Cued by music: Eliciting autobiographical memories across the lifespan and assessing their emotionality

by

Khalil Husein

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Arts

in

Psychology

Waterloo, Ontario, Canada, 2025

© Khalil Husein 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Music holds a unique capacity to involuntarily and rapidly evoke autobiographical memories (AMs), yet the source of its cueing ability is unspecified. In experiment 1, I examined what components of music are most beneficial in triggering AMs. Younger adults were presented with short clips of popular songs from their youth as memory cues, that were either unmodified, instrumental versions, lyrics-only, or a visual presentation of song and artist name. If participants experienced a memory, they provided a description of it. Unmodified song clips were most effective at evoking memories, and recall was enhanced in those high in trait music-related reward sensitivity. I also found evidence of temporal alignment between the song's year of popularity, and timeframe of the evoked AMs. Given the powerful cueing effect of music on AMs, in experiment 2 I examined whether the findings would generalize to an older adult population, and whether music could serve to reduce age-related episodic memory deficits. Strikingly, song cues evoked more AMs in older than younger adults, offering a means of enhancing access to personal memories. I observed a temporal alignment between the year of popularity of the song cue and the time frame of the evoked memory in both age groups. I propose that songs are particularly effective memory cues because they help set temporal context, constraining the search for AMs to a specific point in time, facilitating access. These findings are particularly relevant for older adults who may have difficulty spontaneously recreating context. In chapter 3, I investigated whether computational approaches using natural language processing can be employed to examine emotional sentiment within autobiographical narratives. Current methods of analysis require extensive manual human coding, limiting the sample size that can feasibly be examined. I compared the congruence in classification of two

popular lexicon-based sentiment analysis tools, VADER and TextBlob, with self-reported valence of 3,309 AM narratives from two datasets. Confusion matrices showed better congruence using VADER than TextBlob. Accuracy improved significantly when classification required binning valence into three rather than five categories, regardless of dataset size. Results suggest sentiment analysis is a promising avenue to determine broad classifications of valence within AM narratives. Overall, this thesis shows that songs are powerful cues for eliciting personal memories and can reduce age-related memory decline by setting the temporal context, facilitating access to content. Moreover, this thesis shows that VADER is a sentiment analysis approach that can be reasonably used to determine emotional valence, offering a means of characterizing autobiographical memory narratives.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Dr. Myra Fernandes. Her unwavering support and steady guidance formed the foundation for me to grow and progress as a researcher. I am also thankful for her constant encouragement that was crucial in overcoming the inevitable challenges over the past two years. I cannot imagine having a better graduate school experience elsewhere, and I heartily look forward to continuing my academic journey with her.

I would also like to provide my immeasurable thanks to my family. Thanks to my dad for being one of the few outside PAS to understand and provide support for the tumultuous experience that is academia. Thanks to my mom for never failing to check-in on me and for being a light that keeps me going through the most difficult of times. Thanks to my brother Solana for always sharing your love (and reels) even while living across the country. Despite you being the younger brother, I consider you a role model.

I must also shout-out my labmates (past and present): Patrick, Pelin, Sophia, Yadu and especially my officemate Brian. Thank you for your inspiring ideas and feedback during lab meetings and equally inspiring and warm-hearted banter outside of meetings. You all helped make Waterloo a new home for me (and kept my poster presentations picture perfect). Thank you as well to the close friends I made in the program outside of the lab: Jisoo, Justine and Miranda. You all helped me evolve from ‘going through it’ to ‘got through it’ and were instrumental in creating some of my fondest memories. Finally, thank you to Rico and Ursula for providing the love that only pets can provide.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements v

List of Figures ix

List of Tables x

Chapter 1 General Introduction 1

1.1 Cueing autobiographical memories 1

1.2 Episodic memory declines with age 2

1.3 What makes music an effective cue? 5

1.4 Computational methods in memory analysis 7

1.5 Current study 8

Chapter 2 11

2.1 Experiment 1 11

2.1.1 Methods 12

2.1.2 Materials 13

2.1.3 Procedure 14

2.1.4 Results 16

2.1.5 Discussion 22

2.2 Experiment 2 23

2.2.1 Methods 24

2.2.2 Materials 25

2.2.3 Procedure 27
2.2.4 Results 28
2.2.5 Discussion 38
Chapter 3 40
3.1 Growth of computational methods 40
3.2 Content analysis in psychology 40
3.3 Valence is relevant to content 42
3.4 Use of Sentiment analysis in psychology 43
3.5 Current study 43
3.6 Experiment 3 44
3.6.1 Data preparation 44
3.6.2 Frequency analysis 46
3.6.3 Sentiment analysis - VADER and TextBlob 46
3.6.4 Results 47
3.6.5 Discussion 53
3.7 Experiment 4 54
3.7.1 Participants 54
3.7.2 Materials & Methods 54
3.7.3 Results 56
3.7.4 Discussion 61
3.8 Experiment 5 61
3.8.1 Materials and methods 62

3.8.2 Results 64
3.9 Discussion 67
3.9.1 Congruence of classification 68
3.9.2 Factors influencing sentiment analysis tool classification accuracy 69
3.9.3 Summary of Chapter 71
Chapter 4 General Discussion 72
4.1 The role of cue type 73
4.2 Music enhances recall in older adults 75
4.3 Music sets temporal context to guide retrieval 75
4.4 Trait sensitivity to reward from music positively predicts retrieval rate 77
4.5 VADER can effectively classify sentiment in memory narratives 78
4.6 Limitations and future directions 78
4.7 Conclusions 81
References 82
Appendix A 101
Appendix B 103

List of Figures

Figure 1 *Retrieval Frequency by Cue Type in Experiment 1* 19

Figure 2 *Age of Memory by Year of Cue Popularity for each Cue Type in Experiment 1* 22

Figure 3 *Successful Retrieval Rate by Cue Type and Age Group in Experiment 2* 31

Figure 4 *Age of Memory by Year of Song Popularity for each Cue Type in Young Adults (panel A) and Older Adults (panel B) in Experiment 2* 34

Figure 5 *Word cloud of top 200 most frequently used words in music-evoked autobiographical memories* 49

Figure 6 *Frequency of memories by valence, calculated using self-reports, VADER and TextBlob* 50

Figure 7 *Confusion matrices of valence classifications of VADER (panel A) and TextBlob (panel B) relative to self-reported memory valence* 52

Figure 8 *Word cloud of top 200 most frequently used words in recurrent involuntary autobiographical memories.* 57

Figure 9 *Frequency of memories by valence, calculated using VADER, TextBlob and self-reports.* 58

Figure 10 *Confusion matrices of valence classifications of VADER (panel A) and TextBlob (panel B) relative to self-reported memory valence.* 60

Figure 11 *Average MCC between sentiment categories and self-reported valence at five different subset sizes of the rIAMs dataset (Experiment 2)* 65

Figure 12 *Heat maps of correlations between self-reported valence and the sentiment categories at different category thresholds* 66

List of Tables

Table 1 *Types of song clips presented to participants* 9

Table 2 *Descriptive Characteristics of Evoked Autobiographical Memories in Experiment 1* .. 17

Table 3 *Frequency of MEAM by Cue Type and Song Familiarity in Experiment 1* 18

Table 4 *Descriptives of Memory Characteristics for Younger and Older Adults in Experiment 2* 29

Table 5 *Generalized Linear Mixed Effects Model of Familiarity, Cue Type, eBMRQ score and Age Group on Successful Retrieval Rate in Experiment 2* 32

Table 6 *Age of Memory by Year of Cue Popularity for each Cue Type in Young Adults, with Song Cues of Low, Average, and High Familiarity* 37

Chapter 1
General Introduction

1.1 Cueing autobiographical memories

Recollection of places and events from our past are called autobiographical memories (Tulving, 1983). These can be evoked by a variety of sensory cues (Berntsen, 2009; Willander et al., 2015). For example, a pivotal scene in the movie 'Ratatouille' depicts one of the characters being transported to a tender event from their childhood after tasting a particular dish. Music, as an auditory cue, has emerged as a particularly potent one (Janata et al., 2007; Schulkind et al., 1999), frequently eliciting memories in both experimental and everyday settings (Jakubowski & Ghosh, 2019; Kaiser & Berntsen, 2022). For instance, the song 'As it Was' by Harry Styles may bring back a memory of having a picnic at the park with some friends. When presented with a series of popular music clips, participants in recent studies report experiencing a memory from a music cue approximately 30-50% of the time (Janata et al., 2007; Tanberg et al., 2025). Previous work also found that individuals experience music-evoked autobiographical memories (MEAMs) on a daily basis (Jakubowski & Ghosh, 2019; Jakubowski et al., 2023).

Music also appears to produce a strong phenomenological experience as part of evoked memories, even providing an enhanced experience relative to other memory cues. MEAMs are often rated as feeling more positive in valence compared to memories evoked by environmental sounds and words (Jakubowski & Eerola, 2022), as well as television shows (Jakubowski et al., 2021). MEAMs are also considered more personally important compared to memories evoked by concrete words (Zator & Katz, 2016) and food (Jakubowski et al., 2023). They are on par with memories evoked by abstract words and environmental sounds (Jakubowski & Eerola, 2022).

MEAMs are described as feeling similarly vivid to memories evoked by words (Zator & Katz, 2016), environmental sounds (Jakubowski & Eerola, 2022) and food (Jakubowski et al., 2023), and are more vivid than memories evoked by famous faces (Belfi et al., 2015). Music also seems to ease facilitation of memory access, evoking memories with little retrieval effort (Kaiser & Berntsen, 2022) and more spontaneously than food (Jakubowski et al., 2023) and famous face cues (Belfi et al., 2015).

On the other hand, MEAMs are less likely to occur than memories evoked by famous faces (Belfi et al., 2015), environmental sounds or words (Jakubowski & Eerola, 2022). Furthermore, Cady et al. (2007) found that memory quality did not differ between memories evoked by music and memories elicited by song lyrics (presented visually) or pictures of album covers. Amongst the backdrop of these mixed results, the mechanism of music's cueing ability is currently unclear. Identifying the properties of music that drive its beneficial cueing properties would help pinpoint music's strengths and limitations as a memory cue, providing perspective on these mixed cueing results and in determining music's potential in facilitating recall.

1.2 Episodic memory declines with age

It is currently unclear whether music retains its potential as a memory cue across the lifespan, particularly in cognitively healthy older adults. Older adults commonly report increased difficulty in their ability to learn new information and remember past events (Craik & Rose, 2012). Past studies have consistently identified age-related deficits in autobiographical memory (Light, 1991; Salthouse, 1991). However, these deficits can be mitigated by the use of memory cues (Perlmutter, 1979). Music may be a beneficial tool for this population, and could lessen

memory deficits by facilitating retrieval. Past work suggests age-related memory deficits are eliminated when additional environmental support is provided (Craik, 1983). Craik's 'environmental support' hypothesis proposes that age-related memory deficits can be reduced when older adults are provided with external supports (prompts, cues, or context) or structured environments that help compensate for declines in self-initiated processing. Environmental support can consist of giving hints or reminders, as well as being exposed to sensory cues that reinstate the context of a recalled event. For example, auditory and visual aspects of an environment (like listening to a certain song or seeing a particular building) that relate to the setting of a past event can trigger retrieval of that event, bypassing the need for self-initiating this process.

There is, however, reason to believe music's cueing ability might be limited in this population. Some studies suggest that the effectiveness of music as a memory cue depends on the integrity of the medial prefrontal cortex (Belfi et al., 2018; Ford et al., 2016; Janata, 2009). In older adults, unfortunately, this is precisely the brain region known to deteriorate, with frontal areas experiencing some of the largest volumetric declines (Craik & Rose, 2012; Giorgio et al., 2010). Not all studies suggest a frontal basis for MEAMS: Lesiuk et al. (2025) found that they were primarily associated with activation of the parahippocampal cortex in older adults. Because MEAM-associated activation in this region did not differ for those with versus without structural changes in grey matter volume, they suggest music's cueing abilities may be preserved in older adults. These differing findings suggest there is a need to assess music's function as a memory tool (that can either enhance, lower or show no benefit) across the lifespan.

Previous studies have demonstrated that music can effectively evoke memories in older adults (Platz et al., 2015; Salakka et al., 2021), and that songs dating from participants' reminiscence bump (period of adolescence in which memories get privileged access; Rubin et al., 1986) are particularly effective cues (Martínez-Saez et al., 2024; Salakka et al., 2024). These music-evoked memories are also rated as feeling positive (López-Cano et al., 2020). However, to our knowledge, only work by Jakubowski & Ghosh (2019) and Jakubowski et al. (2023) have directly compared music's cueing ability in cognitively healthy older and younger adults. Music was found to cue somewhat fewer memories in older than younger adults (Jakubowski et al., 2023), and older adults' memories were less specific (Jakubowski et al., 2023). Older adults did however report their memories as feeling more positive (Jakubowski & Ghosh, 2019), consistent with the American positivity effect of memory in aging (Mather & Carstensen, 2005). However, both of these previous studies utilised a diary approach in which, over a few days, participants recorded memories evoked from music they experienced in their everyday life. In this approach, a time lag between when the memory was experienced and when it was recorded in the diary (average delay = 22.17 minutes; Jakubowski et al., 2023) does not allow for capture of the immediate effects of music. Most importantly, there is no means within this method to control for the quantity of music exposure and kind of music listened to (e.g., whether songs are from participants' reminiscence bump or not). In our study, I assess music's effectiveness as a cue across age groups using a fixed number of songs for each participant, chosen from the era corresponding to when they would experience the well-known 'reminiscence bump'. The reminiscence bump is a tendency for individuals to access more memories from when they were teenagers and young adults (Rubin et al., 1986). Here, I conducted the first comparison of

music's cueing ability between younger and older adults, controlling for the quantity and kind of cue presented.

1.3 What makes music an effective cue?

The root reason for music's ability to cue memories is currently ambiguous. Multiple ideas have been proposed regarding its mechanism of action, a large proportion of which are derived from its ability to induce emotions (Juslin & Västfjäll, 2008). The subjective experience of emotion can be sub-divided into two dimensions, valence and arousal (Russell, 1980), both of which provide distinct contributions to memory recall. One idea stems from mood-congruence theory (Bower et al., 1981), suggesting that music can enhance recall for memories congruent in valence to one's mood. (Tesoriero & Rickard, 2012).

Music can be broken down into distinct components such as melody, harmony, rhythm and lyrics (Krumhansl & Cuddy, 2010). The component of music that drives its capacity to induce emotions during autobiographical memory retrieval is currently unspecified. The prosaic elements of music (song lyrics) can induce emotion and facilitate mood-congruence (Ali & Peynircioğlu, 2006; Peck & Grealey, 2020; Stratton & Zalanowski, 1994). However, this ability seems to be modest relative to other music components such as the melody (Ali & Peynircioğlu, 2006; Sousou, 1997). Alternatively, the instrument elements of music have displayed the ability to engender emotions, which then mediate music's ability to evoke memories (Salakka et al., 2021; Salakka et al., 2024). Identifying the unique capacity of the prosaic and instrument elements to evoke memories, separate from each other, could help elucidate which components contribute to music's cueing ability, and whether cue emotion contributes to facilitating recall.

An alternative idea, proposed by Tanberg et al. (2025), for music's cueing ability suggests that music can function as a temporal landmark (Shum, 1998), guiding memory retrieval to a particular point in time. Music can become associated with particular time periods (Christenson et al., 2019; Putter et al., 2022) and with significant lifetime events (Jakubowski & Ghosh, 2019). Other memory cues (e.g., words representing a lifetime period like 'primary school,' 'sixth grade') that contain temporal information have been found to evoke memories more quickly than cues without (Conway & Bekerian, 1987). Tanberg et al. (2025) hence suggested that music, through its temporally-unique characteristics such as genre and stylistic properties (Mauch et al., 2015), can direct memory retrieval to a particular time period. They found that memories evoked by popular music tended to originate from the same time in which that music was most popular. They attributed this temporal alignment between the year of memory and year of cue popularity to the music creating a temporal context which facilitates recall. However, this temporal alignment has currently only been observed in younger adults with music cues popular in the most recent five years. Here, I investigated the replicability of the temporal alignment effect, and whether it also translates to the memories of older adults.

Another characteristic of music that could contribute to its cueing effectiveness is its ability to elicit feelings of reward. Individuals have been found to vary in their propensity to experiencing feelings of reward from music (Mas-Herrero et al., 2013). The eBMRQ is a self-report questionnaire that can capture individual differences in music-related reward sensitivity (Cardona et al., 2022). Feelings of reward from music have been associated with increased arousal (Blood & Zatorre, 2001), which can facilitate memory consolidation (Cahill & McGaugh, 1998). Musical reward is also associated with activity within the mesolimbic

dopaminergic system (Menon & Levitin, 2005) which has been linked to enhanced memory encoding and retrieval (Wittmann et al., 2005; Wolosin et al., 2012).

1.4 Computational methods in memory analysis

Computational tools for text analysis, commonly referred to as Natural Language Processing (NLP) methods, allow for near real-time analysis of mass corpora of text, numbering in the hundreds of millions (Eichstaedt et al., 2021; Golder & Macy, 2011). Despite their widespread embrace in business research, NLP methods have only recently started to filter into psychological realms of research. Past work has shown that narrative texts can provide insight into the mental and affective state of a person, regardless of age (Yeung & Fernandes, 2020; 2021). Here I explored the capacity and competence of NLP methods when applied to understanding autobiographical memories, defined as recollections of events that occurred in a specific time and place (Tulving, 1983). Automation of text analysis could be particularly useful in expanding the scope of data and research questions that can be considered within psychological studies (Belfi et al., 2015; Newby & Moulds, 2012; Schlagman et al., 2006). Online research platforms, such as Prolific and Amazon MTurk, have also made it easier to collect larger datasets of text data, and NLP methods could facilitate analysis of these large datasets. For example, recent work used NLP content analysis tools to identify patterns in the emotional valence of 3345 autobiographical memory transcripts that differentiated depression-prone from boredom-prone individuals, addressing long-standing questions about whether such conditions are unique or interrelated (Yeung et al., 2024). Application to autobiographical

memories elicited by music could help characterize them, thus enabling the use of NLP as a tool in future research.

1.5 Current study

In this thesis I aimed to pinpoint the components of music responsible for its effectiveness in cueing autobiographical memories. In experiment 1, I examined what components of music are most beneficial in triggering AMs. Young adults were presented with short clips of popular songs from their youth as memory cues, which were either unmodified, instrumental versions, lyrics-only, or a visual presentation of song and artist name (see Table 1). If participants experienced a memory, they provided a description of it. In experiment 2, I compared music's cueing ability in another set of younger adults to older adults. With this experiment, I aimed to ascertain the generalisability of the findings from experiment 1, by determining whether the driving components of music's cueing ability varied across age populations. Based on the well-documented age-related episodic memory deficits across aging and the music-specific findings of Jakubowski et al. (2023), I predicted that older adults would demonstrate an observable but small deficit in autobiographical memory recall relative to young adults. However, I did not predict any differences between populations in the relative effectiveness of different song components. Past work suggests that music cues temporally and emotionally aligned with autobiographical memories (Tanberg et al., 2025). To assess the generality of this proposed mechanism of action, I examined its replicability in an older adult population.

