A Lifespan Study of the Confidence-Accuracy Relation in Working Memory and Episodic Long-Term Memory

Nathaniel R. Greene[1,2], Alicia Forsberg[3], Dominic Guitard[4], Moshe Naveh-Benjamin[1], & Nelson Cowan[1]

[1]Department of Psychological Sciences, University of Missouri

[2]Department of Psychology, University of Pennsylvania

[2]Department of Psychology, University of Sheffield

[3]School of Psychology, Cardiff University

Author Note

Correspondence concerning this article should be addressed to Nathaniel R. Greene, Department of Psychological Sciences, University of Missouri, 9J McAlester Hall, Columbia, MO 65211, or Department of Psychology, University of Pennsylvania, 425 S. University Avenue, Levin Building, Room 201, Philadelphia, PA 19104. Email: ngreene@mail.missouri.edu or nrgreene@sas.upenn.edu

This research was supported by a National Institute of Health (NIH) research grant R01 HD-21338 to N. C. and by a Jacob's Foundation grant to A. F. The ideas discussed in this paper have not been disseminated previously. Data and analysis scripts are available at: https://osf.io/vfe9g/

Abstract

The relation between an individual's memory accuracy and reported confidence in their memories can indicate self-awareness of memory strengths and weaknesses. We provide a lifespan perspective on this confidence-accuracy relation, based on two previously published experiments with 320 participants, including children aged 6 to 13, young adults aged 18 to 27, and older adults aged 65 to 77, across tests of working memory (WM) and long-term memory (LTM). Participants studied visual items in arrays of varying set sizes and completed item recognition tests featuring six-point confidence ratings either immediately after studying each array (WM tests) or following a long period of study events (LTM tests). Confidence-accuracy characteristic analyses showed that accuracy improved with increasing confidence for all age groups and in both WM and LTM tests. These findings reflect a universal ability across the lifespan to use awareness of the strengths and limitations of one's memories to adjust reported confidence. Despite this age invariance in the confidence-accuracy relation, however, young children were more prone to high-confidence memory errors than other groups in tests of WM, whereas older adults were more susceptible to high-confidence false alarms in tests of LTM. Thus, although participants of all ages can assess when their memories are weaker or stronger, individuals with generally weaker memories are less adept at this confidence-accuracy calibration. Findings also speak to potential different sources of high-confidence memory errors for young children and older adults, relative to young adults.

Keywords. Lifespan, working memory, long-term memory, metamemory, confidence, childhood development, adult aging

Public Significance Statement

Our ability to accurately evaluate the strengths and weaknesses of our memories is critical to how we make decisions and can influence how much confidence other people can place in our reported experiences, as in eyewitness testimony. The present study reveals a strikingly universal ability of individuals from young childhood to older adulthood in calibrating their confidence in their memory with their observed accuracy in recognition. However, young children and older adults are more susceptible to high-confidence memory errors, but during different phases of memory retention, with young children's high-confidence errors arising in working memory and those of older adults appearing mostly in long-term memory. These findings suggest that different mechanisms underlie why young children and older adults generally have poorer memory than adolescents and young adults.

A Lifespan Study of the Confidence-Accuracy Relation in Working Memory and Episodic Long-Term Memory

Metamemory refers to our awareness of the strengths and limitations of our memories, and this awareness can guide how we make decisions (Dunlosky & Tauber, 2016; Flavell, 1971; Metcalfe & Dunlosky, 2008; Nelson & Narens, 1990), such as deciding not to take a pill because you are confident that you remember doing so an hour ago. Understanding how capable individuals are of gauging the strengths and limitations of their memories can have profound implications for how much certainty we can place in their reported experiences. For example, understanding whether an eyewitness has a keen awareness of the reliability of their memories can inform how much confidence a jury should place in their testimony when they claim to be very certain to have seen something (e.g., Mickes, 2015; Wixted et al., 2015).

In the present study, we concentrate on three key components of the *memory-metamemory relation* that previously have not been considered together. The components are: (1) metamemory for information held in *working memory* (WM), the temporary store of information useful for ongoing cognitive operations (Cowan, 2017); (2) metamemory for information held in *episodic long-term memory* (LTM), memories for past events occurring in specific times and places (Tulving, 1983); and (3) lifespan changes, from young childhood (beginning at age 6) to older adulthood (age 65-80), in metamemory for both WM and LTM. We provide a synthesis across these content areas, using recent advances (Mickes, 2015; cf., Fleming & Lau, 2014) in estimating the relation between individuals' observed memory performance and their subjective retrospective confidence in their memory decisions (i.e., the *confidence-accuracy relation*), applied to lifespan data (Forsberg et al., 2022a, 2022b) from a novel procedure (Forsberg et al., 2021b) that provides independent tests of WM and LTM in a shared paradigm.

