by National Institute on Aging [grant number R01 AG008441] awarded to DLS.

General Discussion

This dissertation examined how episodic retrieval contributes to tasks that are not traditionally thought of as memory tasks. Further, it explored how the methods used for studying episodic retrieval in memories and imagined situations can be automated to make memory research easier, and to enable researchers to analyze larger samples.

Paper 1 of this dissertation focused on the contributions of episodic retrieval to creative writing. Since episodic retrieval is thought to support the construction of novel imagined situations, we expected that manipulating episodic retrieval prior to a creative writing task would increase the amount of episodic content that participants used during the task. We found mixed support for this hypothesis. Results from the first study strongly supported this claim; results from the second study were consistent with these findings and were trending in the same direction. However, more research will be needed to firmly establish that creative writing benefits from episodic retrieval. We expect that episodic retrieval contributes to scene and event construction during creative writing, so, to maximize detecting an effect of ESI, future studies could instruct participants to focus on these aspects of their stories.

Paper 2 of this dissertation tested whether episodic retrieval contributes to mentalizing when individuals imagine specific events. We expected that imagined scenes and remembered events would inform mentalizing judgments, and therefore hypothesized that boosting episodic retrieval via ESI would increase the number of mental-state related details in responses to the mentalizing task. We also hypothesized that, when asked to label which of their mental state inferences were informed by specific memories, participants would report drawing on particular memories for a substantial number of inferences. We found support for the ESI

hypothesis in Experiment 1. We failed to replicate ESI results from Experiment 2, including on our manipulation check. However, we found that more than half of the thoughts generated by participants during Experiment 2 were labeled as having been informed by a specific past event that they remembered while completing the task. Combined with previous research, these results suggest that episodic retrieval may inform mentalizing. Therefore, further exploring the conditions under which episodic retrieval contributes to mentalizing remains a promising area for future research.

Paper 3 of the dissertation focused on automating the scoring procedures used by the Autobiographical Interview for counting the amount of internal and external content in memories and imagined events (Levine et al., 2002). We found that our approach was able to correctly identify internal content and identify external content with little misclassification. These results suggest that large studies using the Autobiographical Interview can now be conducted with relative ease. Instructions for using our model are available at https://github.com/rubenvangenugten.

Together, these studies raise questions about the role that episodic memory plays in other domains of cognition, and questions about how natural language processing can further be used to study memory.

Theoretical implications

The results presented in this dissertation are consistent with the constructive episodic simulation hypothesis, which suggests that we retrieve elements from episodic memory and recombine them to construct new imagined events (Schacter & Addis, 2007, 2020). Because episodic retrieval can be used to imagine specific events and specific scenes, tasks involving this

type of construction likely benefit from episodic retrieval.

As we think about the role of episodic retrieval in other cognitive domains, it is important to keep in mind that tasks can be solved in multiple ways. Some ways of solving a problem may draw on episodic retrieval while other ways do not. For example, participants coming up with creative uses for an object may recall an autobiographical memory in which they saw an object used creatively (e.g. bricks being used as bookends) and report that use (e.g. Gilhooly et al., 2007). Alternatively, they may rely on semantic memory to generate a new use (e.g. Kenett & Faust, 2019; Mednick, 1962) or use other strategies. These considerations suggest that strategy choice impacts whether episodic retrieval contributes to task performance. Importantly, they also suggest that episodic retrieval is not necessary for tasks in which participants can find alternative approaches for solving a problem. This distinction between episodic retrieval being *necessary* versus *helpful under certain conditions* informs the interpretation of our research as well: episodic retrieval may not be necessary for mentalizing and creative writing, but can aid these processes under certain conditions.

As we discussed in Chapter 3, episodic retrieval may be helpful for mentalizing and social cognition more broadly, even if it is not necessary. Previous work shows that theory of mind does not require episodic retrieval, as patients with memory deficits are able to complete standard theory of mind tasks (Rosenbaum et al., 2007). Indeed, the default networks involved in these two processes are separable (DiNicola et al., 2020). While episodic retrieval is not necessary for theory of mind, episodic retrieval and social cognition cooperate during some tasks. For example, work by Gaesser et al. (2018) shows that imagining an event in a familiar place increases empathy judgments relative to imagining the event in a less vivid location. In

addition, theory of mind sometimes elicit scene construction as well. For example, social situations are imagined with more specificity and imagery (e.g. Andrews-Hanna et al., 2013) than non-social situations, and memory regions such as the hippocampus are active during theory of mind judgments when the task involves richly imagining others (e.g. Rabin et al., 2010). Together with our work, these observations suggests that episodic retrieval can cooperate with or aid mentalizing processes in some situations.

Like theory of mind, creative writing does not require episodic retrieval. But, specific types of creative writing likely benefit from episodic retrieval. In our experiment, we asked participants to continue writing stories in the way that felt most natural to them. Some participants wrote long stories involving primarily external content. Elements from these stories, such as the background of a character, can be generated without episodic retrieval because they depend on semantic retrieval and other processes. Other stories involved descriptions of imagined scenes. Because theoretical work suggests that scene construction requires episodic retrieval (e.g. Schacter & Addis, 2020), this component of creative writing likely requires episodic retrieval. As a result, not all creative writing may require episodic simulation, but we expect that future studies will show it is helpful for specific types of creative writing.