Table 1

Types of song clips presented to participants

		Prose	
		with	*without*
Instruments	*with*	Unmodified	Instruments-only
	without	Lyrics-only	Song & artist name

Finally, in Chapter 3 (Husein et al., in preparation), I used NLP methods to assess the sentiment contained within autobiographical narratives as this emerging approach can offer a means to enable analysis for large datasets. Such newly accessible data analyses tools open the door to expanding the characterisation of memory narratives. For example, one could examine the positivity effect in aging, or differences in emotional expressivity described in one's first or second language. These tools could also help provide insight into the emotional experience of music-evoked memories, and the role of different components of music in invoking this experience. In settings where self-reported memory valence is difficult to acquire, such as in naturalistic environments (e.g., music therapy) or from populations who may have biases in interpretation of their emotion (alexithymia), NLP methods could be an alternative instrument to investigate the emotional experience of memories elicited by music. As a first step, in my thesis, I investigated whether there is converging validity in classification of valence as determined by self-assessment and by two different NLP methods: VADER and TextBlob. Sentiment scores generated by each method were compared to participant self-reported valence scores, the current prevailing gold-standard used in psychology for identifying text valence (Barrett, 2004; Bradley & Lang, 1994; Watson et al., 1988). Here, the viability of these tools as an alternative to human coding was considered, as human coding is subjective and requires intensive amounts of time

and labour. In order to modernize the study of narrative transcripts, and better characterize them, analysis of larger samples than in past work is needed. Here I examine whether NLP methods are viable.

Chapter 2

2.1 Experiment 1

I manipulated cue type, specifically whether instruments and prose (lyrics) were present or absent in a song cue. I then evaluated whether these components were differentially effective at cueing memories, and whether there were differences in ease of access and properties of the memories (emotional valence, arousal, sense of re-living). In this way, our study was designed to help clarify the mechanism of action for how music cues memory. I predicted that song cues containing neither prose nor instruments (a visual presentation of the song and artist name) would be least effective, due to a restricted ability to induce retrieval-guiding emotions (Tesoriero & Rickard, 2012). I expected that cues containing prose but no instruments (lyric readings) would be somewhat more effective, due to the presence of lyrics to provoke emotions (Stratton & Zalanowski, 1994). However, I anticipated greater effectiveness from cues containing instruments but no prose (akin to a karaoke background) because of their stronger link to music-induced emotion (Ali & Peynircioğlu, 2006) and demonstrated predictive ability of evoking memories (Salakka et al., 2021; Salakka et al., 2024). Finally, I predicted unmodified song cues to be most potent as a memory cue, benefitting from the emotional features of both lyrics and instruments and comprising of more stylistic properties that could be traceable to a particular time period (Mauch et al., 2015).

I additionally explored whether there is inter-individual variation in songs' ability to cue memories. Particularly, I investigated whether individual differences in propensity for experiencing feelings of reward from music could influence a song's cueing effectiveness.

Feelings of reward have been associated with enhanced arousal (Blood & Zatorre, 2001) and dopamine release (Menon & Levitin, 2005) that can facilitate memory retrieval (Wittmann et al., 2005; Wolosin et al., 2012). As such, I predicted that individuals with a higher sensitivity to music-related reward, indexed by a higher score on the eBMRQ scale, would be more likely to experience memories from music, and that they would report a more emotionally arousing experience.

2.1.1 Methods

2.1.1.1 Participants

Eighty undergraduate students were recruited from the University of Waterloo, who participated in the study in return for course credit. Participation was restricted to individuals with normal or corrected-to-normal vision and hearing, as well as to individuals who had lived in Canada since 2018. Data from two participants who were not fluent in English, as indicated by a score of less than 30% on the Mill-Hill vocabulary test (Raven, 1983), as well as data from one participant who was above the study age range, were excluded. The final analyzed sample consisted of 77 participants. Of these participants, 70.10% were female, 28.60% were male, and 1.30% preferred not to disclose their sex. Mean age was 19.13 (*SD* = 1.23, range = 18-22). Mean years of education was 14.47 years (*SD* = 2.36, range = 4-19 years). Data were collected in 2024 between March and April.

The sample size was based on a calculation using G*Power 3.1.9.7 (Faul et al., 2009). I used a small effect size, f =0.1, to power for a within-factors interaction with a desired power of ~.80 (α =.05, two-tailed).

2.1.2 Materials

2.1.2.1 Song cues

Six songs from each of the years 2019, 2020, 2021 and 2022 were chosen from the Billboard Top 10 most popular songs of each year (Billboard Top 10 Songs; see Wikipedia for an open-access version of the Billboard top songs of each year), and an equal number of songs were chosen with male and female singers (see Appendix A for list). When possible, only one song was chosen from each singer if they appeared multiple times in the Top 10. Songs were purchased from the iTunes store for educational purposes and were shortened into 15-second clips starting at the beginning of the chorus (defined by Genius; Moghadam et al., 2009) using Audacity (Mazzoni & Dannenberg, 2023). Instruments-only clips were created by stripping out the lyrical features of unmodified song clips using vocali.se (*Vocali.Se*, 2025). Spoken clips of the lyrics were created by first transcribing the lyrics within each song's clip. These lyrics were then transformed into auditory stimuli using Speechify Studio text-to-speech software (Weitzman, 2017). Lyrics were spoken with a neutral tone/pacing. To make pacing neutral and naturalistic, I added 0.1-0.25 second pauses between sentences. Clips shorter than 15 seconds were looped using Audacity (Mazzoni & Dannenberg, 2023) to be 15 seconds in length, matching the other cue types, with a 1.5-2.5 second pause between loops. The gender of the (Speechify-generated) speaker was matched to the gender of the original song's vocalist.

2.1.2.2 Musical Sophistication index

Participants reported music-specific demographic information including their years of formal musical training and perceived musicianship (1 = *nonmusician*, 7 = *professional musician*) (question from the Ollen Musical Sophistication index; Ollen, 2006)

2.1.2.3 eBMRQ Scale

I investigated whether individual differences in sensitivity to music-related reward modulate the generalisability of songs as an effective memory cue. Participants completed the extended Barcelona Music Reward Questionnaire (eBMRQ; Cardona et al., 2022), a 24-item scale indexing their trait music-related reward sensitivity.

2.1.2.4 Mill Hill vocabulary test

Participants completed set A of the Mill-Hill vocabulary test (Raven, 1983) to assess their English fluency (M = 15.68, SD = 3.31), as proficiency could influence the quantity and quality of autobiographical transcripts produced. In accordance with standard protocol, two participants were excluded from data analysis since they received a score of less than 30% on the Mill-Hill vocabulary test (i.e., a score of 9 or lower).

2.1.3 Procedure

Participants completed the 60-minute study in-person on lab computers in groups of 1-3. After providing informed consent, participants, on each of 24 intermixed trials, listened to (or looked at for the song and artist name) a song cue for 15 seconds using over-ear OneOdio A70 Fusion Bluetooth headphones. The song cue either consisted of an unmodified song clip, an instruments-only clip, a lyrics-only spoken clip, or a visual presentation of the song and artist

name. As participants were presented with each cue type, they indicated whether they were reexperiencing a memory by pressing the <spacebar> on a QWERTY keyboard. If a participant pressed the <spacebar>, the time elapsed between when the song cue started playing and when the key press occurred was logged (i.e., response time), but the remainder of the cue was always played for the full duration uninterrupted. Participants would also be asked after the clip ended whether the song cue evoked a memory or not, in case participants failed to press the <spacebar> in-time. Participants then rated their familiarity with the cue on a 5-point scale (1 = *not familiar*, 5 = *I completely knew it*).

If a participant had experienced an autobiographical memory, they were prompted to write a short 3-5 sentence description of it in a text box, note the year from which the original event occurred, and then rate the properties of their evoked memory in terms of its emotional valence (1 = *very unpleasant*, 5 = *very pleasant*) and arousal (1 = *very calm*, 5 = *very exciting*) and their subjective feelings of reliving (1 = *no reliving*, 5 = *as if it were happening now*) on 5-point scales. Participants then continued to the next trial.

Participants who did not experience a memory were prompted to answer a series of similarly structured but thematically unrelated questions. These questions were included to maintain a consistent study length, regardless of whether participants reported a memory or not. In trials where participants did not experience a memory, they were presented with an emotionally neutral, concrete word. They were asked to define the word in a text box and then state how many letters are in the word. Following this step, participants were asked to indicate, through a series of yes/no responses, whether the word refers to something living or not, whether

the word refers to something bigger than an iPhone, and whether the word rhymes with 'pen.' After these questions were completed, participants continued to the next trial.

Participants completed all 24 trials in one block, with stimulus type randomized, such that participants experienced six stimuli from each condition. No participant heard the same song twice in a different form (i.e., they only heard one of the unmodified, instruments-only, lyrics-only or the song/artist name version of the same song). Participants completed one of four possible versions of the study, created for counterbalancing of cue type for each song. Song cue presentation and follow-up questions were administered using E-Prime 3.0 (Cernicky, 2017). After completion of all 24 trials, participants completed a survey on Qualtrics in which they were asked to answer demographic questions, complete the eBMRQ scale and the Mill Hill vocabulary test. All procedures were approved by the University of Waterloo's Office of Research Ethics (Protocol #45916).

2.1.4 Results

See Table 2 for descriptives of participants' memory characteristics. Participants on average reported experiencing a memory on 50% of the 24 trials. Memories were typically retrieved within 5-6 seconds after exposure to the cue. Generally, memories were self-rated as moderately arousing and having a positive valence. Additionally, participants typically reported a weak to moderate sense of re-living their memories. Participants reported moderate familiarity with the music cues. Statistical analyses were completed using SPSS (version 30.0.0.0).

Table 2

Descriptive Characteristics of Evoked Autobiographical Memories in Experiment 1

	M	*SD*
Number of memories	11.90 (50%)	4.58
Response time (s)	5.72	2.42
Arousal	3.22	0.52
Valence	3.74	0.40
Sense of re-living	2.68	0.68
Cue familiarity	3.40	0.75

2.1.4.1 Familiarity of music cue by cue type

Familiarity ratings were analyzed in a 2 (Prose: with or without) by 2 (Instruments: with or without) repeated measures ANOVA. Outliers were defined as data or ratings that were greater or lesser than 3 *SD*s from the mean. Using this definition, 1.30% of the cue familiarity data were identified as outliers and removed from analyses (see Table 3 for means).

I found a main effect of Instruments, $F(1,74) = 38.93$, $p < .001$, $\eta^2_p = .35$, and Prose, $F(1,74) = 23.99$, $p < .001$, $\eta^2_p = .25$, along with an Instruments X Prose interaction, $F(1,74) = 41.87$, $p < .001$, $\eta^2_p = .36$. In cues containing Prose, there was a main effect of Instruments such that cues with instruments (unmodified song cues) were rated as more familiar ($M = 4.10$, $SD = 0.65$) than cues without instruments (lyric-only cues), ($M = 3.19$, $SD = 1.05$, $F(1,74) = 84.66$, $p < .001$, $\eta^2_p = .53$). There was no effect of Instruments in cues without prose (instruments-only or song-artist name cues), $F(1,74) = 1.42$, $p = .237$, $\eta^2_p = .02$.

Table 3

Frequency of MEAM by Cue Type and Song Familiarity in Experiment 1

	Number of MEAMs		Familiarity	
	M	*SD*	*M*	*SD*
Song/artist name	2.70	1.72	3.14	1.04
Lyrics-only	2.48	1.71	3.17	1.07
Instruments-only	2.74	1.59	3.25	0.95
Unmodified	3.97	1.51	4.05	0.71

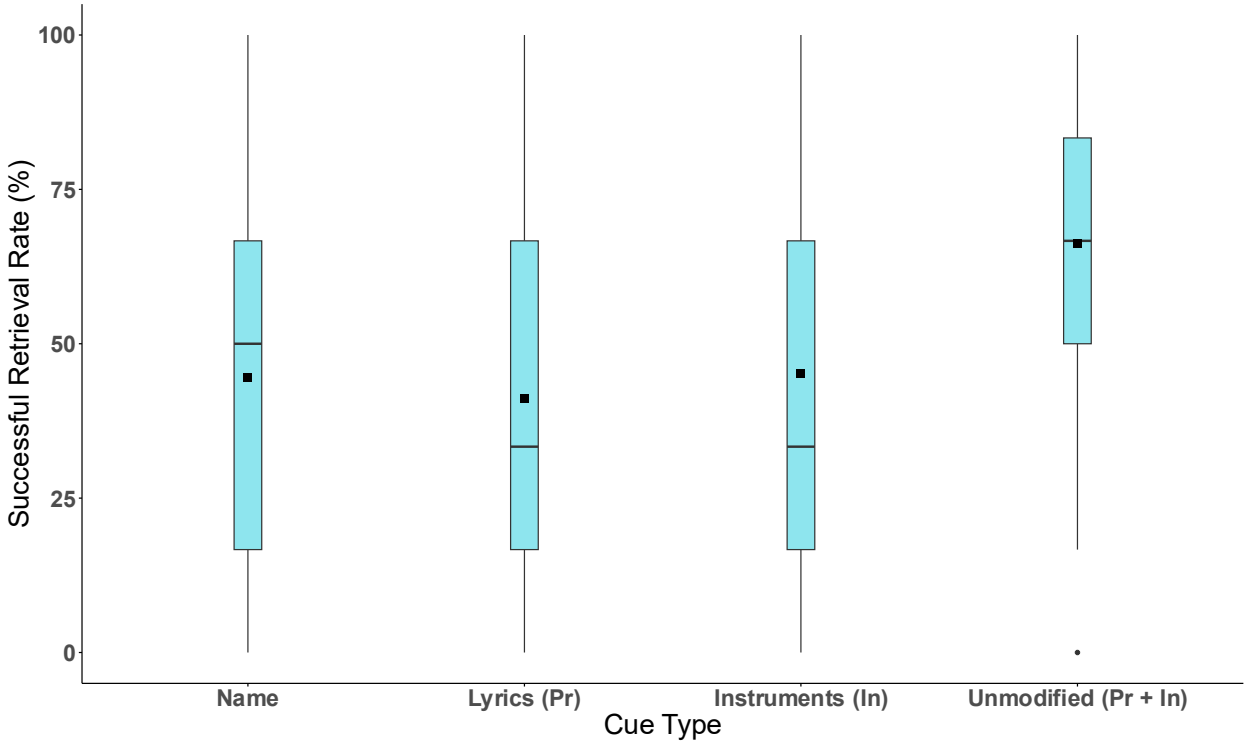
Note: MEAM = music-evoked autobiographical memory; Six clips of each cue type were presented to each participant; Familiarity was rated on a scale of 1 (least) to 5 (most familiar)

2.1.4.2 Retrieval frequency by cue type

Frequency of successful memory retrieval was analyzed in a 2 (Prose: with or without) by 2 (Instruments: with or without) repeated measures ANOVA. See Table 3 for means and Figure 1 for a visualization. I found a main effect of Instruments, $F(1,76) = 25.12$, $p < .001$, $\eta^2_p = .25$, and of Prose, $F(1,76) = 9.25$, $p = .003$, $\eta^2_p = .11$. Furthermore, there was an Instruments X Prose interaction, $F(1,76) = 26.85$, $p < .001$, $\eta^2_p = .26$. Probing this interaction, a simple effect of Instruments emerged in the presence of prose, such that cues with instruments (unmodified song cues) evoked more memories than cues without instruments (lyric-only cues), $F(1,76) = 46.74$, $p < .001$, $\eta^2_p = .38$. In contrast, there was no effect of Instruments in cues without prose (instruments-only or song-artist name cues), $F(1,76) < 1$. Interestingly, the presence and absence of prose and instruments seem to similarly influence cue familiarity and retrieval frequency.

Figure 1

Retrieval Frequency by Cue Type in Experiment 1



Note. Horizontal lines represent the median retrieval frequency. Square points represent the mean retrieval frequency. 'Pr' indicates a cue type that contains prose. 'In' indicates a cue type that contains instruments. Memory retrieval was most frequent when cued from unmodified songs, containing both prose and instruments.

2.1.4.3 Individual differences in memory experiences

The effects of music-specific demographic properties (i.e., years of musical training, perceived musicianship, sensitivity to reward from music (indexed by participants' score on the eBMRQ scale)) on retrieval rate and memory properties were analyzed in separate linear multiple regression analyses. These demographic properties somewhat predicted retrieval frequency, although this ultimately didn't reach significance, $R^2_{Change} = .10$, $F_{Change}\ (3, 73) = 2.65$, $p = .055$. Specifically, participants' trait sensitivity to reward from music positively predicted the number of memories they had, $B = 0.13$, $t(76) = 2.80$, $p = .007$. Furthermore, these

demographic factors predicted participants' memory arousal, R^2_{Change} = .18, F_{Change} (3, 73) = 5.23, p = .003, and their familiarity with the music cues, R^2_{Change} = .14, F_{Change} (3, 73) = 3.98, p = .011. Participants' musical background, indexed by their years of musical training and self-perceived musicianship, was not observed to impact memory experience. However, participants who had higher trait sensitivity to reward from music reported their memories as being more arousing, B = 0.02, t(76) = 3.20, p = .002, and reported higher cue familiarity, B = 0.03, t(76) = 3.43, p < .001. These music-specific individual differences did not predict participants' response time or valence, all F_{Change} (3, 73) < 1, nor their sense of reliving, R^2_{Change} = .05, F_{Change} (3, 73) = 2.24, p = .301.

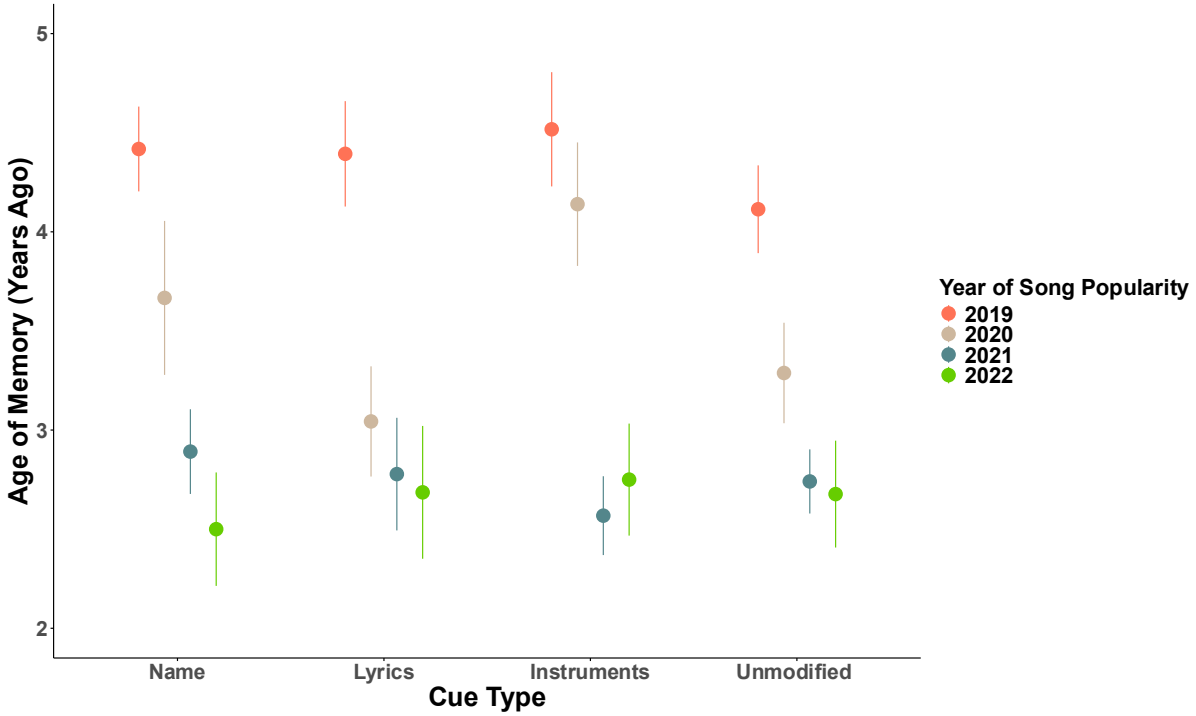
2.1.4.4 Temporal alignment between year of song popularity and age of memory

I investigated whether there was a relationship between the year a song cue was most popular and the year of a participant's reported memory. Results are shown in Figure 2. A temporal relationship was observed, such that AMs produced by recently popular songs were also from a more recent time period, and older songs produced memories from a more remote time period. I also calculated the deviation between cue and memory age (in years) by taking the difference between the year of a memory and the year the corresponding song cue was most popular. Deviations were short: less than a year ($M_{yeardeviation}$ = 0.33, SD = 1.33). I compared year deviation across cue types using a 2 (Prose: with or without), by 2 (Instruments: with or without) repeated measures ANOVA. Outliers were defined as data or ratings that were greater or lesser than 3 SDs from the mean. Using this definition, 7.94% of the temporal deviation data were identified as outliers and removed from analyses.