By independent tests, we mean that this paradigm separately measures memory for information that can still be held within the capacity constraints of WM and memory for information that must be retrieved instead from LTM, a point we return to in the ensuing section.

In what follows, we first clarify our definitions of WM and LTM and how they are related, and we also emphasize why it is important to measure the confidence-accuracy relation. We then distinguish how our aims diverge from previous studies and describe our analytical method for measuring the confidence-accuracy relation in tests of WM and LTM in a common way. Next, we review previous research on the confidence-accuracy relation in WM and in LTM across young adulthood, childhood development, and adult aging. We end the introduction by detailing our specific hypotheses both for whether, within an age group, individuals can adjust their confidence with their memory signals and whether, across age groups, individuals are equally capable of evaluating the strengths and weaknesses of their memories relative to those of other groups.

Terminology Used in the Present Study

First, our definitions of WM and LTM are not intended to imply that these are separate memory systems. According to the embedded processes model (Cowan, 1988, 2019; Cowan et al., 2024), WM is an activated subset of LTM in which new LTM representations can be formed (cf., Oberauer, 2002). Due to its limited capacity, WM acts as an encoding bottleneck, constraining how much information individuals can later access in LTM (Forsberg et al., 2021; Fukuda & Vogel, 2019; cf. Atkinson & Shiffrin, 1968). Thus, our emphasis on "independent" tests of WM and LTM is intended to distinguish memory tests that occur during different periods following encoding. WM tests refer to periods where information can still be active within WM without having been "transferred" or consolidated into LTM, whereas LTM tests refer to periods

where information may be retrieved back into WM but must have been, at some point, transferred into LTM due to severe constraints on the capacity to maintain information in WM (Cowan, 2001). Some may prefer the distinction of short-term versus long-term memory, but in the present usage, WM implies a more active form of temporary representation (i.e., it can involve not only storage but also processing of currently active memoranda) than short-term memory, which implies passive storage of recently acquired information (Cowan, 2017).

Second, our focus on the relation between retrospective confidence and memory accuracy means that the type of metamemory ability under investigation here is one that Nelson and Naren (1990) have classified as a retrieval monitoring process, as opposed to other forms of metamemory, like judgments of learning that occur during memory encoding or maintenance (e.g., Son & Metcalfe, 2005). Providing a common metric for studying the confidence-accuracy relation across the lifespan, as we intend to do, is important given a disconnect between empirical studies that suggest a strong confidence-accuracy relation (e.g., Winsor et al., 2021; Wixted et al., 2018; Wixted & Mickes, 2022) and long-standing views that certain individuals, such as young children or older adults with generally poorer memory capabilities, are unreliable witnesses due to their ineffective ability to evaluate their memory strengths and weaknesses (e.g., Keast et al., 2007; Newcombe & Bransgrove, 2007; Powell et al., 2013). Such views have long dominated in the courtroom, perhaps in part due to limitations in how the confidence-accuracy relation has traditionally been measured (Winsor et al., 2021).

Goals of the Present Study and Connections to Previous Studies

There are two key goals of our approach. First, we aim to understand whether individuals within a given age group are aware of their memory strengths and limitations in both WM and LTM testing situations. This would be reflected by a confidence-accuracy relation in which

individuals are, on average, more accurate when they express higher compared to lower confidence in their recognition responses. Second, we intend to measure whether, across age groups, individuals are equally capable of evaluating the strengths or weaknesses of their memories, relative to those of other groups, in both WM and LTM tests. This would be reflected both in consistent, positive confidence-accuracy relations across age groups and, critically, by an absence of age-related differences in recognition accuracy at highest confidence levels, even as such differences may be present at lower confidence levels. Regarding this second criterion, if individuals with weaker memories (e.g., young children or older adults) are aware that their memories are more impoverished, relative to those of young adults for instance, and take this impoverishment into account, they would rarely express *high* confidence. That is, they would "downregulate" their retrospective confidence ratings, only expressing *high* confidence in situations where they are likely to be accurate. As a result of this downregulating process, individuals with poorer memories but good subjective awareness of these limitations would likely be more error prone at lower confidence levels (both because they rate their confidence as low most of the time and because they commit more recognition errors) but not at higher confidence levels. Thus, age differences at *low* confidence levels would not tell us much about whether individuals of different age groups are equally capable of adjusting their confidence with their observed memory abilities. However, age differences at *high* confidence levels would suggest that some groups are less adept at this confidence-accuracy calibration, overall.