Future directions

Exploring the role of episodic retrieval in other domains

This discussion raises questions about which other tasks may benefit from episodic retrieval. In exploring this topic, Szpunar, Spreng, & Schacter (2014) provide clear hypotheses about the role that semantic and episodic retrieval play in four types of future thinking:

simulation (imagining a specific or abstract hypothetical event), prediction (estimating the likelihood that something will happen), intention (setting a goal), and planning. Szpunar et al. conceptualize these types of future thinking as existing on a continuum from fully episodic to fully semantic. For example, they suggest that episodic planning involves imagining a series of autobiographical steps (e.g. planning steps to prepare for a test), whereas semantic planning may involve imagining more abstract steps (e.g. financial planning for a company). New work could directly test the hypotheses presented in Szpunar et al. (2014). If these four types of future thinking exist on a continuum from semantic to episodic, an ESI should have no effect on task behavior for semantic future thinking trials, some effect on hybrid future thinking trials (i.e., tasks that tap both episodic and semantic retrieval), and a larger effect on fully episodic future thinking trials. For example, an ESI should impact the number of steps participants provide in response to an autobiographical planning prompt, a smaller impact on the number of steps participants provide in response to a mixed episodic and semantic planning prompt, and no impact on the number of steps in response to a semantic planning prompt. To test the hypotheses laid out by Szpunar et al. (2014), researchers could design studies combining the ESI procedure with different kinds of future thinking conditions. We should note that some types of future thinking have already been examined with ESI (e.g. Madore et al., 2014; Madore & Schacter, 2014), though no ESI experiment has included fully episodic, fully semantic, and hybrid trials in the same study to directly test the hypotheses from Szpunar et al. (2014).

Adapting experimental procedures to streamline ESI studies

The approach suggested in the previous section would improve our understanding of episodic retrieval and the role it plays in future thinking. However, the amount of time and

effort required to test these hypotheses would be significant, since the necessary studies would involve at least 12 conditions (four types of future thinking with three conditions per task: fully episodic, fully semantic, and hybrid trials). Significantly more conditions would be needed if more than one task were used per future thinking type. To test these (and other) hypotheses with reasonable time and effort, researchers could adapt current ESI study protocols. As a reminder, a typical experiment involves individuals participating in the ESI (or a control induction), completing the task of interest, and then completing the manipulation check, which involves describing memories or imagined future events. The ESI is administered by a researcher, and the task of interest and manipulation check are scored manually. To streamline these manual components of ESI studies, researchers could use an online alternative to the ESI and automate scoring of narratives using the procedure of Paper 3.

Vollberg et al. (2021) discussed an online episodic retrieval manipulation that produced results in their experiment similar to those produced by the ESI in other studies. In their episodic retrieval manipulation, participants were asked to imagine an event in as much detail as possible, before completing the task of interest. In a control task, participants completed a series of math problem, before they engaged in the task of interest. This manipulation was based on work by Rudoy et al. (2009). Additional validation of this episodic retrieval manipulation (akin to the validation conducted by Madore et al., 2014) would be necessary before widespread use. However, an automated ESI and automated scoring of narratives would make each of the hypotheses discussed by Szpunar et al. easier to test in a series of large online experiments.

Using natural language processing to support memory research

In addition to exploring the role of episodic retrieval in new domains, this dissertation demonstrated that natural language processing can be used to build tools to study memories. As discussed in Paper 3, additional language models could be built to identify other types of content that is present in each narrative, such as the amount of spatial information, the amount of perceptual information, and the amount of time information. Tools could also be built to automatically score events according to common ratings, such as vividness, perceptual richness (Levine et al., 2002), and overall quality judgments (Hassabis et al., 2007).

Recent research illustrates that existing natural language processing tools can also be used to study memory. These tools may be especially useful for studying emotional memories because sentiment analysis (e.g. Socher et al., 2013) allows researchers to automatically label the emotional valence in a piece of text. For example, ongoing work by Sanson et al. (2022) and others is using sentiment analysis to explore the emotional trajectories present in memories. In addition to identifying valence, language models are also able to identify specific emotions in text (e.g. Abdul-Mageed & Ungar, 2017). So, we expect future work to take advantage of these existing tools to study emotional memories as well.

Another promising linguistic measure for studying memories is known as semantic distance. Semantic distance quantifies how dissimilar two pieces of text are, with many models providing scores that closely correspond to dissimilarity judgments from human raters (e.g. Pennington et al., 2014). This measure is especially promising for studying encoding or retrieval of complex material. For example, Chen et al. (2017) showed that the semantic similarity of recall transcripts for different events was related to those events' neural similarity in the posterior medial cortex during retrieval. Semantic distance may be especially useful in

naturalistic fMRI studies, where complex language is present in movies, narratives, or audio. By quantifying semantic distance, researchers can study the encoding and retrieval processes used for naturalistic stimuli as a function of stimuli similarity. While the use of natural language processing to study memory and the brain is relatively new, this approach has a proven track record in other areas such as linguistics (e.g., Schrimpf et al., 2021) and the study of decoding cognitive states from fMRI (e.g., Pereira et al., 2018).

Exploring the link between social cognition and episodic simulation

We discussed the possible contributions of episodic simulation to mentalizing in our second paper. Because much of our imagination is social in nature (e.g. D'Argembeau et al., 2011), it is likely that episodic simulation interacts with other aspects of social cognition as well (e.g. Spreng, 2013; Spreng & Mar, 2012). For example, as we imagine events, such as playing soccer in a park or walking around at a barbeque, we must represent the locations of others in space and simulate their movement. Simulation of others in space is an especially fruitful area for future research at the intersection of episodic simulation and social cognition. However, even though simulating the movement of others through space is critical to imagining social events, little is known about how we do this.