There was a main effect of Prose on temporal alignment such that cues with prose (both unmodified songs and lyrics-only) evoked memories that deviated more temporally from year of cue popularity ($M = 0.54$, $SD = 1.29$) than cues without instruments (both instruments-only and song & artist name), ($M = 0.13$, $SD = 1.37$), $F(1,57) = 5.46$, $p = .023$, $\eta^2_p = .09$. There was no main effect of Instruments, $F(1,57) < 1$, nor was there an interaction between Instruments and Prose, $F(1,57) < 1$.

Figure 2

Age of Memory by Year of Cue Popularity for each Cue Type in Experiment 1



Note. Points represent mean age of memory. Error bars represent ± 1 standard error. Data were collected in 2024 (i.e., a memory from 0 years ago would be from 2024).

2.1.5 Discussion

Song cues were most effective when presented intact, containing both prose and instruments. Results suggest that the effect of instruments on cued recall is beneficial only if prose is also present in the cue. That is, unmodified songs were the most effective cue type to conjure autobiographical memories. I additionally observed a temporal match between the year of popularity of a song cue and the resulting cued memory, replicating the alignment found by Tanberg et al. (2025). This alignment could be a consequence of song cues guiding retrieval by constraining memory search to a particular associated time period. In experiment 2 I applied linear mixed effects models to evaluate the strength of this effect, and included cue type in the

model to determine whether this effect was unique to unmodified song cues. In experiment 2 I also investigated whether the temporal alignment effect perseveres even when song cues are deemed less familiar by the participant. Given the powerful cueing effect of music on AMs, I also aimed to determine whether the findings would generalize to an older adult population and could serve to reduce the known age-related episodic memory deficit associated with cognitive aging.

2.2 Experiment 2

Experiment 1 consisted of participants between the ages of 18 and 22. The generality of these findings across age groups, particularly to older adults, remains unclear. A study by Salakka et al. (2021) identified key instrumental features that predicted the cueing ability of songs in older adults. They conducted a mediation analysis and found that the emotions evoked by instruments mediated the beneficial effects of songs on memory retrieval. Specifically, they suggested that instrumental features facilitated song's cueing ability by inducing a state of higher emotional intensity. This work suggests that the instrument component within a song may play an important role in cueing in this population. Alternatively, one can reason that the semantic information contained in prose may prove more advantageous as a cue because of the well-documented preservation of semantic and global knowledge in this population (Park et al., 2002). To empirically evaluate these competing ideas, as well as to assess the generalizability of experiment 1 findings across age groups, in experiment 2 I compared the capability of prose and instruments cues in eliciting AMs in younger and older adult populations. Different songs cues were presented to younger vs. older adults since, to equivalently support memory retrieval in

both populations, songs were chosen to be from each age group's respective reminiscence bump era.

Cue familiarity and trait sensitivity to feelings of reward from music, indexed by the eBMRQ scale, emerged in experiment 1 as possible factors influencing successful retrieval of AMs. The influence of participants' eBMRQ scores on retrieval suggests consideration of individual differences is necessary to determine the generalizability of song cueing and its mechanism of action. The retrieval benefit associated with a higher eBMRQ score, indicating high music-related reward, also hints that a song's ability to cue memories may be multi-faceted. Songs may not only set temporal context during retrieval, but additionally could induce a reward state that guides memory retrieval to events associated with a similar reward state. In experiment 2, I used a generalized linear mixed effects model to explore whether increased cue familiarity could account for the influence of Prose and Music on successful retrieval rate, and whether trait reward sensitivity to music provides additional explanatory power in accounting for a song's cueing benefit.

2.2.1 Methods

2.2.1.1 Participants

Eighty younger adults (range = 18-24) and eighty older adults (range = 64-78) were recruited using the Prolific online recruitment platform (*Prolific*, 2014), who participated in the study in return for £6.00 GBP. The sample size of experiment 2 was chosen to match that of experiment 1 for each age group. Participation was restricted to individuals from Canada or the USA, as well as to individuals with normal or corrected-to-normal vision and hearing, no

neurological or psychological problems, nor any severe head injury or stroke history. Data from three younger adults and two older adults who were not fluent in English, as indicated by a score of less than 30% on the Mill-Hill vocabulary test (Raven, 1983), were excluded. The final analyzed sample consisted of 77 younger adults and 78 older adults.

Of the younger adult participants, 62.34% were female, and 37.66% were male. Mean age was 20.92 (*SD* = 1.67). Mean years of education was 15.73 years (*SD* = 2.47, range = 5-23 years. Of the older adult participants, 57.69% were female, and 42.31% were male. Mean age was 70.19 (*SD* = 3.39). Mean years of education was 17.06 years (*SD* = 3.18, range = 12-25 years). Data were collected in 2024 between August and October.

2.2.2 Materials

2.2.2.1 Song cues

Younger adult song cues were the same as those used in experiment 1. For older adults, six songs from each of the years 1966, 1967, 1968 and 1969 were chosen from the Billboard Top 20 most popular songs of each year (Billboard Top 10 Songs; see Wikipedia for an open-access version of the Billboard top songs of each year). Songs were chosen to achieve a balance in number of male and female singers whenever possible, though in these years there were relatively few female singers with top songs (proportion of songs with female singers was only 25%). When possible, only one song was chosen from each singer if they appeared multiple times in the Top 20 (see Appendix A for list of songs). Songs were purchased from the iTunes store for educational purposes and were shortened into 15-second clips starting at the beginning of the chorus (defined by Genius; Moghadam et al., 2009) using Audacity (Mazzoni &

Dannenberg, 2023). The same procedure from experiment 1 was used to create the various cues from the selected songs.

2.2.2.2 Musical Sophistication index

Participants reported music-specific demographic information including their years of formal musical training and perceived musicianship (1 = *nonmusician*, 7 = *professional musician*) (question from the Ollen Musical Sophistication index; Ollen, 2006)

2.2.2.3 eBMRQ Scale

I investigated whether individual differences in sensitivity to music-related reward modulate the generalisability of songs as an effective memory cue. Participants completed the extended Barcelona Music Reward Questionnaire (eBMRQ; Cardona et al., 2022), a 24-item scale indexing their trait music-related reward sensitivity.

2.2.2.4 Mill Hill vocabulary test

Participants completed set A of the Mill-Hill vocabulary test (Raven, 1983) to assess their English fluency (young adults: $M = 22.57$, $SD = 5.88$; older adults: $M = 23.44$, $SD = 4.27$), as proficiency could influence the quantity and quality of autobiographical transcripts produced. In accordance with standard protocol, five participants (three young adults, two older adults) were excluded from data analysis since they received a score of less than 30% on the Mill-Hill vocabulary test (i.e., a score of 9 or lower).

2.2.2.5 Cognitive Failures Questionnaire 2.0

As a proxy measure of cognitive capacity, I asked participants to complete the Cognitive Failures Questionnaire 2.0 (CFQ 2.0; Goodhew & Edwards, 2024). The CFQ 2.0 is a 15-item scale which assesses individual differences in experience of cognitive failures. The average score on the CFQ 2.0 in young adults was 22.19 (*SD* = 10.86). The average score on the CFQ 2.0 in older adults was 13.76 (*SD* = 8.06).

2.2.3 Procedure

The procedure was the same as that in experiment 1. The only differences were that participants completed the 60-minute study online via Qualtrics. Participants indicated whether they were reexperiencing a memory by clicking on a button on-screen. If a participant pressed the button, the time elapsed between when the song cue started playing and when the button press occurred was logged (i.e., response time); as in experiment 1, the remainder of the cue was always played for the full duration uninterrupted. If a participant had experienced an autobiographical memory, they were prompted to auditorily record themselves (rather than typing their memory as in experiment 1) describing their memory in 30-45 seconds, and noted when the original event occurred from a series of eight possible year ranges.

After completion of all 24 trials, in addition to the scales and questionnaires from experiment 1, participants were asked to complete the CFQ 2.0. All procedures were approved by the University of Waterloo's Office of Research Ethics (Protocol #45916).

2.2.4 Results

Collapsed across younger and older adults, participants on average reported experiencing a memory on 46% of the 24 trials. Samples of memory narratives can be viewed in Appendix B. Memories were typically retrieved after 10-15 seconds, (M = 12.38, SD = 4.32). Generally, memories were rated as moderately arousing (M = 3.27 out of 5, SD = 0.71) and having a positive valence (M = 3.99 out of 5, SD = 0.50). Furthermore, participants typically reported a moderate sense of re-living their memories, (M = 2.88 out of 5, SD = 0.82). Participants reported moderate familiarity with the song cues, (M = 3.04 out of 5, SD = 0.81).

Memory characteristics, split by age group, are shown in Table 4. Older adults on average reported experiencing more memories than younger adults, $t(153) = 2.23$, $p = .027$, $d = 0.36$. However, younger adults reported retrieving their memories more quickly than older adults, $t(145) = 2.07$, $p = .041$, $d = 0.34$. Older adults also rated their memories as more arousing, $t(149) = 2.45$, $p = .015$, $d = 0.40$, and of a more positive valence than younger adults, $t(151) = 3.26$, $p < .001$, $d = 0.53$. Sense of re-living was not significantly higher in older than younger adults, $t(151) = 1.78$, $p = .077$, $d = 0.29$. Older adults reported significantly higher familiarity with their song cues than did younger adults, $t(153) = 2.74$, $p = .007$, $d = 0.44$.

Table 4

Descriptives of Memory Characteristics for Younger and Older Adults in Experiment 2

	Younger adults (n = 77)		Older adults (n = 78)	
	M	*SD*	*M*	*SD*
Number of memories	9.95 (41%)	5.01	11.92 (50%)	5.98
Response time (s)	11.65	4.61	13.13	4.03
Arousal	3.13	0.70	3.41	0.69
Valence	3.86	0.51	4.12	0.46
Sense of re-living	2.76	0.79	2.99	0.83
Cue familiarity	2.86	0.83	3.21	0.75

2.2.4.1 Retrieval frequency by cue type and age group

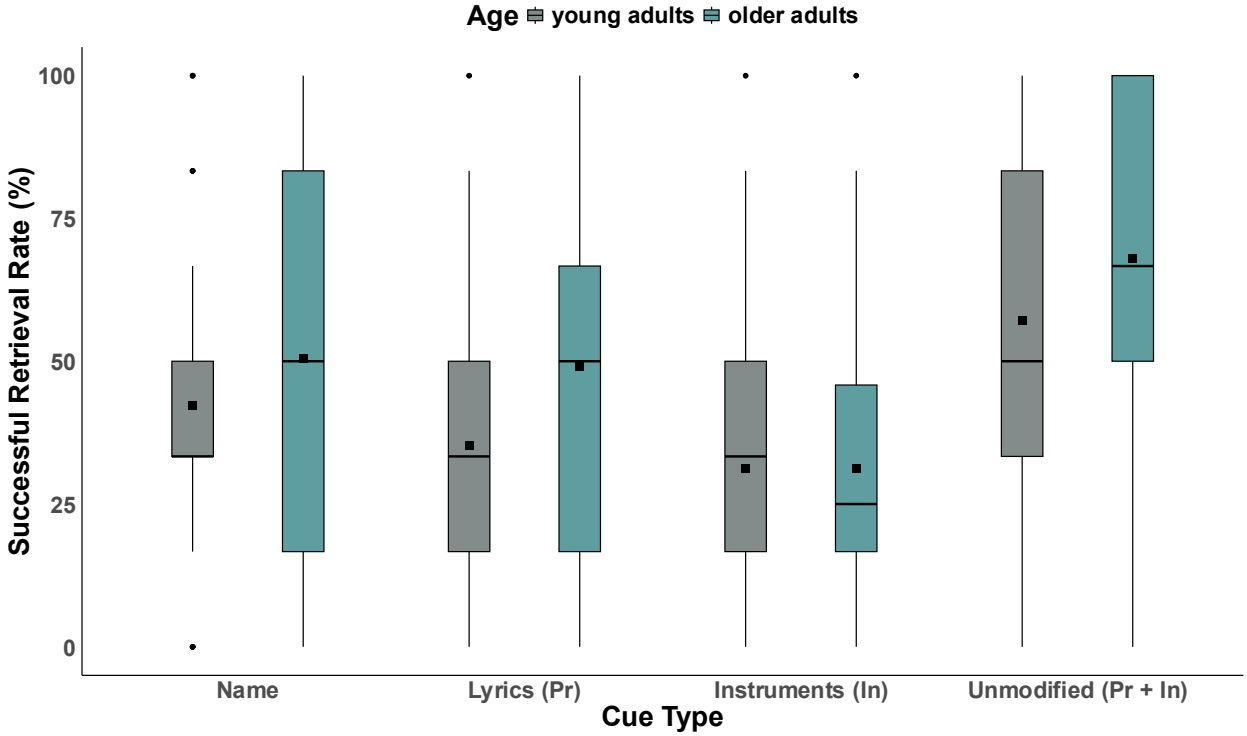
Frequency of successful memory retrieval was analyzed in a 2 (Prose: with and without) by 2 (Instruments: with or without) by 2 (Age: younger and older adults) repeated measures ANOVA (see Figure 3). I found a main effect of Prose, $F(1,153) = 74.27$, $p < .001$, $\eta^2_p = .33$, and of Age, $F(1,153) = 4.96$, $p = .027$, $\eta^2_p = .03$. These were qualified by a Prose X Age interaction, $F(1,153) = 6.74$, $p = .010$, $\eta^2_p = .04$ as well as a Prose X Instrument interaction, $F(1,227.44) = 118.04$, $p < .001$, $\eta^2_p = .44$. There was no significant three-way interaction between Instrument Prose and Age, $F(1,153) < 1$. Notably, the post-hoc power to detect this three-way interaction was quite low ($\beta = .12$).

I probed the interaction between Prose and Age by conducting independent samples t-tests comparing the effect of Age on successful retrieval rate using cues with and without prose. There was no effect of Age from cues without prose (instruments-only and song-artist name cues), $t(153) = -1.07$, $p = .284$, $d = -.17$. In contrast, there was an effect of Age when the cue contained prose (unmodified and lyric-only cues), such that older adults recalled more memories than younger adults, $t(153) = -2.94$, $p = .004$, $d = -.47$.

I probed the interaction between Prose and Instrument by conducting paired t-tests comparing the effect of Instrument on successful retrieval rate using cues with and without prose. There was a simple effect of Instrument in cues without prose, such that cues with instruments (instruments-only cues) were less effective at evoking memories, $t(154) = 6.64$, $p < .001$, $d_{av} = .55$. In contrast, in cues with prose, cues with instruments (unmodified cues) were more effective at evoking memories, $t(154) = -9.17$, $p < .001$, $d_{av} = -.68$.

Figure 3

Successful Retrieval Rate by Cue Type and Age Group in Experiment 2



Note. Horizontal lines represent the median retrieval frequency. Square points represent the mean retrieval frequency. 'Pr' indicates a cue type that contains prose. 'In' indicates a cue type that contains instruments.

2.2.4.2 Factors affecting retrieval frequency

To investigate the possible drivers of song's cueing benefit, I conducted a generalized linear mixed-effects model predicting successful memory retrieval. Cue type (song & artist name, lyrics-only, instruments-only, unmodified), cue familiarity (1 = *not familiar,* 5 = *I completely knew it*), age group (Young, Old) and eBMRQ score were included as fixed effects, and participant was included as a random effect (see Table 5 for results)

Cue familiarity was a strong predictor of memory retrieval, such that, holding other variables constant, each unit increase in cue familiarity (on a 5-unit scale) was associated with an increase of 7.25 in the odds of recalling a memory, $F(1,3713) = 841.37$, $p < .001$. Controlling for all other variables, including familiarity, cue type persisted as a predictor of retrieval rate, $F(3,3713) = 3.06$, $p = .027$. Specifically, unmodified song cues remained more effective than instrument-only cues, $b = 0.47$, $p = .003$. However, the effectiveness of unmodified songs was not significantly higher than song and artist name, $b = 0.19$, $p = .212$, or lyric-only cues, $b = 0.26$, $p = .086$. The eBMRQ, which indexes trait sensitivity to reward from music, remained a significant predictor of memory retrieval. Holding other variables constant, each unit increase in eBMRQ score (on a 120-point scale) was associated with an increase of 1.03 in the odds of recalling a memory, $F(1,3713) = 5.16$, $p = .023$. Controlling for other variables, age was no longer a significant predictor of memory retrieval, $F(1,3713) < 1$.

Table 5

Generalized Linear Mixed Effects Model of Familiarity, Cue Type, eBMRQ score and Age Group on Successful Retrieval Rate in Experiment 2

Predictors	*F*	*df1*	*df2*	*p*
Familiarity	841.37	1	3713	< .001
Cue Type	3.06	3	3713	.027
eBMRQ	5.16	1	3713	.023
Age Group	0.35	1	3713	.554

Note: N = 155; eBMRQ = extended Barcelona Music Reward Questionnaire. df = degrees of freedom

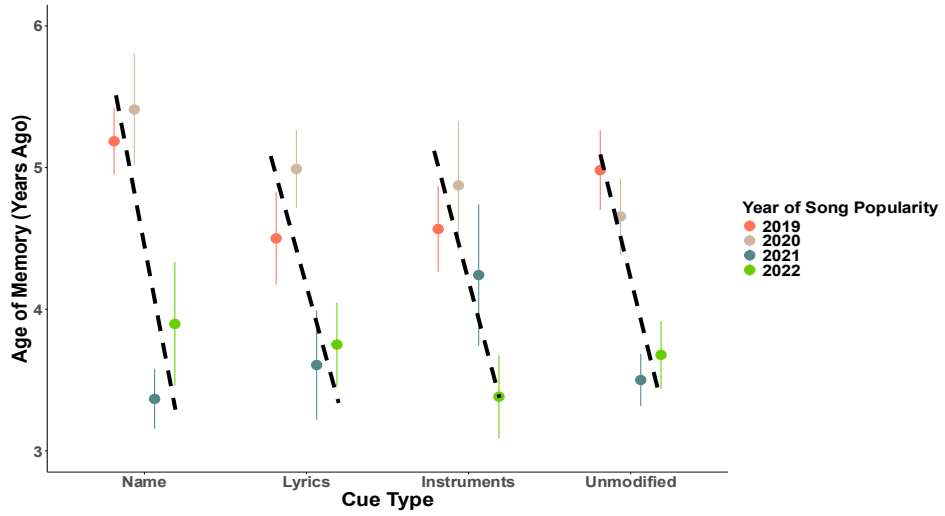
2.2.4.3 Temporal alignment between year of song popularity and age of memory

Year deviation was calculated by taking the difference between the year of a memory and the year the corresponding song cue was most popular. A temporal alignment was again observed between the year of song cue popularity and age of memory. That is, participants' memories tended to be from the time frame when the song cue was most popular. There was deviation of less than a year for younger and older adults (Young: $M_{deviation}$ = 0.15, SD = 0.62; Old: $M_{deviation}$ = 0.90, SD = 1.14). Remarkably, this temporal alignment effect was not only observed in younger adults, but also older adults in memories from over 50 years ago (see Figure 4). Additionally, I observed a temporal gradient effect, where more recently popular songs evoked more recent memories. This effect was seen in memories cued by all cue types in younger adults, but only memories cued by unmodified songs in older adults.