Although numerous studies have examined the confidence-accuracy relation separately for WM and LTM, no previous studies have examined this relation in a common paradigm that includes tests of both WM and LTM. Doing so is important given a growing body of evidence showing that initial WM limitations constrain how much information individuals, from young

children to older adults, can later access from LTM (Forsberg et al., 2021b, 2022a, 2022b, 2023; cf., Fukuda & Vogel, 2019).

Moreover, there have been relatively few efforts to measure lifespan developmental changes, from childhood to older adulthood, in the confidence-accuracy relation (Fandakova et al., 2013; Shing et al., 2009)[1], and none which have measured this relation across the lifespan in both WM and LTM. Our lifespan approach is especially useful because, typically, multiple changes across age groups are confounded. Memory improves across childhood (Courage & Cowan, 2022; Shing et al., 2010) as the result of increases in knowledge and improvements in both storage and processing. In contrast, memory declines in old age (Light, 1991; Naveh-Benjamin & Cowan, 2023; Naveh-Benjamin & Old, 2008; Zacks et al., 2000) as the result of declines in storage and processing, despite the knowledge that has accrued during a lifetime. Noting differences between childhood development and adult aging therefore provides a perspective that can shed light on the mechanisms that account for changes across the lifespan.

On Measuring the Confidence-Accuracy Relation

Numerous methods have been used to measure the relation between participants' retrospective confidence in their memory and their observed accuracy in recall or recognition. This diversity of methods is particularly notable when comparing how the confidence-accuracy relation has been measured in studies of WM compared with studies of LTM. In the present study, we provide a unified approach for both WM and LTM, based on confidence-accuracy-characteristic (CAC) analyses (Mickes, 2015) that overcome the many shortcomings of calculating the correlation between confidence and accuracy (for extensive critiques of the

[1] Hiller and Weber (2013) also measured the confidence-accuracy relation in a LTM associative recognition paradigm among children and adults, though their study did not include participants over the age of 60 where memory declines are more pronounced.

traditional correlational approach, see Busey et al., 2000; Juslin et al., 1996; Mickes, 2015; Winsor et al., 2021; Wixted & Wells, 2017). Traditional correlational approaches provide no information about whether individuals over- or underestimate their memory performance (Juslin et al., 1996) and are subject to both within- and between-subject variations in confidence or accuracy (Busey et al., 2000). CAC analysis overcomes these limitations by decomposing each participant's accuracy at different confidence levels into the proportions of correct and erroneous responses.

CAC analysis measures whether individuals are capable of gauging how reliable their memories are, based on how confident they are when claiming that something is "old" or "new." Person A with good subjective awareness of their memory strengths or limitations will be highly accurate in classifying an item as "old" or "new" when they express *high* confidence in their classification. That is, Person A would express *high* confidence recognition responses only when they are certain that their memory is accurate. In contrast, Person B with poor subjective awareness of their memory strengths or limitations would be more prone to inaccuracies when expressing *high* confidence recognition responses. If Person B forgets an old item that is shown again as a test probe, they may be highly confident that the test probe is new because they believe that they would have remembered the item if they saw it previously. That is, Person B would be unaware of a limitation in their memory and would thus be a poorer judge of their memory limitations than Person A, even if Person A has worse overall memory than Person B. We turn now to considering how this type of confidence-accuracy analysis has been applied in studies with young adults, children, and older adults.

Young Adults' Confidence-Accuracy Relation in WM and LTM

We first consider prior research on the confidence-accuracy relation in young adulthood (usually studied between ages 18 to ~30 years), as this period of the lifespan is typically associated with more capable memory compared to childhood and older adulthood. Young adults can hold more information in limited capacity WM compared than young children (Cowan et al., 2005, 2006, 2018; Riggs et al., 2006; Simmering, 2012) and older adults (Brockmole & Logie, 2013; Gilchrist et al., 2008; Greene et al., 2020; Light & Anderson, 1985; Naveh-Benjamin & Cowan, 2023; Salthouse & Babcock, 1991; Wingfield et al., 1988). Likewise, in tests of LTM, young adults tend to be more accurate than young children (Fitzgerald & Price, 2015; Lindsay et al., 1991; Shing et al., 2010)[2] and older adults (Fraundorf et al., 2019; Old & Naveh-Benjamin, 2008; Rhodes et al., 2019).