Existing human and animal research suggest interesting hypotheses for how we simulate others moving through space. We likely engage the brain regions involved in our own navigation, such as the hippocampus, parahippocampal gyrus, posterior cingulate cortex, medial frontal gyrus, and inferior parietal lobe (e.g., Boccia et al., 2014) for imagining the movement of others as well. We imagine others from an allocentric rather than egocentric perspective, so literature on allocentric and egocentric navigation may provide additional

predictions for the brain regions used to simulate others (T. Brown, personal communication, September 23, 2021). For example, regions such as the posterior parietal cortex, which are involved in egocentric navigation but not allocentric navigation (e.g. Ciaramelli et al., 2010), may be less important for simulating the movement of others. Animal work also provides suggestions for how others may be represented in space. For example, Omer et al. (2018) show that a subset of the hippocampal cells that represent a bat's location in space also fire when a conspecific flies through the same location. These neurons do not fire when an object watched by the bat flies through the same location. These results suggest that animals may represent others in space the same way that they represent themselves in space: with place cells.

Future work could test whether humans simulate others through space by (1) using brain regions implicated in allocentric navigation and (2) using similar hippocampal representations to encode one's own and others' locations. One particularly promising paradigm to test these hypotheses comes from Brown et al., (2016). Brown et al. used virtual navigation to study goal and location representations as participants imagined walking through space. On each trial in the fMRI session, a participant was shown where they would start on a circular track. The participant was then asked to imagine how they would navigate to a specific goal location. Brown et al. showed that the hippocampal patterns for locations between the start and end location are active during this planning period. These results are consistent with Johnson & Redish (2007), who suggest that rodents mentally simulate paths by activating place cells that code for locations on those paths.

An adaptation of Brown et al.'s paradigm could provide a way to examine the neural representation we use to imagine others moving through space. Specifically, researchers could

add simulation trials in which participants imagine others moving along their track. To mimic work from Omer et al. (2018), additional simulation trials could be included in which the participant imagines an object moving along the track. Researchers could then test whether representations for imagining the self in a location are reused for imagining others in a location. Specifically, similar hippocampal patterns are expected to be used for representing the self and other in each location. To test if these patterns are specific to imagining individuals (as Omer et al., 2018, suggest with their work on 'social place cells'), researchers could test whether self and other patterns for each location are more similar to each other than self and object patterns. Last, a univariate contrast between self and baseline, other and baseline, and other and self are expected to reveal that similar brain regions are used for imagined navigation for self and others, with notable differences in brain regions implicated in allocentric versus egocentric navigation. This approach, or other study designs, could be used to explore how we imagine others moving through space.

Conclusion

In this dissertation, we showed that episodic retrieval may play a role in other forms of cognition not typically associated with memory: creative writing and mentalizing. Future work should continue to investigate the link between episodic retrieval and these forms of cognition, and we provide recommendations in this dissertation for doing so. In addition, we demonstrated an approach for automatically scoring internal and external content in autobiographical memories and future simulations. We end by discussing further intersections of episodic retrieval and imagination as well as promising avenues through which natural

language processing can contribute to memory research. Overall, this research helps inform our understanding of the function and measurement of episodic memory.

Appendix: Paper 1

Story prompts (modified from Tamir et al., 2015)

1. Under the trees several pheasants lay about, their rich plumage dabbled with blood; some were dead, some feebly twitching a wing, some staring up at the sky, some pulsating quickly, some contorted, some stretched out—all of them writhing in agony except the fortunate ones whose tortures had ended during the night. Tess's first thought was to put the still living birds out of their torture, and to this end with her own hands she broke the necks of as many as she could find, leaving them to lie where she had found them.
   Thomas Hardy, *Tess of the D'Ubervilles* (1891)
2. He dreamed that the priest whom they had shot that morning was back in the house dressed in the clothes his father had lent him and laid out stiffly for burial. The boy sat beside the bed and his mother read out of a very long book: there was a fish basket at her feet, and the fish were bleeding, wrapped in her handkerchief. He was very bored and very tired and somebody was hammering nails into a coffin in the passage. Suddenly the dead priest winked at him—an unmistakable flicker of the eyelid, just like that.
   - Graham Greene, *The Power and the Glory* (1940)
3. Lloyd shoves off the bedcovers and hurries to the front door in white underwear and black socks. He steadies himself on the knob and shuts his eyes. Chill air rushes under the door; he curls his toes. But the hallway is silent. Only high-heeled clicks from the floor above. A shutter squeaking on the other side of the courtyard. His own breath, whistling in his nostrils, whistling out. Faintly, a woman's voice drifts in. He clenches his eyelids

tighter, as if to drive up the volume, but makes out only murmurs, a breakfast exchange between the woman and the man in the apartment across the hall.

- Tom Rachman, *The Imperfectionists* (2010).

4. My brother was already in school by the time I was born, and my earliest memory is of Jimmy going to school every day, leaving me to think of the future when I could go to big school myself. In the afternoons I would press my nose against the picture window in the den, watching for the big yellow school bus and listening for the screech of air brakes as the bus stopped at the top of the hill to deliver Jimmy home.

   - Cindi Rigsbee, *Finding Mrs. Warnecke* (2010).

5. Meru is a hydra-headed massif, with multiple summits; our goal was to climb the most dramatic of these, a blade of pale, steep granite aptly named the Shark's Fin. But on this afternoon the weather had turned nasty, and we were afforded little rest. Hammered by high winds, our entire world bucked wildly against the cams and pitons holding us to the wall. The ice we'd climbed to reach this point wasn't particularly solid, a bad sign for what lay ahead.

   - Conrad Anker, "Why Am I Here Again?" *Outside* (April 2009)

6. He dropped his oars and felt the weight of the small tuna's shivering pull as he held the line firm and commenced to haul it in. The shivering increased as he pulled in and he could see the blue back of the fish in the water and the gold of his sides before he swung him over the side and into the boat. He lay in the stern in the sun, compact and bullet shaped, his big, unintelligent eyes staring. The old man hit him on the head for kindness and kicked him, his body still shuddering, under the shade of the stern.