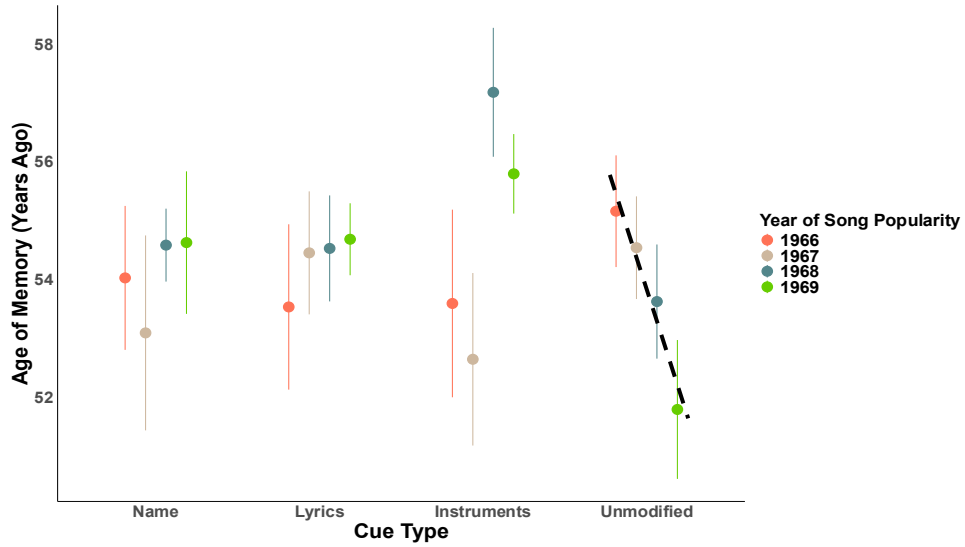
Figure 4

Age of Memory by Year of Song Popularity for each Cue Type in Young Adults (panel A) and Older Adults (panel B) in Experiment 2

A.



B.



Note. Points represent mean age of memory. Error bars represent ± 1 standard error. Data were collected in 2024 (i.e., a memory from 0 years ago would be from 2024). Memories in both young and older adults tended to be from the same time period in which the song cues were most popular. Trend lines are shown in black.

I conducted two mixed effects models, one for younger adults and one for older adults, predicting memory year. Year of song popularity (younger adults: 2019/2020, 2021/2022; older adults: 1966/1967, 1968/1969), cue familiarity, (1 = *not familiar,* 5 = *I completely knew it*), and cue type (song & artist name, lyrics-only, instruments-only, unmodified) and their interactions were included as fixed effects, and participant was included as a random effect.

For younger adults, results indicated a significant main effect of familiarity, $F(4,728) = 4.54$, $p = .001$, and cue type, $F(3,728) = 8.95$, $p < .001$. These main effects were qualified by two-way interactions between cue familiarity and year of song popularity, $F(4,728) = 7.91$, $p < .001$, between cue familiarity and cue type, $F(12,728) = 3.93$, $p < .001$, and between cue type and year of song popularity, $F(3,728) = 3.29$, $p = .020$. These in turn were qualified by a significant three-way interaction between cue familiarity, cue type, and year of song popularity, $F(10,728) = 2.91$, $p = .001$ (see Table 6).

To examine this three-way interaction, I examined the effect of Cue Type and Year of Song Popularity, as well as their interaction, at varying levels of song cue familiarity, split into tertiles representing low, medium and high levels of familiarity. See Table 6 for results. For song cues of relatively low familiarity, there was a main effect of Year of Song Popularity, $F(1,247) = 7.12$, $p = .008$, such that more recently popular song cues evoked more recent memories. There was no observed main effect of Cue Type, $F(3,247) < 1$, nor was there a significant interaction, $F(3,247) < 1$. In memories evoked by song cues of average familiarity, there was a main effect of Year of Song Popularity, $F(1,248) = 27.57$, $p < .001$. Specifically, more recently popular songs evoked more recent memories. There was no main effect of Cue Type, $F(3,248) < 1$, nor was there an interaction between Year of Song Popularity and Cue Type, $F(3,248) = 1.72$, $p = .164$.

Finally, in memories evoked by song cues of relatively high familiarity, I again observed a main effect of Year of Song Popularity, $F(1,247) = 47.49$, $p < .001$. Songs that were more recently popular evoked more recent memories. I observed no main effect of Cue Type, $F(3,247) = 1.27$, $p = .284$, nor was there an interaction between Cue Type and Year of Song Popularity, $F(3,247) = 1.88$, $p = .134$.

Table 6

Age of Memory by Year of Cue Popularity for each Cue Type in Young Adults, with Song Cues of Low, Average, and High Familiarity

		Age of memory (years ago from 2024)					
Cue Type	Year of song popularity	Low familiarity		Medium familiarity		High familiarity	
		M	*SD*	*M*	*SD*	*M*	*SD*
Song and artist name	2019	5.00	2.47	5.12	2.05	5.63	1.36
	2020	5.50	1.51	4.70	1.69	6.05	2.69
	2021	3.59	1.76	2.97	1.41	3.41	1.69
	2022	5.44	3.21	3.04	1.85	3.75	1.28
Lyrics-only	2019	4.93	1.90	4.42	1.93	4.39	2.19
	2020	4.00	1.86	5.38	1.50	5.47	2.21
	2021	4.91	2.97	3.21	1.38	2.83	1.45
	2022	3.77	2.60	4.05	1.29	3.50	1.36
Instruments-only	2019	3.96	2.03	3.32	1.40	5.60	1.84
	2020	5.58	3.18	4.50	2.04	3.00	1.91
	2021	4.25	2.41	4.94	3.66	3.10	1.67
	2022	3.70	2.39	2.79	1.27	3.90	1.26
Unmodified	2019	5.20	2.91	4.86	1.80	4.97	1.60
	2020	5.04	3.28	4.39	2.08	4.50	1.96
	2021	3.50	1.51	3.29	1.32	3.77	1.03
	2022	3.85	2.59	3.42	1.91	3.74	1.67

For older adults, results indicated a main effect of cue type, $F(3,896) = 3.00$, $p = .030$. Using Bonferroni-corrected post-hoc tests, unmodified song cues were found to evoke more recent memories than song and artist name cues, $t(896) = 3.52$, $p = .003$, and lyric-only cues, $t(896) = 3.59$, $p = .002$. There was also a main effect of familiarity, $F(4,896) = 5.96$, $p < .001$, which was qualified by a two-way interaction between familiarity and year of song popularity, $F(3,896) = 3.96$, $p = .008$.

Song cues with the lowest familiarity scores tended to show an effect of year of song popularity, such that more recently popular songs evoked more recent memories. However, these memories tended to be from years after the song cues were most popular, showing a temporal gradient rather than temporal alignment effect. More familiar song cues tended to evoke memories from the years 1968-1969, regardless of which year the song cue released between 1966-1969.

2.2.5 Discussion

Songs were able to successfully evoke memories nearly half of the time, replicating the first experiment. Additionally, like in experiment 1, evoked memories tended to be rated as moderately arousing, positive in valence, and associated with a moderate sense of re-living. Unmodified cues, containing both prose and instruments, again demonstrated the highest effectiveness in evoking memories, highlighting the importance of both these components in song's cueing ability. Older adults experienced more memories than younger adults and were found to uniquely benefit from the presence of prose in song cues. While cue familiarity was found to strongly predict cueing ability, unmodified cues persisted as a uniquely strong cue even when familiarity was controlled statistically. Trait reward sensitivity (eBMRQ) to music was found to provide additional explanatory power in song cueing, beyond cue familiarity.

I replicated, across experiments, a temporal alignment between the year of memory and year of song popularity. Notably, this alignment was seen in both younger and older adults, marking the first time it has been observed in the older population. This effect also persisted across different levels of cue familiarity in younger adults. In older adults, memories evoked

from low familiarity song cues showed a temporal gradient such that more recently popular songs evoked more recent memories. However, these memories tended to be from after the songs were most popular, whereas high familiarity song cues showed temporal alignment, evoking memories from around the same time period as when the songs were most popular. I return to discuss these findings further in the general discussion.

Chapter 3

Note. Experiments 3-5 of this thesis are in preparation for publication. That article is co-authored by myself, Lakshay Goel, Eric Anderson, Ryan Yeung, Marek Stastna and Myra A. Fernandes. I was involved in the conceptualization of the study, data curation, formal analysis, methodology development, project administration, providing of resources, visualization of data, writing of the original draft as well as review and editing. Portions of this thesis have been derived from or taken verbatim from this manuscript.

3.1 Growth of computational methods

The development of text analysis tools has grown significantly in recent years, driven by advancements in natural language processing (NLP), machine learning, and artificial intelligence. These tools have become increasingly accessible, allowing businesses and individuals to analyze large volumes of text data for insights into consumer behavior and social trends. At the same time, the quantity of text data available for analysis has soared in the past two decades, in a phenomenon known as 'big data' (Borgman, 2015). The recently available volume of text data has diminished the practicality of time-consuming and laborious manual methods of text analysis, spurring the development of automated text analysis tools (Dhaoui et al., 2017; Wang et al., 2012). To modernize the study of narrative transcripts, and better characterize them, analysis of larger samples than in past work is needed. Application of NLP to human memory narratives can offer insights into yet undiscovered age differences in music's cueing ability.

3.2 Content analysis in psychology

Recent work has attempted to use people's writings to provide insight into their cognitive capacity (Tausczik & Pennebaker, 2010), affective states (Rathje et al., 2024), or mental health (Yeung & Fernandes, 2020; 2021). Past and current popular techniques in the field, such as the

Thematic Apperception Test (Morgan & Murray, 1935), General Inquirer (Stone et al., 1962), and LIWC (Tausczik & Pennebaker, 2010) have shared a common approach. Short pieces or individual words within a text are characterized based on membership in multiple a priori defined psychologically relevant categories (e.g., 'admirable' would be categorized as positive and a virtue). These word-category associations are stored and drawn from a pre-defined 'dictionary,' inspiring the name for this type of analysis: dictionary methods. The frequency distribution of words in each category is then used to characterize a text or corpus of texts. While a prominent procedure for content analysis today in psychology, many of these methods require extensive manual human work to classify such writings (Krans et al., 2015; Newby & Moulds, 2012; Schlagman et al., 2006), limiting the sample size of text that can feasibly be analyzed.

While these dictionary methods have demonstrated success in gleaning the properties of text content, the use of pre-defined categories can bias the analysis of text themes to researcher expectations (Yeung et al., 2022). Frequency analyses have since developed as a complementary way to capture the content within texts. Instead of placing text words into pre-defined categories, frequency analysis utilises counts of each word in a corpus of text to describe its themes. Word clouds (graphic representations of words where the size of the word corresponds to its frequency of use) have been used to visualize frequencies of words (Eichstaedt et al., 2021; Sreekumar et al., 2018; Yeganeh et al., 2020), and praised for their ability to display content information, as well as their ease of creation using computational tools (DePaolo & Wilkinson, 2014; Henderson & Segal, 2013). However, word clouds lack quantitative metrics. To compensate, frequency analysis methods have been recommended to be paired with dictionary methods (DePaolo & Wilkinson, 2014; Henderson & Segal, 2013; Yeung et al., 2022). Here, I utilised both dictionary

methods and frequency analyses to characterize text from cognitive studies, as well as to explore the effectiveness of automated methods of analysis.

3.3 Valence is relevant to content

The sentiment, or valence of text data is a particularly useful feature to characterize the content of text (Barrett, 2006). Valence is conceptualized as a key dimension of affect (Russell, 1980), used to describe the level of positivity or negativity being conveyed within a piece of text. The valence of a text has been of particular interest to psychological researchers, as it is believed to reveal the writer's past emotional experience (Charland, 2005) or their present affective state (Ortony et al., 1988). Self-reporting has been the principal method used to determine valence level (Barrett, 2004; Bradley & Lang, 1994; Watson et al., 1988). However, use of varying methods for self-reporting valence makes it difficult to compare results across studies, and these methods are fundamentally dependent on text data being labelled with an associated self-reported valence (Zentner & Eerola, 2010). NLP methods to determine text valence may be a promising tool for addressing these limitations and elucidating the meaning behind text data.

Sentiment analysis algorithms are a subclass of NLP methods purposed to detect and quantify the valence of text within large corpora (Paltoglou, 2014). One subset of these algorithms utilises a dictionary approach, resembling that of the General Inquirer (Benoit, 2020). Each word within a text is labelled with a valence score (i.e., level of positivity or negativity) based on that word's pre-defined score in the algorithm's dictionary of words (Provoost et al., 2019). The other common approach employs machine learning algorithms that are 'trained' and learn from manually annotated example classified texts (Paltoglou, 2014; Provoost et al., 2019).

3.4 Use of Sentiment analysis in psychology

The Valence Aware Dictionary and sEntiment Reasoner (VADER) is a sentiment analysis method developed by Hutto & Gilbert (2014) that utilises a lexicon (i.e., dictionary), rule-based algorithm to detect text valence. While designed as a specialist tool to analyze social media text, it has become popular for a range of applications (Asri & Fajri, 2023; Borg & Boldt, 2020; Kavitha & Celina, 2024). VADER has been used in psychological research to evaluate feature similarity of stimuli during stimulus development (Grall et al., 2021) as well as to assess risk of psychological distress (Viviani et al., 2021).

TextBlob is another popular lexicon, rule-based analysis tool that has been applied in psychological research (Loria, 2024). It has been used to determine affective content within children's writing (Dong et al., 2024), and has been compared to VADER in its effectiveness in tracing the emotional effects of the COVID-19 pandemic (Razzaq et al., 2022). While the validity of VADER and TextBlob has been compared (Abiola et al., 2023; Min & Zulkarnain, 2020), with results suggesting VADER edges out TextBlob in validity, these comparisons have largely been conducted solely on social media posts. There is currently no comparison of their accuracy in psychological research.

3.5 Current study

In Experiment 3, I used NLP methods to assess the sentiment contained within autobiographical narratives. Specifically, I investigated whether there is converging validity in classification of valence as determined by self-assessment and by two different NLP methods: VADER and TextBlob. Sentiment scores generated by each method were compared to

participant self-reported valence scores, the current prevailing gold-standard used in psychology for identifying text valence (Barrett, 2004; Bradley & Lang, 1994; Watson et al., 1988).

VADER and TextBlob were chosen from a range of possible sentiment analysis methods due to their popularity compared to other lexicon methods, and their demonstrated effectiveness across domains (Abiola et al., 2023). These methods also utilise the same lexicon approach that has been applied in prior content analysis work in psychology (Stone et al., 1962; Tausczik & Pennebaker, 2010). Previous works suggest machine learning sentiment tools could be even more effective than lexicon dictionary methods (Paltoglou, 2014; van Atteveldt et al., 2021; Weissman et al., 2019). I did not include this approach here since these require manual coding of a large number of sample texts to train the algorithm before implementation. Our goal was to provide psychological scientists with a rapid means of classifying valence in text that is congruent with human self-reports, using currently available sentiment analysis tools.

3.6 Experiment 3

3.6.1 Data preparation

I examined narratives taken from experiment 1 of this thesis. In total, 909 memories were elicited, collapsed across cue type, with an average of 11.96 memories per participant (SD = 4.57). Prior to analysis, spelling mistakes were corrected manually by authors KH and EA. The endings of memory narratives with incomplete sentences (e.g., no noun-verb connection, ending on an article) were trimmed manually by author KH to the last complete phrase so that each sentence in the narrative ended in a complete sentence.

I then used a minimum word count method to remove invalid text from the dataset, a step recommended to enhance data quality (Banks et al., 2018; Chen, 2011). As the length of text

responses can vary by study (Banks et al., 2018), I chose to use a word count cut-off tailored to our dataset. I removed texts containing a word count less than 'the mean of the dataset minus 1 standard deviation', as done by Yeung & Fernandes (2022). This led to the removal of text responses that were less than 11 words in length. Using this approach, 9.1% of the transcripts were classified as invalid and removed, leaving 825 narratives for analysis.

As recommended by Banks and colleagues (2018), I conducted additional preprocessing steps. First, I tokenized each text into separate words and symbols. I then used the spaCy library (Honnibal et al., 2020) to conduct lemmatization: the reduction of morphological variants of a word (e.g., 'changing,' 'changed') into a single base form (e.g., 'change'). The word 'singing' was lemmatized manually into 'sing,' instead of the automated default of 'singe.' I then removed stop words: words from the narratives that appeared frequently but added little value in understanding their content (e.g., 'the,' 'a,' 'in'), using the Snowball stop word list (Benoit et al., 2018; Porter, 2001). However, I did not remove negators that are utilised by the sentiment analysis tools (e.g., 'not,' 'doesn't'). Finally, I applied vocabulary pruning to remove words that are too frequent to be analytically useful, and removed a total of ten words that appeared at a frequency of 4 SDs greater than $M_{word\ frequency}$ (e.g., 'remember', 'remind', 'listen'). Many of these processes, including tokenization, lemmatization, removal of stop words and vocabulary pruning, were automated using Python, and conducted within a couple minutes. Removal of invalid texts was conducted using MS Excel, and was semi-automated and completed within five minutes.

3.6.2 Frequency analysis

I conducted frequency analyses to attain a coarse perspective on the topic and content of the narratives; this was also done to determine whether topics differed by self-reported memory valence (Ball, 1994). In a frequency analysis, the number of times each word occurs within a corpus of text is counted. I used word clouds to visualize word use frequency, across the whole dataset.

3.6.3 Sentiment analysis - VADER and TextBlob

I used sentiment analyses to characterize the valence of participants' memory narratives (Paltoglou, 2014). I utilised two different sentiment analysis methods – VADER and TextBlob – and compared their results against participants' self-reports of memory valence, to evaluate the accuracy of these tools in classifying memory narratives. Both methods utilise a dictionary, rule-based method, in which each word is given an assigned valence score based on a pre-defined 'dictionary'. Sentiment scores are then totalled to obtain a score for each narrative. Both methods also consider relevant nearby words, such as negation words (e.g., 'not,' 'doesn't') and modifier words (e.g., 'strongly,' 'slightly') to tailor the assignment of valence for a word (e.g., 'fun'). However, the way these nearby words are considered varies between programs: First, VADER contains a larger dictionary of potential modifiers than TextBlob. However, while VADER applies a uniform effect of modifier on word valence, regardless of the modifier, TextBlob considers the intensity of the modifier and adjusts its effect accordingly. TextBlob also only adjusts word valence in response to negators/modifiers that occur immediately before the word, whereas VADER considers other possible negators/modifiers. For instance, the phrase 'not the

best' would be classified as negative by VADER but be classified as positive by TextBlob, since the negator 'not' does not appear immediately before the word being scored: 'best.' As well, each tool uses different dictionaries, and they utilise different methods to total assigned valence scores for each narrative. TextBlob calculates a weighted average, while VADER calculates a summed score and normalizes it. Finally, VADER classifies any words in text that are not in its dictionary as neutral, whereas TextBlob ignores these words. Such differences leave the possibility for varying results across methods.