There has been much more focus on young adults' confidence-accuracy relation in tests of LTM than in tests of WM. Much of this LTM research has come from the eyewitness memory literature, primarily from studies using suspect identification line-ups or face recognition procedures (Busey et al., 2000; Dodson & Dobolyi, 2016; Lindsay et al., 1998; Mickes, 2015; Palmer et al., 2013; but for studies using more traditional verbal or visual stimuli commonly encountered in studies of memory, see Mickes et al., 2011; Tekin & Roediger, 2017). Early studies suggested that young adults' confidence poorly tracked their accuracy in LTM tests (Bothwell et al., 1987; Lindsay et al., 1981), but those studies relied on traditional correlational methods that can over- or underestimate this relation (Busey et al., 2000; Juslin et al., 1996). Using calibration-based methods and more recently CAC analyses, the prevailing evidence suggests that young adults' retrospective confidence is strongly and positively associated with

[2] However, there are some surprising instances of "developmental reversals" in LTM, in which young children, with less acquired knowledge, are less sensitive to semantic false memories than adolescents, young adults, and older adults (Brainerd & Reyna, 2015; Brainerd et al., 2002).

their recognition accuracy in tests of LTM (Busey et al., 2000; Mickes, 2015; Wixted & Wells, 2017). However, this positive confidence-accuracy relation is mediated by individual differences among young adults (e.g., differences in face recognition abilities; Grabman et al., 2019).

There is more limited research on young adults' metamemory for the contents of their WM, but a few recent studies have produced some insightful results. Using a modified color reproduction task (e.g., Wilken & Ma, 2004), Suchow et al. (2017) showed that when young adults were allowed to reproduce the color of an item from a just-presented visual array that they remembered best, their accuracy in color reproduction was superior (92% correct) compared to the standard situation in which they were tasked with reproducing the color of a randomly selected item from the array (71% correct). These findings suggest that young adults can evaluate the strengths of different memory representations for items in WM. Studies of WM using retrospective confidence ratings have shown that, as in LTM, young adults' confidence tracks their accuracy on a trial-by-trial basis (Adam & Vogel, 2017; Bona & Silvanto, 2014; Vandenbroucke et al., 2014), though these studies have not relied upon CAC analyses recommended in the LTM literature (Mickes, 2015). These studies have also occasionally documented important limitations in young adults' abilities to evaluate the contents of their WM. For example, young adults sometimes express overconfidence in the number of items that they can effectively remember from a just-presented study set (i.e., they express overconfidence in their WM capacity; Adam & Vogel, 2017; Cowan et al., 2016; Forsberg et al., 2021a). This overestimation of one's WM capacity tends to occur most often on trials where attention lapsed (Adam & Vogel, 2017) and can lead young adults to erroneously assume that no change in a display occurred on trials where changes actually occurred (Cowan et al., 2016). Nevertheless, on average, young adults' accuracy in WM and LTM recognition improves with increases in

their retrospective confidence. To date, no study has employed a common metric (e.g., CAC analyses) for comparing young adults' confidence-accuracy relation on tests of WM and LTM, though the present study intends to fill this gap.

Development of the Confidence-Accuracy Relation in WM and LTM

Much of the focus on the development, across childhood to young adulthood, of metamemory has been limited to studies of LTM. These studies have typically documented improvements in the LTM confidence-accuracy relation with childhood development (Schneider, 1985; Flavell et al., 1993; Schneider & Pressley, 2013). A long-standing view is that young children (from about age 3 to 8, and occasionally also reported for children up to age 10) overestimate how much information they can remember, but this overconfidence in one's LTM declines after about age 10 (e.g., Brewer & Day, 2005; Ghetti et al., 2008; Howie & Roebers, 2007; Pressley et al., 1987). However, children up to age 12 also sometimes express overconfidence in their LTM accuracy, resulting in *high-confidence* memory errors, when they are given misleading questions about an event or feedback about their performance (Allwood et al., 2005; Howie & Roebers, 2007; Roebers, 2002; Roebers & Howie, 2003).