   - Ernest Hemingway, *The Old Man and the Sea* (1952)

7. Entering through a window, I gathered up all the household chemicals, and, believe me, he had a lot, more than I did, more than he needed, thinner, paint, lye, gas, solvents, etc. I got it all in like nine Hefty bags and was just starting up the stairs with the first bag when here comes the whole damn family, falling upon me, even his kids, whipping me with coat hangers and hitting me with sharp-edged books and spraying hair spray in my eyes, the dog also nipping at me, and rolling down the stairs of the basement I thought, They are trying to kill me.

   - George Saunders, "Adams," *In Persuasion Nation* (2006)

8. Lily, the caretaker's daughter, was literally run off her feet. Hardly had she brought one gentleman into the little pantry behind the office on the ground floor and helped him off with his overcoat than the wheezy hall-door bell clanged again and she had to scamper along the bare hallway to let in another guest. Miss Kate and Miss Julia were there, gossiping and laughing and fussing, walking after each other to the head of the stairs, peering down over the banisters and calling down to Lily to ask her who had come.

   - James Joyce, "The Dead," *The Dubliners* (1914)

9. John Reed was a schoolboy of fourteen years old: large and stout for his age, with a dingy and unwholesome skin; thick lineaments in a spacious visage, heavy limbs and large extremities. He gorged himself habitually at table, which made him bilious, and gave him a dim and bleared eye and flabby cheeks. He ought now to have been at school; but his mama had taken him home for a month or two, on account of his delicate health.

   - Charlotte Brontë, *Jane Eyre* (1847)

10. Roger gathered a handful of stones and began to throw them. Yet there was a space round Henry, perhaps six yards in diameter, into which he dare not throw. Here, invisible yet strong, was the taboo of the old life. Round the squatting child was the protection of parents and school and policemen and the law.

    - William Golding, *Lord of the Flies* (1954)

Example Responses with Originality Ratings

Story Prompt

My brother was already in school by the time I was born, and my earliest memory is of Jimmy going to school every day, leaving me to think of the future when I could go to big school myself. In the afternoons I would press my nose against the picture window in the den, watching for the big yellow school bus and listening for the screech of air brakes as the bus stopped at the top of the hill to deliver Jimmy home.

Example Response with Low Originality Rating Response

I was so excited for Jimmy to get home. I would ask him questions about what he learned, how his teachers and friends are like, what he eats for lunch, and so on. He wasn't too excited to talk to school when he got back which was disappointing but I understand now after going to school myself. I have to wake up early at 7am to get on the school bus and I love sleeping in. When I get to school, there's assembly and we have five classes everyday - Math, English, Science, Social Studies, and Spanish. Math is hard and my teacher is not so nice. She gives us so much homework everyday and I am struggling. Thankfully I have my big brother Jimmy to help me with my homework when I get stuck. The only part of school that I enjoy is lunch break. We have a cafeteria in school and the menu changes daily. Some

of my friends don't like the food and would rather bring their own lunch but I'm not a picky eater and I think the food at the cafeteria isn't bad. After lunch, me and my friends usually play soccer or dodge ball. I wish lunch break was longer. After lunch, we have more classes and it's sometimes hard for me to focus because I get sleepy after eating food. On Monday, Wednesday, and Friday, I have volleyball practice. Volleyball is fun but it can be stressful and tiring sometimes. My arms get bruised and it's hard for me to

Example Response with Medium Originality Rating

I wanted to be just like Jimmy when I was younger. I wanted to follow him around, go to school and meet his friends. When he was in middle school, I would mimic his mannerisms and habits so that he would think I was cool enough to hang out with his friends. When he was in high school, and I was in seventh grade, I would try to tell him about my "girlfriends" to show him that I was really mature for my age. My image of Jimmy was that he was perfect. He was well-liked by his friends and teachers, a successful football player and did well in every class. However, I did not realize until later that this was all just an image he created for us to see to please us.

Only in the past couple of years has Jimmy really opened up to me. He told me about how he struggled with his self-esteem and while outwardly he seemed content, he was often not. Depression, he told me, is like a cut so deep that you feel like you are always bleeding even if no one can actually see it. He feels like his emotions were always seeping out of him and that it made

Example Response with High Originality Rating

> Even then, he walked like he does now: slow, loping; you'd think of panthers, or the hunters that hang their heads on walls. He was never made for Kansas, I think. The squareness of the state extends to the people, men built like refrigerators… all-muscle oxen squared off next to barns and silos that barely last the winter. Jimmy was softer on the edges. You'd almost say graceful.
>
> When he talked about California for the first time, we were eight and fourteen, and crowded around the woodstove waiting for our parents to get home. Of course I'd studied the state in passing, heard about its voting habits now as new election cycles rolled around, but to me it seemed far-off and mystical. I'd heard, vaguely, of New Age, so I imagined that they'd come up with a different

Example Responses with Detail Scoring

The following condensed examples contain both internal and external details. Internal details are event and scene details, which includes the objects, actions, locations, thoughts of people in the scene, and other similar details. External details are details that are not specific to an event and are mostly made up of factual information. In some cases, they are used to provide context to the story. In other cases, they are the main focus of the story. In the examples below, details are separated with a forward slash, and the external details are surrounded in square brackets. For more information on how pieces of text are separated into details, see Levine et al. (2002).

Example 1:

[I was 4/ when it happened. / The day/ started like every other one/.] Momma /yelled /at Jimmy/ for 20 minutes/ to get out of bed,/ or else he's be late /for the bus./ [It always went like this /- Jimmy/ didn't wake up/ particularly well/.] Momma /sent me/ into his room /to wake him/ so, I jumped /on top of him,/ screaming /in his ear /"Momma /says /up/! Time to get up/!" Groaning, /he playfully /pushed me off/. "Tell /mom/ I'm on my way down./" Ten minutes later /he ran down/, slinging/ his bag /over his shoulder,/ and grabbed/ a granola bar /from the tin/ on his way out/ and onto the bus /that took him /to school/.