Both methods assign a valence score to each word and calculate a compound score for each narrative by totalling the individual valence scores, obtaining a compound score between [-1.0, 1.0], where -1.0 is most negative and 1.0 is most positive. Thresholds by compound score were then used to categorize each narrative. Narratives with a compound score of [-1.0, -0.6), [-0.6, -0.2), [-0.2, 0.2], (0.2, 0.6], or (0.6, 1.0], were classified as very negative, somewhat negative, neutral, somewhat positive, or very positive, respectively.

3.6.4 Results

3.6.4.1 Prevalence of autobiographical memories by valence

All participants reported experiencing at least two memories. Participants on average reported experiencing a memory on 50% of the 24 trials. Out of the valid text responses, 1% were self-rated as very unpleasant, 7% were self-rated as somewhat unpleasant, 27% as neutral, 41% as somewhat pleasant, and 24% as very pleasant. Average length of memory narratives after applying a minimum word count of 11 words for removal of invalid texts was 28.04 words (*SD* = 10.99).

3.6.4.2 Frequency analysis

3.6.4.2.1 Overall frequency

Frequency analyses were conducted using the ‘wordcloud’ package in Python (Mueller, 2023). Frequency was determined by post-lemmatization word count across the entire corpus. I see similar word usage to a previous study on the content of music-evoked autobiographical memories (Jakubowski & Ghosh, 2019).

As shown in Figure 5, the music itself was often referenced (‘music’), and participants commonly cited media where music is often played (e.g., ‘radio,’ ‘tiktok,’ ‘video,’ ‘car,’ ‘concert’) and activities involving music (e.g., ‘dance,’ ‘sing’). The singers of the audio cues were frequently mentioned. Narratives tended to reference ‘summer.’ ‘One’ and ‘first’ were additionally among the most frequently used words.

Figure 5

Word cloud of top 200 most frequently used words in music-evoked autobiographical memories



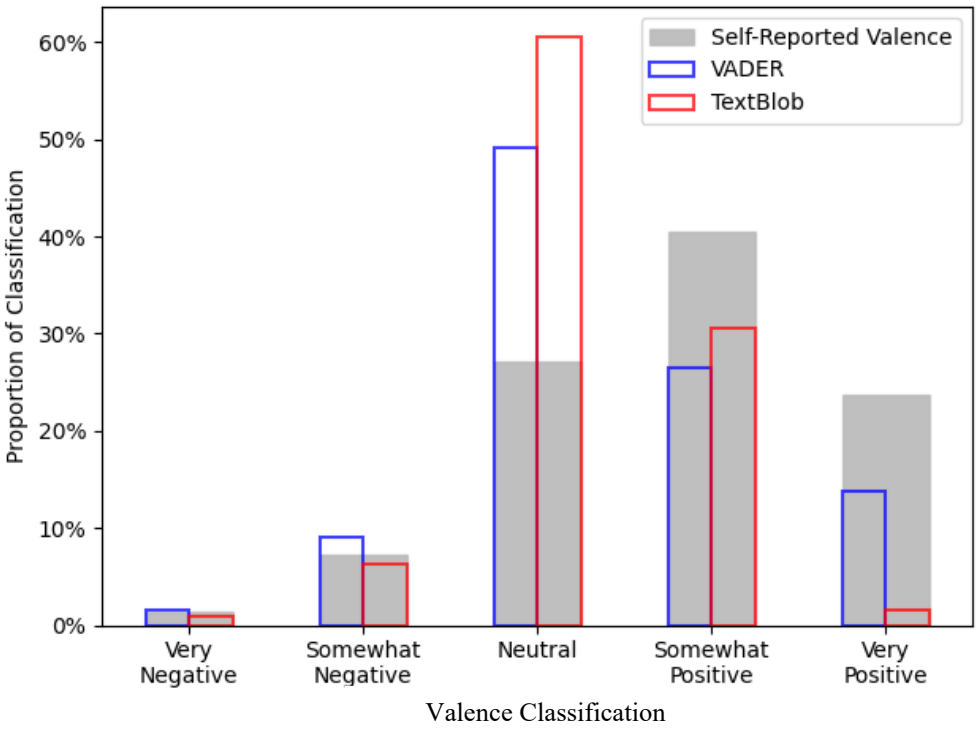
Note. Larger word size corresponds to higher token frequency.

3.6.4.3 Sentiment analysis

I utilised VADER (Hutto & Gilbert, 2014) and TextBlob (Loria, 2024) to computationally assign sentiment scores to each of the 825 narratives. The frequency of memories classified by valence using VADER, TextBlob and self-reports are displayed in Figure 6.

Figure 6

Frequency of memories by valence, calculated using self-reports, VADER and TextBlob



Note. Continuous sentiment scores by VADER and TextBlob were split into five categories using equally spaced thresholds.

Using self-reported valence as the gold standard, by visual inspection, TextBlob appears to be slightly more effective at classifying memories rated as somewhat negative or somewhat positive, whereas VADER displays greater accuracy in classifying memories rated as neutral or very positive. Both VADER and TextBlob display a considerable bias for classifying memories as neutral, while classifying substantially fewer memories as somewhat or very positive.

3.6.4.3.1 Model Validation

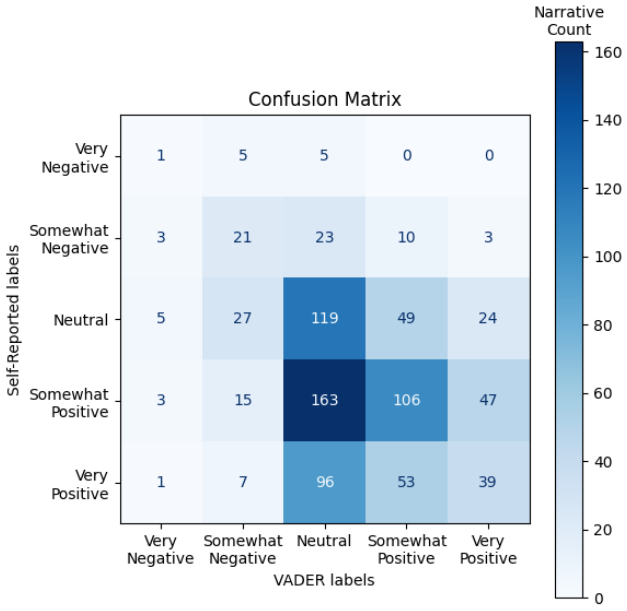
To more rigorously evaluate these two tools, I calculated confusion matrices for each, comparing their sentiment classifications to self-reported valence scores (see Figure 7).

Figure 7

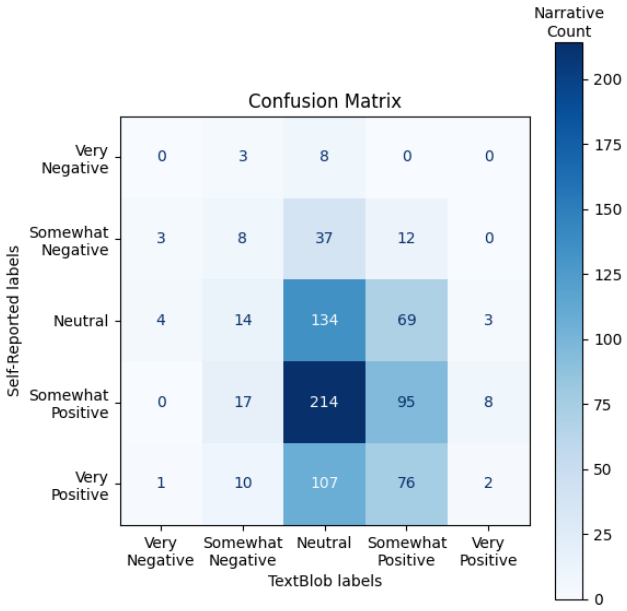
Confusion matrices of valence classifications of VADER (panel A) and TextBlob (panel B) relative to self-reported memory valence

A

.



B



I used the Matthews Correlation Coefficient (MCC; Matthews, 1975) to assess the relationship between computationally derived and self-reported valence scores. This coefficient suits the purposes of this study best since it accounts for one class (somewhat positive memories) in this dataset being far more populous than another class (very or somewhat negative memories), which would otherwise allow a package to be rated as effective, despite it only being able to classify the populous class (Chicco & Jurman, 2020). Using the MCC, these methods were evaluated based on their concordance in classification of memories. VADER had a stronger relationship with self-reports (MCC = .10) than TextBlob (MCC = -.01), $z = 2.23$, $p = .013$. While TextBlob's sentiment scores particularly failed to produce positive correlations with self-reports, both methods displayed strikingly low correlations with self-reports.

3.6.5 Discussion

The results from this dataset indicated that VADER was slightly better than TextBlob in its congruence with self-reported valence classification of memory narratives, although both methods demonstrated poor alignment with self-reports. This dataset, however, was dominated by texts with self-reported positive valence, making it difficult to ascertain the effectiveness of these tools in evaluating narratives that are more neutral or negative. Additionally, the narratives in this dataset were quite short (averaging around 28 words); this contrasts with the longer narratives often found in qualitative data collected using more formal interview methods in psychology. Finally, our dataset was relatively small compared to other datasets used for computational analysis (Dong et al., 2024; Tanana et al., 2021). As such, in experiment 4, I applied these evaluation methods to a larger dataset of autobiographical memories from Yeung et

al. (2024), which contained a more balanced distribution of self-reported valence and consisted of longer narratives.

3.7 Experiment 4

The second dataset in this thesis chapter was obtained from a study by Yeung et al., (2024) that identified the content of recurrent, involuntary and spontaneous recalled autobiographical memories (rIAMs) as a tool to distinguish the experiences of boredom proneness versus depression symptoms.

3.7.1 Participants

Undergraduate students were recruited from the University of Waterloo, who participated in the study in return for course credit. Data were collected between September 2018 and January 2020. That study recruited 6,187 unique participants, with 2,626 participants reporting having experienced a rIAM. In the final sample, 71% were women, 28% were men, and 1% were nonbinary, genderqueer or gender nonconforming. Mean age was 19.9 (*SD* = 3.3, range = 16-49).

3.7.2 Materials & Methods

3.7.2.1 Recurrent Memory Scale

Participants indicated if they had experienced at least one recurrent rIAM within the past year, not within the past year, or never. If a participant had experienced a rIAM within the past year, they were asked to write a 3-5 sentence description of it, noted how long ago the original event occurred, and then rated the properties of their evoked memory (e.g., frequency of

recurring, perceived detail/completeness, vividness of visual imagery, valence, and emotional intensity) on a set of 5-point Likert scales.

Participants self-registered and completed the study online. After providing informed consent, participants completed a series of questionnaires, including the Recurrent Memory Scale and self-report scales of trait boredom proneness and depression symptoms. For the current study, I analyzed only the written memory narratives and the valence self-reports from the Recurrent Memory Scale; all other measures collected were unrelated to the current study. All procedures were approved by the University of Waterloo's Office of Research Ethics (Protocol #40049).

3.7.2.2 Data preparation

Prior to receiving the dataset, text responses underwent validation screening (removal of meaningless or irrelevant text) using a supervised machine learning program (Yeung & Fernandes, 2022). Using this approach, 5.4% of responses were classified as invalid and removed, resulting in a final total of 2,484 narratives.

I applied the same preprocessing procedures as in experiment 1, such as lemmatization, stop word removal and vocabulary pruning (Banks et al., 2018), not including spelling correction, editing of incomplete sentences, and removal of invalid texts using a minimum word count method.

3.7.3 Results

3.7.3.1 Prevalence of autobiographical memories by valence

Of the 6,187 participants recruited for this study, 3345 (54%) reported experiencing a rIAM within the past year. Out of the valid text responses, 25% were self-rated as very negative, 29% were self-rated as negative, 17% as neutral, 15% as positive, and 14% as very positive. Average length of memory narratives was 32.22 words (*SD* = 22.21).

3.7.3.2 Frequency analysis

3.7.3.2.1 Overall frequency

Frequency analyses were conducted using the 'wordcloud' package in Python (Mueller, 2023). Frequency was determined by raw word count across the entire corpus. The most frequently used words tended to reference other individuals (e.g., 'boyfriend,' 'person,' 'people,' 'someone'). Words that could be providing temporal context were commonly used (e.g., 'childhood,' 'moment,' 'old'). 'Walk' was frequently used. Other common words included 'good,' 'life,' 'often,' and 'relationship'. Figure 8 shows a word cloud of the most frequently used words.

Figure 8

Word cloud of top 200 most frequently used words in recurrent involuntary autobiographical memories.

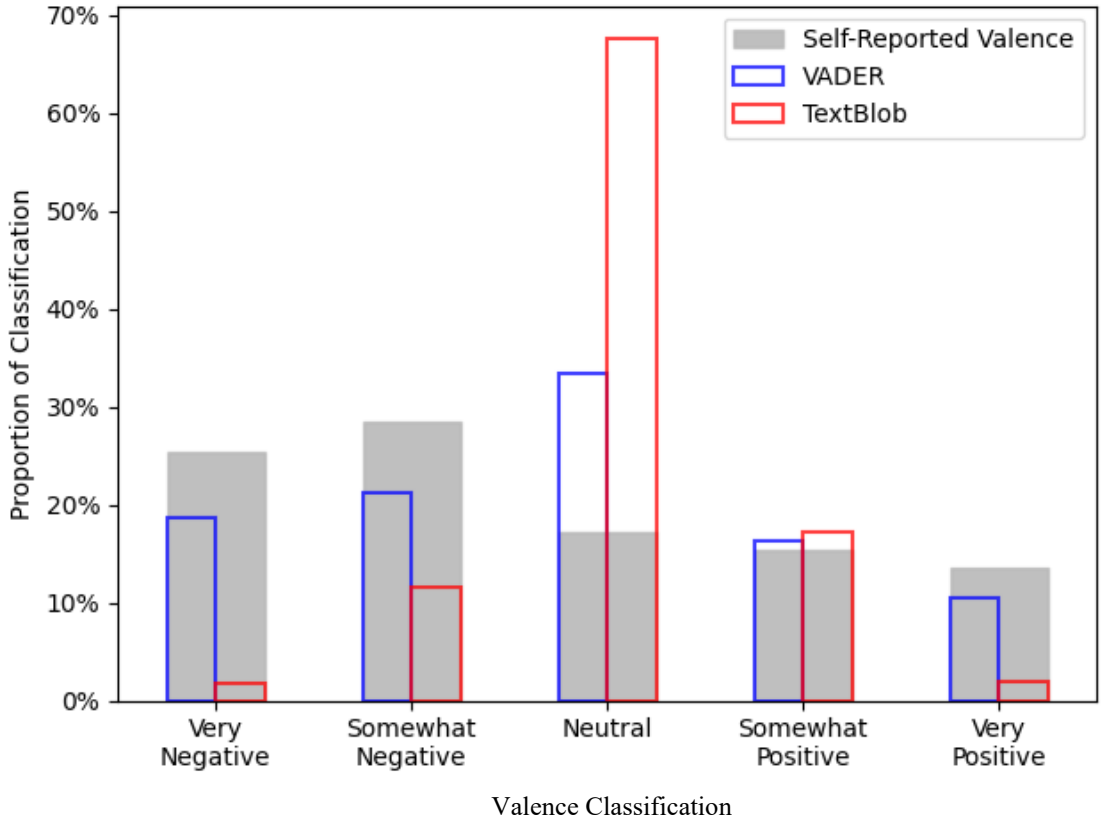


3.7.3.3 Sentiment analysis

I utilised VADER (Hutto & Gilbert, 2014) and TextBlob (Loria, 2024) to computationally discern the valence of memory narratives. As in experiment 1, thresholds for compound scores were then used to categorize each narrative as a certain valence. Narratives with a compound score of [-1.0, -0.6), [-0.6, -0.2), [-0.2, 0.2], (0.2, 0.6], or (0.6, 1.0], were classified as very negative, somewhat negative, neutral, somewhat positive, or very positive respectively. The frequency of memories classified by valence using VADER, TextBlob and self-reports are displayed in Figure 9.

Figure 9

Frequency of memories by valence, calculated using VADER, TextBlob and self-reports.



Note. Continuous valence scores by VADER and TextBlob were split into five categories using equally spaced thresholds.

Using self-reported valence as the gold standard, apart from somewhat positive memories, where they are nearly equivalently accurate, VADER appears to be more accurate than TextBlob in classification of nearly all valences. As also seen in the previous dataset, both TextBlob and VADER displayed a considerable bias towards neutral categorizations, although this was more pronounced in TextBlob classifications.

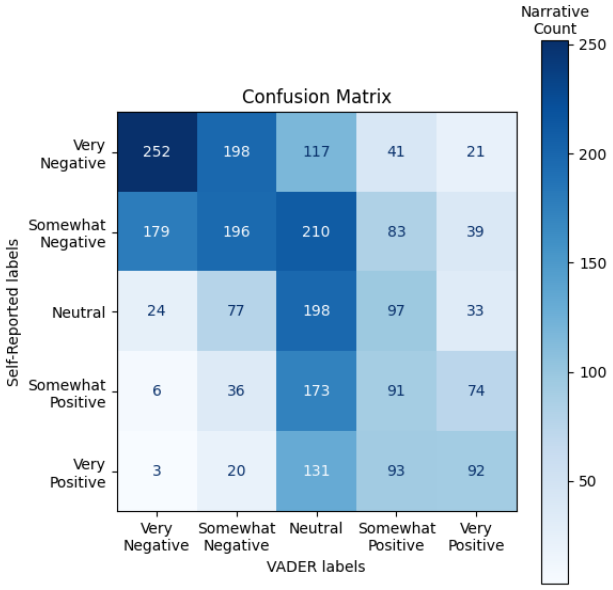
3.7.3.3.1 Model Validation

To more rigorously evaluate these two methods, I computed confusion matrices for each tool, comparing their valence classifications to self-reported valence scores (see Figure 10).

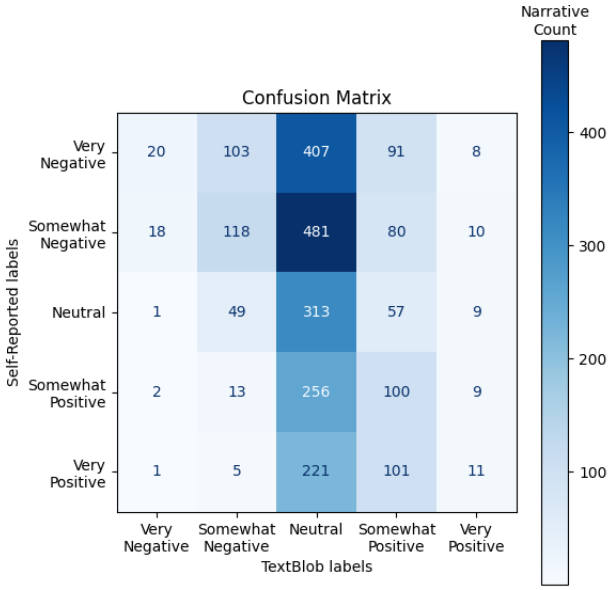
Figure 10

Confusion matrices of valence classifications of VADER (panel A) and TextBlob (panel B) relative to self-reported memory valence.

A.



B.