[Momma /would receive a call/ at about 20 past 4./ We never saw it coming/]

Example 2:

Why was the priest / winking at him/?, he thought./ Dead people/ don't blink! / Maybe he wasn't really dead?/

[So, why did they shoot/ the priest?/ It's because he was really a bad man/. Priests /can be bad men too/. They found that out/ the hard way/. They started out/ trusting/ a man of the cloth/, because they are usually good men/. Some of the best/. But this was a bad man/. He had swindled people/ out of money/]

Appendix: Paper 2

Scoring: Mentalizing Trial

Details are separated with forward slashes. Brackets surround external content.

Prompt:

Your friend has taken to boxing. She is about to enter her first real match.

What could she be thinking?

Response:

She could be thinking, I could get hurt/ because it's very physical/ [which I don't know much about/, but I know it's physical./] She might be worried/ about who she is competing against/. Or maybe she is excited/ about who she could be competing against/…

Scores:

Internal detail count: 6

External detail count: 2

Internal word count: 34

External word count: 11

Scoring: Episodic Simulation Trial

Prompt:

Imagine Near Future Event

Your first ever boxing match

Response:

I am a bit dizzy/ because I feel the adrenaline/. I am wearing red/ gloves/. I am in an arena/. I see people cheering me on/. My roommate is here to cheer me on/. [My roommate has always been supportive of me/. She's more supportive than my family/] …

Scores:

Internal detail count: 7

External detail count: 2

Internal word count: 34

External word count: 14

Appendix: Paper 3

Example scoring

During the AI, a participant might provide a response such as the following:

> We went to the beach because it was my birthday. I was sitting on the beach with a beer. We had driven up last night. We ran into the surf.

After applying the rules described in the scoring manual, we obtained the scored response below. We used forward slashes to separate details and we have surrounded the external details with brackets.

> [We went to the beach because it was my birthday/]. I was sitting on the beach/ with a beer /. [We had driven up last night/]. We ran into the surf/.

We then counted the details to summarize the narrative: 2 external details and 3 internal details. If we were using our automated approach, we would summarize the narrative by counting the number of words in external segments (16 words in the brackets) and internal segments (14 words outside of the brackets).

References

Abdul-Mageed, M., & Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 718–728.

Addis, D. R., Musicaro, R., Pan, L., & Schacter, D. L. (2010). Episodic simulation of past and future events in older adults: Evidence from an experimental recombination task. *Psychology and Aging*, *25*(2), 369-376.

Addis, D. R., Pan, L., Musicaro, R., & Schacter, D. L. (2016). Divergent thinking and constructing episodic simulations. *Memory*, *24*(1), 89-97.

Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, *47*(11), 2222-2238.

Addis, D.R., Wong, A.T. & Schacter, D.L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*, 1363-1377.

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. *Psychological Science, 19*, 33–41.

Andrews-Hanna, J., Kaiser, R., Turner, A., Reineberg, A., Godinez, D., Dimidjian, S., & Banich, M. (2013). A penny for your thoughts: Dimensions of self-generated thought content and relationships with individual differences in emotional wellbeing. *Frontiers in Psychology*, *4*.

Andrews-Hanna, J. R., Saxe, R., & Yarkoni, T. (2014). Contributions of episodic retrieval and mentalizing to autobiographical thought: Evidence from functional neuroimaging, resting-state connectivity, and fMRI meta-analyses. *NeuroImage*, *91*, 324–335.

Azunre, P. (2021). *Transfer learning for natural language processing*. Simon and Schuster.

Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, *20*(2), 87-95.

Beaty, R. E., Chen, Q., Christensen, A. P., Kenett, Y. N., Silvia, P. J., Benedek, M., & Schacter, D. L. (2020). Default network contributions to episodic and semantic processing during divergent creative thinking: A representational similarity analysis. *NeuroImage*, *209*, 116499.

Beaty, R.E. & Johnson, D.R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53,* 757-780.

Beaty, R.E., Thakral, P.P., Madore, K.P., Benedek, M., & Schacter, D.L. (2018). Core network contributions to remembering the past, imagining the future, and thinking creatively. *Journal of Cognitive Neuroscience, 30,* 1939-1951.

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A.C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*, 341-349.

Benoit, R.G. & Schacter, D.L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia,75,* 450-457.

Boccia, M., Nemmi, F., & Guariglia, C. (2014). Neuropsychology of environmental navigation in humans: Review and meta-analysis of FMRI studies in healthy participants. *Neuropsychology Review*, *24*(2), 236–251.

Braga, R. M., & Buckner, R. L. (2017). Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron*, *95*(2), 457-471.e5.

Brown, T. I., Carr, V. A., LaRocque, K. F., Favila, S. E., Gordon, A. M., Bowles, B., Bailenson, J. N., & Wagner, A. D. (2016). Prospective representation of navigational goals in the human hippocampus. *Science*, *352*(6291), 1323–1326.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*(1), 1-38

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*, 49–57.

Carpenter, A.C. & Schacter, D.L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 335-349.

Carpenter, A.C. & Schacter, D.L. (2018). False memories, false preferences: Flexible retrieval mechanisms supporting successful inference bias novel decisions. *Journal of Experimental Psychology: General, 147,* 988-1004.

Christensen, R.H.B. (2015). ordinal - Regression Models for Ordinal Data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

Ciaramelli, E., Rosenbaum, R. S., Solcz, S., Levine, B., & Moscovitch, M. (2010). Mental space travel: Damage to posterior parietal cortex prevents egocentric navigation and reexperiencing of remote spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 619–634.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125.

D'Argembeau, A., Renaud, O., & Van der Linden, M. (2011). Frequency, characteristics and functions of future-oriented thoughts in daily life. *Applied Cognitive Psychology*, *25*, 96–103. h

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401-2414.

Dede, A.J., Wixted, J.T., Hopkins, R.O., & Squire, L.R. (2016). Autobiographical memory, future imagining, and the medial temporal lobe. *Proceedings of the National Academy of Sciences*, *113*, 13474-13479.

Devitt, A. L., & Schacter, D. L. (2018). An optimistic outlook creates a rosy past: the impact of episodic simulation on subsequent memory. *Psychological Science, 29*, 936–946.

Devitt, A. L., & Schacter, D. L. (2020). Looking on the bright side: aging and the impact of emotional future simulation on subsequent memory. *The Journals of Gerontology: Series B, 75*, 1831–1840.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*. http://arxiv.org/abs/1810.04805

DiNicola, L. M., Braga, R. M., & Buckner, R. L. (2020). Parallel distributed networks dissociate episodic and social functions within the individual. *Journal of Neurophysiology*, *123*(3), 1144–1179.

Ditta, A.S. & Storm, B.C. (2018). A consideration of the seven sins of memory in the context of creative cognition. *Creativity Research Journal*, *30*, 402-417.

Duff, M.C., Kurczek, J., Rubin, R., Cohen, N.J., & Tranel, D. (2013). Hippocampal amnesia disrupts creative thinking. *Hippocampus*, *23*, 1143-1149.

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, *30*(1), 527–554.

Epley, N., & Waytz, A. (2010). Mind perception. In *Handbook of social psychology, Vol. 1, 5th ed* (pp. 498–541). John Wiley & Sons, Inc.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics,* 363–370.

Fisher, R.P., & Geiselman, R.E. (1992). *Memory enhancing techniques for investigative*

interviewing: The cognitive interview. Springfield, IL: Charles C. Thomas Publisher.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129-139.

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, *53*, 559-575.

Gaesser, B. (2020). Episodic mindreading: Mentalizing guided by scene construction of imagined and remembered events. *Cognition*, 203, 104325.

Gaesser, B., Keeler, K., & Young, L. (2018). Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind. *Cognition*, 171, 180-193.

Gaesser, B., Sacchetti, D. C., Addis, D. R., & Schacter, D. L. (2011). Characterizing age-related changes in remembering the past and imagining the future. *Psychology and Aging, 26*(1), 80-84.

Gaesser, B. & Schacter, D.L. (2014). Episodic simulation and episodic memory can increase intentions to help others. *Proceedings of the National Academy of Sciences USA, 111,* 4415-4420.

Gerlach, K. D., Spreng, R. N., Gilmore, A. W., & Schacter, D. L. (2011). Solving future problems: default network and executive activity associated with goal-directed mental simulations. *Neuroimage*, *55*(4), 1816-1824.

Gilbert, D. T., & Wilson, T. (2007). Prospection: Experiencing the future. *Science*, *317*, 1351–1354.

Gilhooly, K.J., Fioratou, E., Anthony, S.H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98*, 611-625.

Guilford J.P. (1967) The nature of human intelligence. New York: McGraw Hill.

Hassabis, D., Kumaran, D., Vann, S.D., & Maguire, E.A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104*, 1726-1731.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, *11*(7), 299-306.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences, 33*, 61-83.

Ingvar, D. H. (1985). "Memory of the future": an essay on the temporal organization of

conscious awareness. *Human neurobiology*, *4*(3), 127-136.

Irish, M., Addis, D. R., Hodges, J. R., & Piguet, O. (2012). Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. *Brain*, *135(7),* 2178-2191.

Irish, M., Hornberger, M., Lah, S., Miller, L., Pengas, G., Nestor, P. J., Hodges, J. R., & Piguet, O. (2011). Profiles of recent autobiographical memory retrieval in semantic dementia, behavioural-variant frontotemporal dementia, and Alzheimer's disease. *Neuropsychologia, 49*, 2694–2702.

Irish, M., & Piguet, O. (2013). The pivotal role of semantic memory in remembering the past and imagining the future. *Frontiers in Behavioral Neuroscience*, *7*, 27.

Jing, H. G., Madore, K. P., & Schacter, D. L. (2016). Worrying about the future: An episodic specificity induction impacts problem solving, reappraisal, and well-being. *Journal of Experimental Psychology: General*, *145*(4), 402-418.

Jing, H.G., Madore, K.P., & Schacter, D.L. (2017). Preparing for what might happen: an episodic specificity induction impacts the generation of alternative future events. *Cognition*, *169*, 118-128.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27*(45), 12176–12189.

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, *23*(4), 271–274.

King, C.I., Romero, A.S.L., Schacter, D.L., & St. Jacques, P.L. (2021). The influence of shifting perspective on episodic and semantic details during autobiographical memory recall. Submitted for publication.

Krienen, F. M., Tu, P.-C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *The Journal of Neuroscience*, *30*(41), 13906–13915.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1-26.

Levine, B. (2021) Memory. The Levine Lab. Retrieved November 9, 2021 from https://levinelab.weebly.com/memory.html

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging, 17*, 677–689.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*.

http://arxiv.org/abs/1907.11692

Madore, K.P., Addis, D.R., & Schacter, D.L. (2015). Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. *Psychological Science*, *26*, 1461-1468.

Madore, K. P., Gaesser, B., & Schacter, D. L. (2014). Constructive episodic simulation: Dissociable effects of a specificity induction on remembering, imagining, and describing in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 609-622.

Madore, K.P., Jing, H.G., & Schacter, D.L. (2016). Divergent creative thinking in young and older adults: Extending the effects of an episodic specificity induction. *Memory & Cognition*, *44*, 974-988.