As done with the previous dataset, I used the MCC (Matthews, 1975) to assess the relationship between computationally derived and self-reported valence scores. Using the MCC, VADER and TextBlob were evaluated based on their classification accuracy. As in experiment 3, VADER had a stronger relationship with self-reports (MCC = .17) than TextBlob (MCC = .07). VADER was found to be statistically more effective than TextBlob in its classifications, $z = 3.49$, $p < .001$.

3.7.4 Discussion

The results from this dataset indicate that VADER outperformed TextBlob in classification of sentiment within memory narratives. Both methods also displayed increased accuracy when applied to the dataset from experiment 4 than 3. The low congruence with self-reported valence in our experiments, stands in contrast to when VADER and TextBlob were applied to datasets from other fields (Bonta et al., 2019; Nariman, 2024), and some clinical studies in psychology (Eberhardt et al., 2024; Provoost et al., 2019; Tanana et al., 2021; Weissman et al., 2019).

In experiment 5, I investigated potential reasons for the discrepancy in classification accuracy across the two datasets, and explored how threshold choices for valence categories may impact effectiveness of these tools.

3.8 Experiment 5

In this experiment, I sought to identify possible causes for variations in classification accuracy. First, I aimed to determine whether differences in dataset size might have influenced sentiment analysis performance. To investigate this, I conducted an ablation analysis using the

larger dataset in experiment 4 and compared classification accuracy of VADER and TextBlob across varying sample sizes of the rIAMs dataset.

I then investigated whether the number of categories for classification of valence influenced the effectiveness of VADER and TextBlob in classifying sentiment of texts. In the previous experiments, computationally assigned sentiment scores, varying between -1 and 1, were placed into equally spaced bins across five categories. Since three-category classification is more common in the sentiment analysis literature (Bonta et al., 2019; Elbagir & Yang, 2019; Guha Majumder et al., 2022; Hutto & Gilbert 2014; Lossio-Ventura et al., 2024), I explored whether there was a difference in software effectiveness using 3-category versus 5-category classification of our rIAMs dataset. Thirdly, while a couple thresholds, such as ±0.001 and ±0.05, have been recommended and frequently used in the literature for 3-category classification (Bonta et al., 2019; Elbagir & Yang, 2019; Hutto & Gilbert, 2014; Lossio-Ventura et al., 2024), there is no standard specifying appropriate thresholds for psychology application. Moreover, the two methods in our study, Vader and TextBlob, have differences in their methodologies that may require different thresholds for optimization. Hence, I wanted to determine whether there were differing optimal thresholds for each method using 3-category classification in the rIAMs dataset.

3.8.1 Materials and methods

3.8.1.1 Ablation analysis

Ablation analyses were conducted on the rIAMs dataset (Yeung et al., 2024) to determine if tool performance varied across different subset sizes. Ablation analysis involved splitting the dataset into smaller subsets of varying sizes and comparing the effectiveness of VADER and

TextBlob within each of these different subsets. Sentiment analysis was conducted in subsets of sizes 1800 (75%), 1200 (50%), 825 (33%), 500 (20%), 250 (10%) texts. A subset of size 825 (33%) was specifically chosen because this size coincides with that of the MEAM dataset from experiment 1. Subsets were randomly sampled 50 times per subset size to determine an average MCC score for each sample size. This was conducted using 3-class and 5-class categorization with the thresholds described below.

3.8.1.2 Three-class vs. five-class analysis

For 3-category valence classification, memory narratives were classified as either negative, neutral, or positive. Self-reported scores of "very negative" and "somewhat negative" were grouped as "negative" and similarly, "very positive" and "somewhat positive" were grouped as "positive" narratives. Computational valence scores in the ranges [-1.0, -0.01), [-0.01, 0.01], and (0.01, 1.0] were classified as negative, neutral, and positive respectively. 5-category valence classification was conducted as done in previous experiments. See the explanation of sentiment analysis in experiment 3 for more information.

3.8.1.3 Threshold analysis

To evaluate possible thresholds, I defined the 'positive threshold' as the threshold to classify 'positive' from "neutral" memory narratives and the 'negative threshold' as the threshold to classify 'negative' from 'neutral' narratives. To find the optimal threshold to classify narratives from the rIAMs dataset, I incrementally varied the positive threshold by 0.01 between 0 and 1.0. Likewise, I incrementally varied the negative threshold by 0.01 between 0 and -1.0. MCC scores were computed for each negative and positive pair. This method was conducted

using VADER and TextBlob. For each computational method, I found the negative and positive pair that gave the maximum MCC and thus found the optimal threshold.

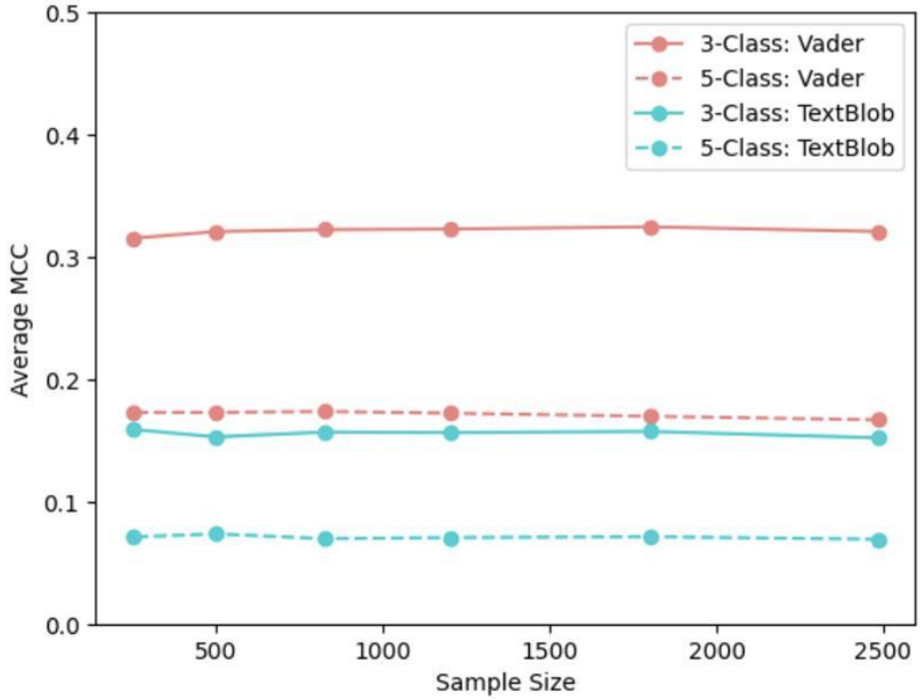
3.8.2 Results

3.8.2.1 Ablation analysis

Congruence between scores from each NLP method and self-reported valence was evaluated at different sample sizes of the dataset from experiment 4 (Yeung et al., 2024). Results are shown in Figure 11. By inspection, results of this ablation analysis indicated no considerable difference in congruence for either tool across different subset sizes (all lines are flat).

Figure 11

Average MCC between sentiment categories and self-reported valence at five different subset sizes of the rIAMs dataset (Experiment 2)



Note. Subsets were randomly sampled 50 times. Results are shown for 3-class and 5-class valence categorization.

3.8.2.2 Three-class vs. five-class analysis

I replicated the previous ablation analysis but manipulated the number of valence categories, assigning narratives into one of three valence classes rather than five valence classes. Performance of both sentiment analysis tools was found to improve considerably when valence categorization was constrained to three classes. Congruence between TextBlob and self-reports increased from r_{MCC} = .07 to .15, z = 2.96, p = .002, and VADER experienced a bigger jump from r_{MCC} = .17 to .32, z = 5.78, p < .001. Like with 5-category classification, performance did not differ when applied to different subset sizes of the rIAMs dataset.

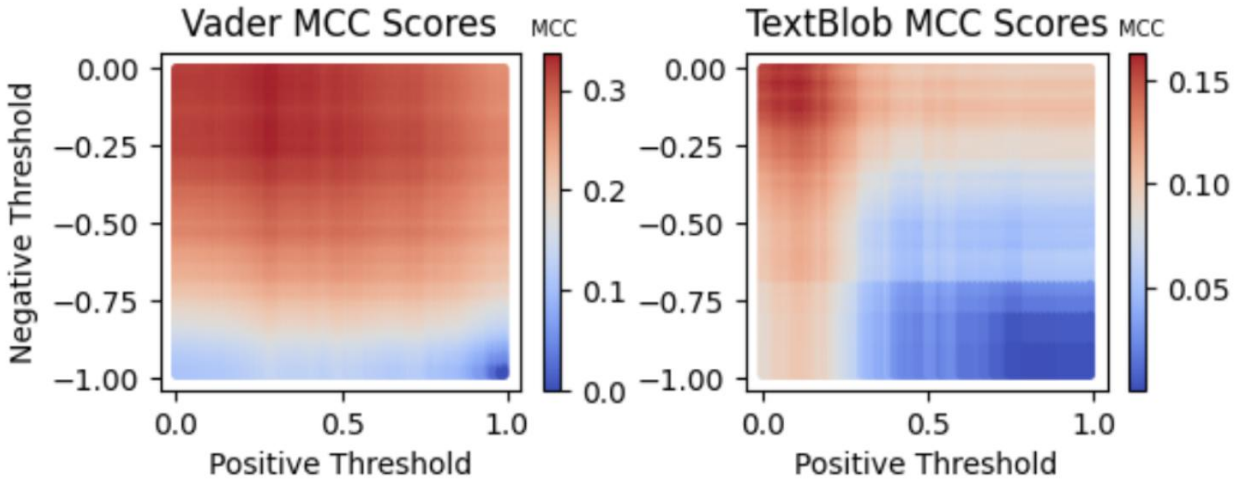
3.8.2.3 Threshold analysis

I appraised congruence of VADER and TextBlob with self-reports in 3-category valence classification using all possible symmetrical threshold combinations. Heat maps of congruence of VADER (A) and TextBlob (B) with self-reports in the rIAMs dataset using different negative and positive bounds for valence classification are shown in Figure 12.

Figure 12

Heat maps of correlations between self-reported valence and the sentiment categories at different category thresholds

A. B.



Note. Colour map shows MCC values. Results are separated by VADER (Panel A) and TextBlob (Panel B), utilising all possible negative and positive thresholds**.**

VADER operates effectively across a wider range of threshold combinations than TextBlob. VADER functions very effectively with a negative threshold range of 0.0 – -0.30, and a positive threshold range between 0.00 – 0.70. VADER is most effective with a negative threshold of -0.03 and a positive threshold of 0.28, achieving a r_{MCC} = .34. This correlation is not

significantly higher than the correlation of r_{MCC} = .32, z = 0.62, p = .267 obtained using the thresholds from the ablation analysis. TextBlob has relatively stricter threshold ranges, with an effective negative threshold range between 0.00 – -0.20, and a positive threshold range of 0.00 – 0.20. The optimal combination for TextBlob was a negative threshold of -0.05, and a positive threshold of 0.12. With these thresholds, TextBlob demonstrates a correlation of r_{MCC} of .16, which is not significantly higher than the correlation of r_{MCC}= .15, z = 0.39, p = .350 observed using the thresholds from the ablation analysis.

3.9 Discussion

Sentiment analysis tools have been employed to extract valence within large corpora of text. While the success of these tools has been substantiated in industry applications and in data analysis for multiple scientific fields, their effectiveness in characterising psychology data is unverified. I compared the congruence of two prominent dictionary-based sentiment analysis tools, VADER and TextBlob, with self-reported valence in classification of memory narratives taken from two different datasets. VADER consistently produced results that were more congruent with self-rated valence than did TextBlob. Effectiveness of both tools was lower than seen in other applications, but performance could be preserved depending on the procedure applied for their use, and the characteristics of the analysed dataset. For example, I found that congruence improved to nearly the same effectiveness as human coders when narratives were binned into only three rather than five valence categories. Below I suggest practices to preserve effectiveness of these tools as they permeate into psychology.

3.9.1 Congruence of classification

I first applied these computational tools to a dataset of 825 autobiographical memory narratives evoked by different components of recent popular songs (the MEAM dataset from experiment 1). Frequency analysis of the dataset revealed that the most frequently used words tended to pertain to the song cue itself. These results replicate findings from previous thematic analyses of memories evoked by music (Jakubowski & Gosh, 2019). Comparison of sentiment analysis tools revealed that VADER (r_{MCC}= .10) was more congruent with self-reported valence than TextBlob (r_{MCC}= -.01), which effectively showed no correlation with self-reported valence. Both tools, especially TextBlob, displayed a bias to rating memories as neutral, especially those self-rated as positive or very positive.

In experiment 4, I applied these tools to a second, larger dataset to investigate whether the features of the narrative dataset, such as dataset size, narrative length and valence skew could impact performance. First, I conducted a frequency analysis, finding that commonly used words often referred to social themes and other individuals. Both sentiment analysis tools demonstrated improved performance, showing modest congruence with self-reports (r_{MCC} = .17 and .07 respectively). A neutral bias was again observed, largely at the cost of narratives self-rated as somewhat or very negative. VADER again outperformed TextBlob.

One distinctive use of sentiment analysis tools in our study was the categorization of narratives into one of five valence categories rather than three. While 5-point Likert scales are common in self-report scales, sentiment analysis tools have typically been applied for 3-category valence classification. To investigate whether this impacted tool performance, I investigated whether this factor influenced VADER and TextBlob accuracy in classifying narratives.

Congruence of VADER and TextBlob increased to around 0.32 and 0.15, an increase of around 0.15 and 0.08 respectively. Like before, sample size did not have an impact on performance. The improvements with 3-class models suggests that sentiment tools have difficulty capturing nuanced valence classifications. Given this, I suggest sentiment analysis is best when applied to determine broad valence classifications (negative, neutral, positive).

3.9.2 Factors influencing sentiment analysis tool classification accuracy

In experiments 3 and 4, a large discrepancy in performance was observed. To determine whether dataset size could explain this, I compared sentiment analysis performance across different size subsets within the larger dataset from experiment 4. Average performance of both tools was consistent regardless of subset size, suggesting that size of a dataset does not impact their performance. These results align with expectations; unlike other sentiment analysis approaches like machine learning that improve in performance with a larger training dataset, each narrative here is analysed independently of other narratives in the dataset. Encouragingly, results suggest that these tools can be utilised to analyse datasets containing at least 250 narratives without compromise in performance. However, it remains unclear why there were differences in performance across datasets. One possible explanation is the large skew in valence characteristic of the MEAM dataset (experiment 1). Both tools showed a tendency to assign neutral sentiment to documents, and this may have exaggerated performance loss in the smaller MEAM dataset in experiment 3, where a large proportion of the narratives were self-rated as somewhat or very positive. The topics of each dataset may have also influenced effectiveness. As seen in the frequency analyses, the most used words in each dataset differed from each other, and

it is possible that words characteristic of one dataset may have better aligned with words in the pre-defined tool dictionaries. Further explorations are necessary to isolate what dataset factors influence the performance of these lexicon-based sentiment analysis tools.

In an effort to improve performance, I investigated how the choice in thresholds for valence categories could impact effectiveness. Each tool provides a continuous valence score between [-1, 1] for each narrative, and cut-off thresholds are required to categorize these scores as representing a certain valence. While there are general guidelines in the literature on thresholds for 3-category valence classification (Bonta et al., 2019; Elbagir & Yang, 2019; Hutto & Gilbert, 2014; Lossio-Ventura et al., 2024), there is no standard specifying appropriate thresholds for psychology applications. I used threshold analysis to determine the most effective threshold combination, and whether it differs for each sentiment analysis tool. I compared the effectiveness of VADER and TextBlob across possible threshold combinations. Performance of VADER and TextBlob did not significantly improve from experiment 2 to experiment 3, which applied optimal threshold combinations (VADER: neg. = -0.03, pos. = 0.28; TextBlob: neg. = -0.05, pos. = 0.12). Using suboptimal thresholds only minimally lowered peak performance, as long as they fell within the 'operating window' of that tool. As such, there is potential for guidelines to be developed for choosing thresholds typically within the 'operating window' of a tool. The 'operating window' of effective threshold combinations was larger for VADER than TextBlob. Further accumulation of research in the field will help pinpoint sets of effective thresholds to extract performance.

3.9.3 Summary of Chapter

In this chapter I chose to focus specifically on two lexicon-based sentiment analysis tools. Previous research has found evidence for machine learning tools or LLMs being more effective than lexicon tools. I chose to evaluate the two lexicon tools due to their wide prevalence, especially in the context of a bias in psychology towards dictionary methods. As such, these tools are likely some of the first a researcher would encounter when seeking to find sentiment analysis tools. Additionally, these two tools are open-source and transparent in their mechanism, and their dictionaries can be modified to improve performance in the long-term. However, future research should explore other tools that might be more effective. While these lexicon tools have demonstrated promising performance in analyzing sentiment, they are currently unable to completely substitute for self-reporting of memory valence by participants.

Chapter 4 General Discussion

Music has a distinctive ability to rapidly evoke emotional and vivid autobiographical memories (AMs). However, it is currently unclear what properties of music underlie its cueing capability. In experiment 1, I examined which components of music (prose (lyrics) or instruments) are most effective in facilitating recall. Young adults were presented with short clips of popular songs from their youth as memory cues, which were either unmodified, instrumental versions, lyrics-only, or a visual presentation of the song and artist name. Unmodified song clips, containing both prose and instruments, evoked the most memories, suggesting the combination of these music components promotes cueing. Individuals higher in trait music-related reward sensitivity (indexed using eBMRQ scores) also experienced more memories. Music's ability to induce feelings of reward may therefore contribute to guiding retrieval processes. I also found evidence of temporal alignment between the song's year of popularity, and timeframe of the evoked AMs. Given the powerful cueing effect of music, in experiment 2 I examined whether its prowess would generalize to an older adult population, and whether music could serve to reduce age-related episodic memory deficits. Strikingly, older adults experienced more memories than younger adults, particularly from song cues containing prose, implicating music's potential in countering age-related deficits. I also observed a temporal alignment between the year of popularity of the song cue and the time frame of the evoked memory in both age groups. I propose that songs are particularly effective memory cues because they help set temporal context, constraining the search for AMs to a specific time frame, facilitating access. These findings are particularly relevant for older adults who may have difficulty spontaneously recreating context.

In chapter 3, I investigated the capability of computational approaches using natural language processing to examine emotional sentiment within autobiographical narratives. I compared the congruence in classification of two popular lexicon-based sentiment analysis tools, VADER and TextBlob, with self-reported valence of 3,309 AM narratives from two datasets. VADER consistently showed greater reliability than TextBlob. Accuracy improved significantly when classification required binning valence into three rather than five categories. Results suggest sentiment analysis is beneficial in swiftly determining coarse classifications of valence within AM narratives.

Overall, this thesis shows that music is a powerful cue for eliciting personal memories, even in populations known to experience memory deficits. Music may work by setting the temporal context during retrieval, facilitating access to content. Moreover, this thesis showed that VADER is a sentiment analysis approach that can be used to rapidly determine emotional valence, offering a means of more comprehensively characterizing autobiographical memory narratives.

4.1 The role of cue type

I first aimed to identify which characteristics of music, the prosaic elements (lyrics) or the instrument elements, feed its cueing capacity. To do this, I determined the effectiveness of these musical characteristics in eliciting memories both in isolation and presented together. I presented participants with song clips that were either unmodified, instrumental versions (no lyrics), lyrics-only (no instruments), or a visual presentation of song and artist name (no lyrics or instruments). In experiments 1 and 2, and across younger and older adults, unmodified song cues

were most effective at evoking memories. This consistent finding suggests that musical components work in concert with each other to support recall. The addition of prosaic and instrumental elements separately did not improve music's cueing benefit relative to the song & artist name – only these components presented together enhanced recall. The one exception was the benefit of prose for recall in older adults, which will be discussed in following paragraphs. Harkening back to prior predictions, based on our results, emotion induced by lyrics-alone or instruments-alone were found to be unable to account for a song's cueing effectiveness. Future work could evaluate whether or not combining prose and instruments enhances music's ability to induce emotion. This could clarify the conditions needed to elicit a beneficial pairing of these components in guiding retrieval.