Madore, K.P., Jing, H.G., & Schacter, D.L. (2019). Episodic specificity induction and scene construction: Evidence for an event construction account. *Consciousness and Cognition,*

Madore, K. P., & Schacter, D. L. (2014). An episodic specificity induction enhances means-end problem solving in young and older adults. *Psychology and Aging, 29*(4), 913-924.

Madore, K.P., & Schacter, D.L. (2016). Remembering the past and imagining the future: Selective effects of an episodic specificity induction on detail generation. *The Quarterly Journal of Experimental Psychology*, *69*(2), 285-298.

Madore, K.P., Szpunar, K.K., Addis, D.R., & Schacter, D.L (2016) Episodic specificity induction impacts activity in a core brain network during construction of imagined future experiences. *Proceedings of the National Academy of Sciences*, *113,* 10696-10701.

Madore, K.P., Thakral, P.P., Beaty, R.E., Addis, D.R., & Schacter, D.L. (2019). Neural mechanisms of episodic retrieval support divergent creative thinking. *Cerebral Cortex*, *29*, 150-166.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, *315*(5810), 393-395.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*, 220–232.

Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research*, *1079*(1), 66–75.

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1309–1316.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315.

Nusbaum, E.C., Silvia, P.J., & Beaty, R.E. (2014). Ready, set, create: What instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 423-432.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, *16*(10), 1746-1772.

Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji, K., Suzuki, K., ... & Yamadori, A. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *Neuroimage*, *19*(4), 1369-1380.

Omer, D. B., Maimon, S. R., Las, L., & Ulanovsky, N. (2018). Social place-cells in the bat hippocampus. *Science*, *359*(6372), 218–224.

Pennington, J., Socher, R., & Manning, C D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP)*, *12*, 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Peters, J., Wiehler, A., & Bromberg, U. (2017). Quantitative text feature analysis of autobiographical interview data: Prediction of episodic details, semantic details and temporal discounting. *Scientific Reports, 7*, 14989.

Plucker, J.A., & Beghetto, R.A. (2004). Why creativity is domain general, why it looks domain specific, and why the distinction does not matter. In R.J. Sternberg, E.L. Grigorenko, & J.L. Singer (Eds.), *Creativity: From potential to realization* (p. 153–167). American Psychological Association.

Rabin, J. S., Gilboa, A., Stuss, D. T., Mar, R. A., & Rosenbaum, R. S. (2010). Common and unique neural correlates of autobiographical memory and theory of mind. *Journal of Cognitive Neuroscience*, *22*(6), 1095–1111.

Race, E., Keane, M. M., & Verfaellie, M. (2011). Medial temporal lobe damage causes deficits in episodic memory and episodic future thinking not attributable to deficits in narrative construction. *Journal of Neuroscience, 31*, 10262–10269.

Rameson, L. T., Morelli, S. A., & Lieberman, M. D. (2012). The neural correlates of empathy: experience, automaticity, and prosocial behavior. *Journal of Cognitive Neuroscience*, *24*(1), 235–245.

Rosenbaum, R. S., Stuss, D. T., Levine, B., & Tulving, E. (2007). Theory of mind is independent of episodic memory. *Science*, *318*(5854), 1257.

Rudoy, J. D., Weintraub, S., & Paller, K. A. (2009). Recall of remote episodic memories can appear deficient because of a gist-based retrieval orientation. *Neuropsychologia*, *47*(3), 938–941.

Sadvilkar, N., & Neumann, M. (2020). PySBD: Pragmatic sentence boundary disambiguation. *ArXiv*. http://arxiv.org/abs/2010.09657

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*. ArXiv:1910.01108.

Sanson, M., Devitt, A. L., Bonning, E., Moffat, K., Calude, A. S., Van Genugten, R., Rasmussen, A. S. & Garry, M. (in prep.) Tracing the Emotional Trajectories of Autobiographical Memories.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72.

Schacter, D.L. & Addis, D.R. (2007a). Constructive memory: The ghosts of past and future. *Nature*, *445*, 27.

Schacter, D. L., & Addis, D. R. (2007b). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*, 773-786.

Schacter, D.L. & Addis, D.R. (2020). Memory and imagination: Perspectives on constructive episodic simulation. In A. Abraham (Ed.), *The Cambridge Handbook of the Imagination*. (pp. 111-131). Cambridge: Cambridge University Press.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*, 657-661.

Schacter, D.L., Addis, D.R., Hassabis, D., Martin, V.C., Spreng, R.N., & Szpunar, K.K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron*, *76*, 677- 694.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, *117*, 14-21.

Schacter, D.L., Devitt, A.L., & Addis, D.R. (2018). Episodic future thinking and cognitive aging. In *Oxford Research Encyclopedia of Psychology*. Oxford University Press.

Schacter, D. L., & Madore, K. P. (2016). Remembering the past and imagining the future: Identifying and enhancing the contribution of episodic memory. Memory Studies, *9*, 245–255.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34.

Shah, C., Erhard, K., Ortheil, H. J., Kaza, E., Kessler, C., & Lotze, M. (2011). Neural correlates of creative writing: an fMRI study. *Human brain mapping*, *34*, 1088-1101.

Sheldon, S., McAndrews, M.P., & Moscovitch, M. (2011). Episodic memory processes mediated by the medial temporal lobes contribute to open-ended problem solving. *Neuropsychologia*, *49*, 2439-2447.

Sheldon, S., Williams, K., Harrington, S., & Otto, A. R. (2020). Emotional cue effects on accessing and elaborating upon autobiographical memories. *Cognition*, *198*, 104217.

Silvia, P. & Benedek, M. (2019, August 22). Creativity & Arts Tasks and Scales: Free for Public Use. Retrieved from osf.io/4s9p6.

Silvia, P.J., Winterstein, B.P., Willse, J.T., Barona, C.M., Cram, J.T., Hess, K.I., Martinez, J.L., & Richard, C.A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2* ,68-85.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Söderlund, H., Moscovitch, M., Kumar, N., Daskalakis, Z., Flint, A., Herrmann, N., & Levine, B. (2014). Autobiographical episodic memory in major depressive disorder. *Journal of Abnormal Psychology*, *123*, 51–60.

Spreng, R. N. (2013). Examining the role of memory in social cognition. *Frontiers in Psychology*, *4*, 437.

Spreng, R. N., Gerlach, K. D., Turner, G. R., & Schacter, D. L. (2015). Autobiographical planning and the brain: activation and its modulation by qualitative features. *Journal of Cognitive Neuroscience*, *27*(11), 2147-2157.

Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of Cognitive Neuroscience*, *22*(6), 1112–1123.

Spreng, R. N., & Mar, R. A. (2012). I remember you: A role for memory in social cognition and the functional neuroanatomy of their interaction. *Brain Research*, *1428*, 43–50.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *Neuroimage*, *53*(1), 303-317.

Squire, L. R., van der Horst, A. S., McDuff, S. G., Frascino, J. C., Hopkins, R. O., & Mauldin, K. N. (2010). Role of the hippocampus in remembering the past and imagining the future. *Proceedings of the National Academy of Sciences*, *107*(44), 19044-19048.

Strikwerda-Brown, C., Mothakunnel, A., Hodges, J. R., Piguet, O., & Irish, M. (2019). External details revisited – A new taxonomy for coding 'non-episodic' content during autobiographical memory retrieval. *Journal of Neuropsychology*, *13*, 371–397.

Strikwerda-Brown, C., Williams, K., Lévesque, M., Brambati, S., & Sheldon, S. (2021). What are your thoughts? Exploring age-related changes in episodic and semantic autobiographical content on an open-ended retrieval task. Memory, 29(10), 1375–1383.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans?. *Behavioral and brain sciences*, *30*(3), 299-313.

Szpunar, K.K., Watson, J.M., & McDermott, K.B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, *104*, 642-647.

Szpunar, K. K., Spreng, R. N., & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences*, *111*(52), 18414–18421. https://doi.org/10.1073/pnas.1417144111

Takano, K., Gutenbrunner, C., Martens, K., Salmon, K., & Raes, F. (2018). Computerized scoring algorithms for the Autobiographical Memory Test. *Psychological Assessment*, *30*(2), 259–273.

Takano, K., Hallford, D. J., Vanderveren, E., Austin, D. W., & Raes, F. (2019). The computerized scoring algorithm for the autobiographical memory test: Updates and extensions for analyzing memories of English-speaking adults. *Memory*, *27*, 306–313.

Takano, K., Ueno, M., Moriya, J., Mori, M., Nishiguchi, Y., & Raes, F. (2017). Unraveling the linguistic nature of specific autobiographical memories using a computerized classification algorithm. *Behavior Research Methods*, *49*, 835–852.

Tamir, D.I., Bricker, A.B., Dodell-Feder, D., & Mitchell, J.P. (2015). Reading fiction and reading minds: The role of simulation in the default network. *Social cognitive and affective neuroscience*, *11*, 215-224.

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences*, *107*(24), 10827–10832.

Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*, 151–162.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. In *Treebanks* (pp. 5-22). Springer, Dordrecht.

Thakral, P.P., Madore, K.P., Kalinowski, S.E., & Schacter, D.L. (2020). Modulation of hippocampal brain networks produces changes in episodic simulation and divergent thinking. *Proceedings of the National Academy of Sciences USA, 117,* 12729-12740.

Thakral, P. P., Madore, K. P., & Schacter, D. L. (2017). A role for the left angular gyrus in episodic simulation and memory. *Journal of Neuroscience*, *37*(34), 8142–8149.

Toutanova, K., & Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference EMNLP/VLC, 63-71, 2000*.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1* (pp. 173-180). Association for Computational Linguistics.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1-12.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1-25.

van Genugten, R. D., Beaty, R. E., Madore, K. P., & Schacter, D. L. (2021). Does episodic retrieval contribute to creative writing? an exploratory study. *Creativity Research Journal*, 1–14.

van Genugten, R., & Schacter, D. L. (2022). Automated Scoring of the Autobiographical Interview with Natural Language Processing. PsyArXiv.

Van Tilburg, Sedikides, & Wildschut (2015). The mnemonic muse: Nostalgia fosters creativity through openness to experience. *Journal of Experimental Social Psychology*, *59*, 1-7.

Vollberg, M. C., Gaesser, B., & Cikara, M. (2021). Activating episodic simulation increases affective empathy. *Cognition*, *209*, 104558.

Wardell, V., Esposito, C. L., Madan, C. R., & Palombo, D. J. (2021). Semi-automated

transcription and scoring of autobiographical memory narratives. *Behavior Research Methods*, *53*, 507–517.

Wardell, V., Madan, C. R., Jameson, T. J., Cocquyt, C. M., Checknita, K., Liu, H., & Palombo, D. J. (2021). How emotion influences the details recalled in autobiographical memory. *Applied Cognitive Psychology*, *35*, 1454–1465.

Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: dissociations between mirroring and self-projection. *Current Directions in Psychological Science*, *20*(3), 197–200.

Wickner, C., Englert, C., Addis, D.R. (2015). Developing a tool for autobiographical interview scoring. Kiwicam Conference, Wellington, New Zealand. https://github.com/scientific-tool-set/scitos

Williams, J. M., & Broadbent, K. (1986). Autobiographical memory in suicide attempters. *Journal of Abnormal Psychology*, *95*, 144–149.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *ArXiv*. http://arxiv.org/abs/1909.00161