Unmodified song cues were also rated as the most familiar of all cue types in both experiments 1 and 2, emerging as a possible explanatory factor for its augmented effectiveness. To evaluate the cueing benefit of unmodified song clips, separate from familiarity, I conducted a generalized linear mixed effects model containing familiarity, cue type, eBMRQ score and age group predicting probability of successful memory retrieval. While familiarity was the strongest predictor of retrieval, cue type remained a significant explanatory variable. That said, I acknowledge that post-hoc tests revealed unmodified song cues were only significantly better than instruments-only cues at predicting memory retrieval. Future replication is needed to corroborate the unique merit of unmodified cues in facilitating recall, beyond familiarity.

4.2 Music enhances recall in older adults

In this study, I also sought to assess music's ability to facilitate retrieval in individuals known to have age-related memory deficits, as well as determine whether instruments and prose equally contributed to music's cueing mechanism in younger and older adults. Remarkably, there were no observable memory deficits in older adults. In fact, they demonstrated greater recall of autobiographical memories compared to younger adults! Older adults also reported a more positive, arousing memory experience with more re-living relative to younger adults, consistent with the American positivity effect in older adults (Mather & Carstensen, 2005). These results spotlight music's unique ability to facilitate memory access across the lifespan, making a case for music's promise as a memory tool in older adults. Probing the effectiveness of different cue types revealed older adults specifically benefitted from cues containing prose (lyrics-only and unmodified cues). This finding provides support to the idea that the relative preservation of semantic knowledge in older adults (Park et al., 2002) allows semantic information contained in music lyrics to elicit privileged access to memory. Our work suggests that future use of music to facilitate recall in this population should emphasize prosaic elements to optimize cueing effectiveness.

4.3 Music sets temporal context to guide retrieval

In this work, I wished to determine whether the temporal alignment observed by Tanberg et al. (2025), between time frame of an evoked memory and year of song popularity, replicates in older adult samples. In experiment 2 I witnessed this temporal alignment, even for older adults using song cues from over 50 years ago. Specifically, the time frame of an evoked memory

tended to overlap with the same time period in which the song was most popular. I also witnessed a temporal gradient effect in younger adults, such that more recently popular songs evoked more recent memories. Notably, a temporal gradient was only observed in older adult memories cued by unmodified song cues. Remarkably, the difference in time between when a memory was from and when a song cue was most popular was generally small: less than a year difference in younger adults and less than four years difference in older adults. These results support the hypothesis from Tanberg et al. (2025) that music facilitates recall by providing a temporal context to guide memory retrieval.

I interrogated the temporal gradient effect further using linear mixed models to determine if it is preserved at varying levels of cue familiarity and across cue types. I ran two linear mixed models, one for younger and one for older adults, containing year of song popularity, cue familiarity and cue type, predicting year of memory. In younger adults, there was a significant three-way interaction. However, in post-hoc analyses, I observed a persistence in the presence of a temporal gradient for cues of low, average and high familiarity and across all cue types. This finding tentatively supports the idea proposed by Tanberg et al. (2025) that properties of the music, such as genre and musical style, can set this temporal context, regardless of familiarity with the cue. However, an encoding benefit of music can also not be ruled out. In experiment 2, 72% of younger adult memories and 51% of older adult memories included the song as part of the memory. In an encoding-based explanation, songs could be providing privileged access to those memories in which they feature, thereby functioning as a means of reinstating context (Smith et al., 1978; Tulving, 1983), rather than cueing the memory. Such memories are likely to be formed when the song cue was most popular. Peculiarly in older adults, for lower familiarity

cues, more recently popular songs evoked more recent memories (temporal gradient effect), but memories were not from the time period in which songs were most popular (no alignment effect). In contrast, for cues of higher familiarity, memories were from the same time period in which songs were most popular (temporal alignment effect), but no temporal gradient was observed. I speculate that the temporal gradient observed for cues of lower familiarity perhaps represents a retrieval benefit, in which genre and stylistic characteristics of the music guide memory retrieval to separate time periods depending on the recency of the song. In contrast, the temporal alignment observed for cues of higher familiarity could reflect an encoding benefit, in which songs reinstate context for memories they are a part of, which are likely to be from when the song cue was most popular.

4.4 Trait sensitivity to reward from music positively predicts retrieval rate

In this study, I explored whether there are individual differences in music's ability to cue memories. No relationship was found between general demographic characteristics (e.g., sex, years of education) and music's cueing effectiveness, suggesting music has widespread applicability as a cue. However, while musical experience (indexed by years of musical training and perceived musicianship) did not relate to cue effectiveness, trait sensitivity to music-related reward (indexed by eBMRQ scores) positively predicted the number of memories participants experienced. Notably, eBMRQ scores retained explanatory power in predicting retrieval rate even after accounting for the effects of familiarity and cue type. These results suggest music's ability to induce reward may be a contributing factor in its cueing capacity. This reward-related benefit could be the outcome of an induced feeling of reward (and dopamine release) reinstating

a certain motivational state, which then provides privileged access during retrieval for memories with a similar motivational state (Kennedy & Shapiro, 2009).

4.5 VADER can effectively classify sentiment in memory narratives

In chapter 3, I investigated the capacity of natural language processing tools to effectively determine emotional sentiment within autobiographical memory narratives. Specifically, I compared the congruence of two prominent lexicon-based sentiment analysis methods, VADER and TextBlob, with participants' self-reported memory valence scores across 3309 memory narratives from two datasets. I found that VADER, and to a lesser degree TextBlob, as sentiment analysis tools could achieve moderate levels of congruence in valence classification with human self-reported valence (r_{MCC} = .34 and .16 respectively). VADER's congruence in particular is comparable to that of manual coding of emotionality (Rubin & Feeling, 2013), although only in its ability to broadly classify valence. Accuracy improved significantly by limiting classifications to three, rather than five, valence types (negative, neutral or positive), with further but minimal improvements by adjusting the thresholds for inclusion in each valence category. Based on these findings I suggest that sentiment analysis is a promising avenue for classifying valence of human memory narratives, and offers a rapid means of determining the emotional valence contained therein.

4.6 Limitations and future directions

This body of work provides one of the first demonstrations of music's promise in countering age-related memory deficits and enhancing recall in older adults. However, I specifically looked at a cognitively healthy population, and music's cueing capacity in

individuals with cognitive impairment remains unclear. I initially broached the question of whether cognitive status influences music's cueing effectiveness by assessing participants' risk of cognitive failures, using the CFQ 2.0, and correlating test scores with retrieval rate. While no correlation was found, test scores surprisingly suggested that younger adults were more likely to experience cognitive failures than older adults by a large margin. This could possibly reflect a self-presentation bias in reporting (Bradley, 1978), such that older adults were reluctant to endorse memory problems in our study. It is also possible that using a crowdsourcing platform like Prolific biased our older adult sample to include individuals who are uniquely technologically active and well-educated (Turner et al., 2021). It is then possible that this sample may not effectively represent the broader older adult population, limiting the generalizability of these findings .Future work should determine if this effect replicates in a sample not collected from a crowdsourcing platform. Additionally, previous work has implicated music as a uniquely potent cue for individuals with Alzheimer's disease and dementia (Cuddy et al., 2017; El Haj, Fasotti, et al., 2012; El Haj, Postal, et al., 2012; Foster & Valentine, 2001; García et al., 2012; Irish et al., 2006). However, future work should directly compare music's cueing capabilities in younger adults and individuals with Alzheimer's disease to further assess its effectiveness.

Here, I did not find any differences in music's cueing effectiveness by general demographic characteristics, such as sex and years of education. However, future research should investigate whether there are cultural differences in music's cueing abilities. Cultural context plays an important role in constructing autobiographical memories (Wang & Brockmeier, 2002), and music's ability to facilitate context may be subject to cultural differences. Additionally, while the reminiscence bump is largely consistent across cultures

(Conway et al., 2005), cultural differences have been observed in its age range (Janssen et al., 2005), possibly confounding our choice of songs in this study to evoke memories from this time period.

I examined the use of computational approaches to determine emotional sentiment within autobiographical narrations. Another possible computational application not explored here is the use of structural topic modelling to discern and identify content themes within music-evoked memories, and whether there are age differences in the kinds of memories elicited by music. Computational approaches may also be potentially useful in assessing music's ability to preserve memory specificity, which has been reported to decline with age (Levine et al., 2002). Future research should expand on our preliminary investigation of the uses of computational approaches in characterizing autobiographical memories. We additionally limited our evaluation of lexicon-based sentiment analysis tools here to two algorithms, VADER and TextBlob. I am currently collaborating with researchers at Western University to evaluate a broader set of lexicon tools (General Inquirer, Harvard IV-4, QDAP, VADER, and TextBlob) in classifying narratives from both younger and older adults.

Here, I consistently identified a temporal alignment between the year of a song's popularity and the time frame of evoked memories. However, this result could be confounded by our use of songs from participants' reminiscence bump. These songs were chosen from the reminiscence bump to optimize their cueing ability. However, it is possible that the observed temporal alignment here is simply because memories from the reminiscence bump are inherently more likely to be retrieved (Rubin et al., 1998), regardless of the song age. While this is unlikely considering the additional temporal gradient effects observed, which should not be confounded

by the reminiscence bump, follow-up work could determine the replicability of temporal alignment when using songs from outside the reminiscence bump.

4.7 Conclusions

I found that music is a powerful cue to evoke autobiographical memories, and that it preserves and even heightens in its potency across the lifespan. I determined that music must be presented intact, with both prose and instruments, to fully harness its cueing capabilities. I propose that music facilitates recall by creating a temporal context to direct retrieval processes. Music's capacity to induce feelings of reward also appears to contribute to music's unique cueing effectiveness. Additionally, I identified VADER as a promising tool to rapidly identify emotional sentiment in autobiographical memory narratives. Computational approaches such as VADER can provide the opportunity to more feasibly and efficaciously use narrative information to further characterize memory experiences.

References

Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. *Journal of Electrical Systems and Information Technology*, *10*(1), 5. https://doi.org/10.1186/s43067-023-00070-9

Ali, S. O., & Peynircioğlu, Z. F. (2006). Songs and emotions: Are lyrics and melodies equal partners? *Psychology of Music*, *34*(4), 511–534. https://doi.org/10.1177/0305735606067168

Asri, Y., & Fajri, M. (2023). Sentiment analysis of PLN mobile review data using lexicon Vader and Naive Bayes Classification. *2023 International Conference on Networking, Electrical Engineering, Computer Science, and Technology (IConNECT)*, 132–137. https://doi.org/10.1109/IConNECT56593.2023.10327064

Ball, C. N. (1994). Automated text analysis: cautionary tales. *Literary and Linguistic Computing*, *9*(4), 295–302. https://doi.org/10.1093/llc/9.4.295

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, *33*(4), 445–459. https://doi.org/10.1007/s10869-017-9528-3

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, *87*(2), 266–281. https://doi.org/10.1037/0022-3514.87.2.266

Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality*, *40*(1), 35–55. https://doi.org/10.1016/j.jrp.2005.08.006

Belfi, A. M., Karlan, B., & Tranel, D. (2015). Music evokes vivid autobiographical memories. *Memory*, *24*(7), 979–989. https://doi.org/10.1080/09658211.2015.1061012

Belfi, A. M., Karlan, B., & Tranel, D. (2018). Damage to the medial prefrontal cortex impairs music-evoked autobiographical memories. *Psychomusicology: Music, Mind, and Brain*, *28*(4), 201–208. https://doi.org/10.1037/pmu0000222

Benoit, K. (2020). Text as data: An overview. In *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 461–497). SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Berntsen, D. (2009). *Involuntary autobiographical memories: An introduction to the unbidden past* (pp. xi, 228). Cambridge University Press. https://doi.org/10.1017/CBO9780511575921

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11818–11823. https://doi.org/10.1073/pnas.191355898

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), Article S2. https://doi.org/10.51983/ajcst-2019.8.S2.2037

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, *162*, 113746. https://doi.org/10.1016/j.eswa.2020.113746

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Bower, G. H., Gilligan, S. G., & Monteiro, K. P. (1981). Selectivity of learning caused by affective states. *Journal of Experimental Psychology: General*, *110*(4), 451–473. https://doi.org/10.1037/0096-3445.110.4.451

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, *36*(1), 56–71. https://doi.org/10.1037/0022-3514.36.1.56

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Cady, E. T., Harris, R. J., & Knappenberger, J. B. (2007). Using music to cue autobiographical memories of different lifetime periods. *Psychology of Music*, *36*(2), 157–177. https://doi.org/10.1177/0305735607085010

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294–299. https://doi.org/10.1016/s0166-2236(97)01214-9

Cardona, G., Ferreri, L., Lorenzo-Seva, U., Russo, F. A., & Rodriguez-Fornells, A. (2022). The forgotten role of absorption in music reward. *Annals of the New York Academy of Sciences*, *1514*, 142–154. https://doi.org/10.1111/nyas.14790

Cernicky. (2017). *E-Prime* (Version 3.0) [Computer software]. Psychology Software Tools. https://pstnet.com/products/e-prime/

Charland, L. C. (2005). The heat of emotion: Valence and the demarcation problem. *Journal of Consciousness Studies*, *12*(8–10), 82–102.

Chen, P.S. D. (2011). Finding quality responses: The problem of low-quality survey responses and its impact on accountability measures. *Research in Higher Education*, *52*(7), 659–674. https://doi.org/10.1007/s11162-011-9217-4

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Christenson, P. G., de Haan-Rietdijk, S., Roberts, D. F., & ter Bogt, T. F. M. (2019). What has America been singing about? Trends in themes in the U.S. top-40 songs: 1960–2010. *Psychology of Music*, *47*(2), 194–212. https://doi.org/10.1177/0305735617748205

Conway, M. A., & Bekerian, D. A. (1987). Organization in autobiographical memory. *Memory & Cognition*, *15*(2), 119–132. https://doi.org/10.3758/BF03197023

Conway, M. A., Wang, Q., Hanyu, K., & Haque, S. (2005). A cross-cultural investigation of autobiographical memory: On the universality and cultural variation of the reminiscence bump. *Journal of Cross-Cultural Psychology*, *36*(6), 739–749. https://doi.org/10.1177/0022022105280512

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *302*(1110). https://doi.org/10.1098/rstb.1983.0059

Craik, F. I. M., & Rose, N. S. (2012). Memory encoding and aging: A neurocognitive perspective. *Neuroscience and Biobehavioral Reviews*, *36*(7), 1729–1739. https://doi.org/10.1016/j.neubiorev.2011.11.007

Cuddy, L. L., Sikka, R., Silveira, K., Bai, S., & Vanstone, A. (2017). Music-evoked autobiographical memories (MEAMs) in Alzheimer disease: Evidence for a positivity effect. *Cogent Psychology*, *4*(1), 1277578. https://doi.org/10.1080/23311908.2016.1277578

DePaolo, C. A., & Wilkinson, K. (2014). Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. *TechTrends*, *58*(3), 38–44. https://doi.org/10.1007/s11528-014-0750-9

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Dong, Y., Hsiao, Y., Dawson, N., Banerji, N., & Nation, K. (2024). The emotional content of children's writing: A data-driven approach. *Cognitive Science*, *48*(3), e13423. https://doi.org/10.1111/cogs.13423

Eberhardt, S. T., Schaffrath, J., Moggia, D., Schwartz, B., Jaehde, M., Rubel, J. A., Baur, T., André, E., & Lutz, W. (2024). Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy. *Psychotherapy Research*, 1–16. https://doi.org/10.1080/10503307.2024.2322522

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison,

and recommendations. *Psychological Methods*, *26*(4), 398–427. https://doi.org/10.1037/met0000349

El Haj, M., Fasotti, L., & Allain, P. (2012). The involuntary nature of music-evoked autobiographical memories in Alzheimer's disease. *Consciousness and Cognition*, *21*(1), 238–246. https://doi.org/10.1016/j.concog.2011.12.005

El Haj, M., Postal, V., & Allain, P. (2012). Music enhances autobiographical memory in mild Alzheimer's disease. *Educational Gerontology*, *38*(1), 30–41. https://doi.org/10.1080/03601277.2010.515897

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and Vader sentiment. *Proceedings of the International Multiconference of Engineers and Computer Scientists*, *122*.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ford, J. H., Rubin, D. C., & Giovanello, K. S. (2016). The effects of song familiarity and age on phenomenological characteristics and neural recruitment during autobiographical memory retrieval. *Psychomusicology: Music, Mind, and Brain*, *26*(3), 199–210. https://doi.org/10.1037/pmu0000152

Foster, N. A., & Valentine, E. R. (2001). The effect of auditory stimulation on autobiographical recall in dementia. *Experimental Aging Research*, *27*(3), 215–228. https://doi.org/10.1080/036107301300208664

García, J. J. M., Iodice, R., Carro, J., Sánchez, J. A., Palmero, F., & Mateos, A. M. (2012). Improvement of autobiographic memory recovery by means of sad music in Alzheimer's Disease type dementia. *Aging Clinical and Experimental Research*, *24*(3), 227–232. https://doi.org/10.3275/7874

Giorgio, A., Santelli, L., Tomassini, V., Bosnell, R., Smith, S., De Stefano, N., & Johansen-Berg, H. (2010). Age-related changes in grey and white matter structure throughout adulthood. *NeuroImage*, *51*(3), 943–951. https://doi.org/10.1016/j.neuroimage.2010.03.004

Golder, S. A., & Macy, M. W. (2011). Diurnal and season mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881. https://doi.org/10.1126/science.1202775

Goodhew, S. C., & Edwards, M. (2024). The Cognitive Failures Questionnaire 2.0. *Personality and Individual Differences*, *218*, 112472. https://doi.org/10.1016/j.paid.2023.112472

Grall, C., Tamborini, R., Weber, R., & Schmälzle, R. (2021). Stories collectively engage listeners' brains: Enhanced intersubject correlations during reception of personal narratives. *Journal of Communication*, *71*(2), 332–355. https://doi.org/10.1093/joc/jqab004

Guha Majumder, M., Dutta Gupta, S., & Paul, J. (2022). Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. *Journal of Business Research*, *150*, 147–164. https://doi.org/10.1016/j.jbusres.2022.06.012

Henderson, S., & Segal, E. H. (2013). Visualizing qualitative data in evaluation research. *New Directions for Evaluation*, *2013*(139), 53–71. https://doi.org/10.1002/ev.20067

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy* (Version 3.8.4) [Python]. https://doi.org/10.5281/zenodo.1212303 (Original work published 2014)

Husein, K., Goel, L., Anderson, E., Yeung, R., Stastna, M., Fernandes, M. (in preparation). Can computational sentiment analysis be applied to autobiographical memories? Comparing VADER and TextBlob.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14550

Irish, M., Cunningham, C. J., Walsh, J. B., Coakley, D., Lawlor, B. A., Robertson, I. H., & Coen, R. F. (2006). Investigating the enhancing effect of music on autobiographical memory in mild Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *22*(1), 108–120. https://doi.org/10.1159/000093487

Jakubowski, K., Belfi, A. M., & Eerola, T. (2021). Phenomenological differences in music- and television-evoked autobiographical memories. *Music Perception*, *38*(5), 435–455. https://doi.org/10.1525/mp.2021.38.5.435

Jakubowski, K., Belfi, A. M., Kvavilashvili, L., Ely, A., Gill, M., & Herbert, G. (2023). Comparing music- and food-evoked autobiographical memories in young and older adults: A diary study. *British Journal of Psychology*, *114*(3), 580–604. https://doi.org/10.1111/bjop.12639

Jakubowski, K., & Eerola, T. (2022). Music evokes fewer but more positive autobiographical memories than emotionally matched sound and word cues. *Journal of Applied Research in Memory and Cognition*, *11*(2), 272–288. https://doi.org/10.1016/j.jarmac.2021.09.002

Jakubowski, K., & Ghosh, A. (2019). Music-evoked autobiographical memories in everyday life. *Psychology of Music*, *49*(3), 649–666. https://doi.org/10.1177/0305735619888803

Janata, P. (2009). The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex*, *19*(11), 2579–2594. https://doi.org/10.1093/cercor/bhp008

Janata, P., Tomic, S. T., & Rakowski, S. K. (2007). Characterisation of music-evoked autobiographical memories. *Memory*, *15*(8), 845–860. https://doi.org/10.1080/09658210701734593

Janssen, S., Chessa, A., & Murre, J. (2005). The reminiscence bump in autobiographical memory: Effects of age, gender, education, and culture. *Memory*, *13*(6), 658–668. https://doi.org/10.1080/09658210444000322

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *The Behavioral and Brain Sciences*, *31*(5), 559–621. https://doi.org/10.1017/S0140525X08005293

Kaiser, A. P., & Berntsen, D. (2022). The cognitive characteristics of music-evoked autobiographical memories: Evidence from a systematic review of clinical investigations. *WIREs Cognitive Science*, *14*(3), e1627. https://doi.org/10.1002/wcs.1627

Kavitha, K., & Celina, A. (2024). Lexicon-based sentiment analysis-VADER sentiment analysis for food apps. In A. E. Hassanien, S. Anand, A. Jaiswal, & P. Kumar (Eds.), *Innovative Computing and Communications* (pp. 423–434). Springer Nature. https://doi.org/10.1007/978-981-97-3588-4_35

Kennedy, P. J., & Shapiro, M. L. (2009). Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10805–10810. https://doi.org/10.1073/pnas.0903259106

Krans, J., de Bree, J., & Moulds, M. L. (2015). Involuntary cognitions in everyday life: Exploration of type, quality, content, and function. *Frontiers in Psychiatry*, *6*. https://doi.org/10.3389/fpsyt.2015.00007

Krumhansl, C. L., & Cuddy, L. L. (2010). A theory of tonal hierarchies in music. In *Music perception* (pp. 51–87). Springer Science + Business Media. https://doi.org/10.1007/978-1-4419-6114-3_3

Lesiuk, T., Dillon, K., Ripani, G., Iliadis, I., Perez, G., Levin, B., Sun, X., & McIntosh, R. (2025). Fractional amplitude of low-frequency fluctuations during music-evoked autobiographical memories in neurotypical older adults. *Frontiers in Neuroscience*, *18*. https://doi.org/10.3389/fnins.2024.1479150

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*(4), 677–689. https://doi.org/10.1037/0882-7974.17.4.677

Light, L. L. (1991). Memory and aging: Four hypotheses in search of data. *Annual Review of Psychology*, *42*, 333–376. https://doi.org/10.1146/annurev.ps.42.020191.002001

López-Cano, M. A., Navarro, B., Nieto, M., Andrés-Pretel, F., & Latorre, J. M. (2020). Autobiographical emotional induction in older people through popular songs: Effect of reminiscence bump and enculturation. *PLOS ONE*, *15*(9), e0238434. https://doi.org/10.1371/journal.pone.0238434

Loria, S. (2024). *TextBlob* (Version 0.18.0) [Python]. https://textblob.readthedocs.io/en/dev/authors.html

Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E., & Pereira, F. (2024). A comparison of ChatGPT and fine-tuned Open Pre-Trained Transformers (OPT) against

widely used sentiment analysis tools: Sentiment analysis of COVID-19 survey data. *JMIR Mental Health*, *11*, e50150. https://doi.org/10.2196/50150

Martínez-Saez, M. C., Ros, L., López-Cano, M., Nieto, M., Navarro, B., & Latorre, J. M. (2024). Effect of popular songs from the reminiscence bump as autobiographical memory cues in aging: A preliminary study using EEG. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1300751

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception*, *31*(2), 118–138. https://doi.org/10.1525/mp.2013.31.2.118

Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences*, *9*(10), 496–502. https://doi.org/10.1016/j.tics.2005.08.005

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: USA 1960-2010. *Royal Society Open Science*, *2*(5), 150081. https://doi.org/10.1098/rsos.150081

Mazzoni, D., & Dannenberg, R. (2023). *Audacity* (Version 3.4.2) [Computer software]. https://www.audacityteam.org/

Menon, V., & Levitin, D. J. (2005). The rewards of music listening: Response and physiological connectivity of the mesolimbic system. *NeuroImage*, *28*(1), 175–184. https://doi.org/10.1016/j.neuroimage.2005.05.053

Min, W. N. S. W., & Zulkarnain, N. Z. (2020). Comparative evaluation of lexicons in performing sentiment analysis. *Journal of Advanced Computing Technology and Application (JACTA)*, *2*(1), Article 1.

Moghadam, M., Lehman, T., & Zechory, I. (2009, October 20). *Genius*. Genius. https://genius.com/

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The thematic apperception test. *Archives of Neurology & Psychiatry*, *34*, 289–306. https://doi.org/10.1001/archneurpsyc.1935.02250200049005

Mueller, A. C. (2023). *Wordcloud* (Version 1.9.3) [Python]. https://github.com/amueller/wordcloud (Original work published 2012)

Nariman, D. (2024). Sentiment analysis of hotel reviews using lexicon-based methods: A comparative study of VADER and TextBlob. In L. Barolli (Ed.), *Advances on Broad-Band Wireless Computing, Communication and Applications* (pp. 263–274). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76452-3_25

Newby, J. M., & Moulds, M. L. (2012). A comparison of the content, themes, and features of intrusive memories and rumination in major depressive disorder. *British Journal of Clinical Psychology*, *51*(2), 197–205. https://doi.org/10.1111/j.2044-8260.2011.02020.x

Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings* [Doctoral dissertation, Ohio State University]. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear=10&p10_accession_num=osu1161705351

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571299

Paltoglou, G. (2014). Sentiment analysis in social media. In N. Agarwal, M. Lim, & R. T. Wigand (Eds.), *Online Collective Action: Dynamics of the Crowd in Social Media* (pp. 3–17). Springer. https://doi.org/10.1007/978-3-7091-1340-0_1

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299–320. https://doi.org/10.1037/0882-7974.17.2.299

Peck, L. S. L., & Grealey, P. (2020). Autobiographical significance of meaningful musical experiences: Reflections on Youth and Identity. *Music & Science*, *3*. https://doi.org/10.1177/2059204320974221

Perlmutter, M. (1979). Age differences in adults' free recall, cued recall, and recognition. *Journal of Gerontology*, *34*(4), 533–539. https://doi.org/10.1093/geronj/34.4.533

Platz, F., Kopiez, R., Hasselhorn, J., & Wolf, A. (2015). The impact of song-specific age and affective qualities of popular songs on music-evoked autobiographical memories (MEAMs). *Musicae Scientiae*, *19*(4), 327–349. https://doi.org/10.1177/1029864915597567

Porter, M. F. (2001, October). *Snowball: A language for stemming algorithms—Snowball*. Snowball. https://snowballstem.org/texts/introduction.html

Prolific (Version October 2025). (2014). [Computer software].

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01065

Putter, K. C., Krause, A. E., & North, A. C. (2022). Popular music lyrics and the COVID-19 pandemic. *Psychology of Music*, *50*(4), 1280–1295. https://doi.org/10.1177/03057356211045114

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Raven, J. (1983). The progressive matrices and Mill Hill Vocabulary Scale in Western societies. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 107–114). Springer US. https://doi.org/10.1007/978-1-4899-2151-2_8

Razzaq, A., Abbas, T., Hashim, S., Qadri, S., Mumtaz, I., Saher, N., Ul-Rehman, M., Shahzad, F., & Nawaz, S. A. (2022). Extraction of psychological effects of COVID-19 pandemic through topic-level sentiment dynamics. *Complexity*, *2022*(1), 9914224. https://doi.org/10.1155/2022/9914224

Rubin, D. C., & Feeling, N. (2013). Measuring the Severity of Negative and Traumatic Events. *Clinical Psychological Science*, *1*(4), 375–389. https://doi.org/10.1177/2167702613483112

Rubin, D. C., Wetzler, S. E., & Nebes, R. D. (1986). Autobiographical memory across the lifespan. In *Autobiographical memory* (pp. 202–221). Cambridge University Press. https://doi.org/10.1017/CBO9780511558313.018

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Salakka, I., Pitkäniemi, A., Pentikäinen, E., Mikkonen, K., Saari, P., Toiviainen, P., & Särkämö, T. (2021). What makes music memorable? Relationships between acoustic musical features and music-evoked emotions and memories in older adults. *PLoS ONE*, *16*(5), e0251692. https://doi.org/10.1371/journal.pone.0251692

Salakka, I., Pitkäniemi, A., Pentikäinen, E., Saari, P., Toiviainen, P., & Särkämö, T. (2024). Emotional and musical factors combined with song-specific age predict the subjective autobiographical saliency of music in older adults. *Psychology of Music*, *52*(3), 305–321. https://doi.org/10.1177/03057356231186961

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging* (pp. xi, 438). Lawrence Erlbaum Associates, Inc.

Schlagman, S., Schulz, J., & Kvavilashvili, L. (2006). A content analysis of involuntary autobiographical memories: Examining the positivity effect in old age. *Memory*, *14*(2), 161–175. https://doi.org/10.1080/09658210544000024

Schulkind, M. D., Hennis, L. K., & Rubin, D. C. (1999). Music, emotion, and autobiographical memory: They're playing your song. *Memory & Cognition*, *27*(6), 948–955. https://doi.org/10.3758/bf03201225

Shum, M. S. (1998). The role of temporal landmarks in autobiographical memory processes. *Psychological Bulletin*, *124*(3), 423–442. https://doi.org/10.1037/0033-2909.124.3.423

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, *6*(4), 342–353. https://doi.org/10.3758/BF03197465

Sousou, S. D. (1997). Effects of melody and lyrics on mood and memory. *Perceptual and Motor Skills*, *85*(1), 31–40. https://doi.org/10.2466/pms.1997.85.1.31

Sreekumar, V., Nielson, D. M., Smith, T. A., Dennis, S. J., & Sederberg, P. B. (2018). The experience of vivid autobiographical reminiscence is supported by subjective content representations in the precuneus. *Scientific Reports*, *8*(1), 14899. https://doi.org/10.1038/s41598-018-32879-0

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484–498. https://doi.org/10.1002/bs.3830070412

Stratton, V. N., & Zalanowski, A. H. (1994). Affective impact of music vs. lyrics. *Empirical Studies of the Arts*, *12*(2), 173–184. https://doi.org/10.2190/35T0-U4DT-N09Q-LQHW

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods*, *53*(5), 2069–2082. https://doi.org/10.3758/s13428-020-01531-z

Tanberg, P., Yeung, R. C., & Fernandes, M. A. (2025). Evidence of temporal and emotional alignment between song cues and their evoked autobiographical memories. *Memory & Cognition*. https://doi.org/10.3758/s13421-025-01717-w

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tesoriero, M., & Rickard, N. S. (2012). Music-enhanced recall: An effect of mood congruence, emotion arousal or emotion function? *Musicae Scientiae*, *16*(3), 340–356. https://doi.org/10.1177/1029864912459046

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.

Turner, A. M., Engelsma, T., Taylor, J. O., Sharma, R. K., & Demiris, G. (2021). Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic. *AMIA Annual Symposium Proceedings*, *2020*, 1230–1238.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Viviani, M., Crocamo, C., Mazzola, M., Bartoli, F., Carrà, G., & Pasi, G. (2021). Assessing vulnerability to psychological distress during the COVID-19 pandemic through the analysis of microblogging content. *Future Generation Computer Systems*, *125*, 446–459. https://doi.org/10.1016/j.future.2021.06.044

Vocali.se. (2025). [Computer software]. https://vocali.se/en

Wang, Q., & Brockmeier, J. (2002). Autobiographical remembering as cultural practice: Understanding the interplay between memory, self and culture. *Culture & Psychology*, *8*(1), 45–64. https://doi.org/10.1177/1354067X02008001618

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "big data" for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 587–592. https://doi.org/10.1109/SocialCom-PASSAT.2012.119

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R., & Halpern, S. D. (2019). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical

illness. *Journal of Biomedical Informatics*, *89*, 114–121. https://doi.org/10.1016/j.jbi.2018.12.001

Weitzman, C. (2017). *Speechify:* [Computer software]. https://speechify.com/

Willander, J., Sikström, S., & Karlsson, K. (2015). Multimodal retrieval of autobiographical memories: Sensory information contributes differently to the recollection of events. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01681

Wittmann, B. C., Schott, B. H., Guderian, S., Frey, J. U., Heinze, H.-J., & Düzel, E. (2005). Reward-related FMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron*, *45*(3), 459–467. https://doi.org/10.1016/j.neuron.2005.01.010

Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2012). Reward modulation of hippocampal subfield activation during successful associative encoding and retrieval. *Journal of Cognitive Neuroscience*, *24*(7), 1532–1547. https://doi.org/10.1162/jocn_a_00237

Yeganeh, L., Boyle, J. A., Gibson-Helm, M., Teede, H., & Vincent, A. J. (2020). Women's perspectives of early menopause: Development of a word cloud. *Climacteric*, *23*(4), 417–420. https://doi.org/10.1080/13697137.2020.1730318

Yeung, R. C., Danckert, J., van Tilburg, W. A. P., & Fernandes, M. A. (2024). Disentangling boredom from depression using the phenomenology and content of involuntary autobiographical memories. *Scientific Reports*, *14*(1), 2106. https://doi.org/10.1038/s41598-024-52495-5

Yeung, R. C., & Fernandes, M. A. (2020). Recurrent involuntary autobiographical memories: Characteristics and links to mental health status. *Memory*, *28*(6), 753–765. https://doi.org/10.1080/09658211.2020.1777312

Yeung, R. C., & Fernandes, M. A. (2021). Recurrent involuntary memories are modulated by age and linked to mental health. *Psychology and Aging*, *36*(7), 883–890. https://doi.org/10.1037/pag0000630

Yeung, R. C., Stastna, M., & Fernandes, M. A. (2022). Understanding autobiographical memory content using computational text analysis. *Memory*, *30*(10), 1267–1287. https://doi.org/10.1080/09658211.2022.2104317

Zator, K., & Katz, A. N. (2016). The language used in describing autobiographical memories prompted by life period visually presented verbal cues, event-specific visually presented verbal cues and short musical clips of popular music. *Memory*, *25*(6), 831–844. https://doi.org/10.1080/09658211.2016.1224353

Zentner, M., & Eerola, T. (2010). Self-report measures and models. In *Handbook of Music and Emotion: Theory, Research, Applications*. OUP Oxford.

Appendix A

Table A1: Popular Song Stimuli for Young Adults, by Song Name, Artist(s), and Release Year used in experiment 1 and 2

Song Name	Artist(s)	Year on Top Charts
Without Me	Halsey	2019
Bad Guy	Billie Eilish	2019
7 Rings	Ariana Grande	2019
Sunflower	Post Malone and Swae Lee	2019
Old Town Road	Lil Nas X	2019
Happier	Marshmello and Bastille	2019
Don't Start Now	Dua Lipa	2020
The Bones	Maren Morris	2020
Blinding Lights	The Weeknd	2020
The Box	Roddy Ricch	2020
Adore You	Harry Styles	2020
Memories	Maroon 5	2020
Good 4 U	Olivia Rodrigo	2021
Kiss Me More	Doja Cat and SZA	2021
Levitating	Dua Lipa	2021
Drivers License	Olivia Rodrigo	2021
Save Your Tears	Ariana Grande and The Weeknd	2021
Mood	24kGoldn	2021
Cold Heart	Dua Lipa and Elton John	2022
Easy on Me	Adele	2022
As it Was	Harry Styles	2022
Heat Waves	Glass Animals	2022
Stay	The Kid LAROI and Justin Bieber	2022
Shivers	Ed Sheeran	2022

Table A2: Popular Song Stimuli for Older Adults, by Song Name, Artist(s), and Release Year used in experiment 2

Song Name	Artist(s)	Year on Top Charts
These Boots are Made for Walking	Nancy Sinatra	1966
You Can't Hurry Love	The Supremes	1966
96 Tears	Question Mark & the Mysterians	1966
What Becomes of the Brokenhearted	Jimmy Ruffin	1966
Last Train to Clarkesville	The Monkees	1966
Reach Out I'll Be There	Four Tops	1966
To Sir with Love	Lulu	1967
Ode to Billie Joe	Bobbie Gentry	1967
Respect	Aretha Franklin	1967
The Letter	The Box Tops	1967
Windy	The Association	1967
I'm a Believer	The Monkees	1967
Harper Valley PTA	Jeannie C. Riley	1968
Hey Jude	The Beatles	1968
Honey	Bobby Goldsboro	1968
(Sittin' On) The Dock of the Bay	Otis Redding	1968
People Got to Be Free	The Rascals	1968
Sunshine of Your Love	Cream	1968
I Can't Get Next to You	The Temptations	1969
Honky Tonk Women	The Rolling Stones	1969
I'll Never Fall in Love Again	Tom Jones	1969
Build Me Up Buttercup	The Foundations	1969
Crimson and Clover	Tommy James and The Shondelles	1969
One	Three Dog Night	1969

Appendix B

Figure B1: Sample of memory narratives experienced by younger and older adults from experiment 2

Younger adults	Older adults
I remember listening to the song um when I was working on an essay for English class in my grade 12 English. And I was using this song as an example in my essay.	My memory is of uh watching the Ed Sullivan show in my parents uh library, the house I grew up in um and Nancy Sinatra was uh on the Ed Sullivan Show and she performed this song.
So, since the recording was about toasting, I remembered, um, going out to drink with my friends and it was a late night and I were all at this, um, restaurant just drinking for a long time until I were all basically sick and had to go outside.	Mm. Being in high school, going to school dances, hanging out with my friends, meeting girls.
The song was super popular a while back. I do remember this one day my friend and I went shopping, I went to Zara, they had the song playing on the radio. I went to H&M, they had this song playing on the radio. I got in the car, put on the radio and the song was playing on the radio. So I heard this song three times back to back um in one day and it was just super funny.	I was spending time with uh, a young man named George who uh reminded me of Otis Redding because he loved him so much. And I were sitting, uh listening to Otis and uh talking about the concert going on. Um, at school, I didn't have tickets for, but I decided to go and break to, well, not break in but walk in and crash the concert, which turned out to be the who and I saw them from the side of the stage.
I actually heard this song on my last drive to, to my college a week ago. Uh I was channel surfing all the Sirius XM channels and I happened to catch this song and it sounded very familiar to me and very catchy and it gave me the best mood I can be to, to succeed in my, in my class that night.	I remember that song, um, and all of the Supreme songs when I was young. Uh, a few girls, uh, I would get together and I would act it out, sing it out. Uh, and also, well, anyway, I had other memories but that's, that's the first one that came to mind.