This traditional view that young children are poor judges of their memories (e.g., Keast et al., 2007; Knutsson & Allwood, 2014; Powell et al., 2013) has led legal experts to question the suitability of children as eyewitnesses. However, many of these studies relied on limited correlational methods for measuring the confidence-accuracy relation. Using CAC analysis, Winsor et al. (2021) documented a consistently positive confidence-accuracy relation among children and adolescents aged 4 to 17, though children younger than 7 were somewhat more prone to *high-confidence* recognition errors when making explicit confidence judgments, as opposed to more implicit measures of confidence (e.g., shrugging when they were uncertain).

These findings echo those of other studies which have relied on more implicit measures of assessing children's confidence in their memories. For example, children as young as 3 are often highly accurate in their LTM recognition responses when they are able to choose whether to respond on a given trial (Balcomb & Gerken, 2008; cf., Koriat et al., 2001; Liu et al., 2018). Such findings suggest that even very young children can monitor the strengths of their LTM representations under certain conditions.

There has been much more limited research on the development of metamemory for WM, despite the important role of WM in childhood development (Cowan, 2016), due to its relation to academic success (Gathercole et al., 2004) and knowledge acquisition (Cowan & Alloway, 2008). An early study by Flavell et al. (1970) found that children aged 3 to 10 overestimated their immediate visual WM spans, but the extent of this overestimation was greatest for children aged 3 to 6. Similar findings were obtained in a more recent study by Forsberg et al. (2021a). They found that both children (aged 6 to 13) and young adults (aged 18 to 26) overestimated how many items they could remember from an array of colored squares, but the extent of this overestimation (relative to observed capacity limits) was greatest among the youngest children (up to age 8). Although these results suggest that younger children are less aware of the limitations of their WM compared to older children and young adults, recent evidence suggests that children as young as 5 can effectively monitor their WM when incentivized to do so. In a study by Applin and Kibbe (2021), 5-to-6-year-olds placed bets (e.g., betting candy that could be earned or lost) on whether they could remember the locations of between two and five just-presented objects. When children took risky bets (e.g., betting up to three of their candies), they were more accurate in remembering the locations of the objects compared to trials on which they made safe bets (e.g., betting only one of their candies). Such findings suggest that there is a

strong confidence-accuracy relation in WM among young children, at least when reward incentives encourage careful responding. It remains unknown, however, whether young children's confidence tracks their WM accuracy in more standard memory testing situations that do not incur a risk of winning or losing rewards. Furthermore, as in the young adult literature, in the development literature, there has not been a common method for measuring the confidence-accuracy relation in tests of WM and LTM. The present study intends to fill this gap.

Older Adults' Confidence-Accuracy Relation in WM and LTM

As noted earlier, older adulthood (typically studied from about age 65 to 80) is a period generally marked with declines in both WM and LTM. Older adults appear to be aware of these declines (Hertzog & Hultsch, 2000), as is evident by effects of stereotype threat (i.e., reminding older adults that aging leads to poorer memory) on older adults' memory performance (Barber & Mather, 2014; Brubaker & Naveh-Benjamin, 2018; Levy, 1996). Yet, this apparent awareness of age-related memory declines does not always translate into preserved metamemory processing in older adulthood. In studies of LTM, older adults exhibit a strong, positive confidence-accuracy relation (occasionally being even more accurate than younger adults at *high* confidence levels) when rating their confidence in memory responses to questions about general knowledge (Lachman et al., 1979; Marquie & Huet, 2000; Perlmutter, 1978; Pliske & Mutter, 1996), which presumably involves more semantic rather than episodic memory. Older adults' confidence in their recognition for specific episodes (i.e., episodic LTM) also tends to increase with improvements in accuracy (Colloff et al., 2017), but older adults are often prone to *high-confidence* memory errors (Dodson et al., 2007; Dodson & Krueger, 2006; Fandakova et al., 2013; Greene et al., 2022; Kelley & Sahakyan, 2003; Shing et al., 2009). Typically, these errors arise when older adults express *high* confidence that they previously saw something (e.g., a face

in a line-up) that was never presented (i.e., *high-confidence* false alarms). To explain this divergence in older adults' confidence-accuracy relation for more semantic versus more episodic LTMs, Dodson et al. (2007) proposed a misrecollection account (cf., Dodson, 2017). According to this account, older adults struggle to monitor their memories for specific episodes because they often erroneously bind elements from different episodes (e.g., Naveh-Benjamin, 2000) and retrieve general features that are shared across multiple similar episodes rather than specific features belonging to a particular episode (e.g., Henkel et al., 1998; Stark et al., 2013; cf., Greene & Naveh-Benjamin, 2023). In a recent test of the misrecollection account, Greene et al. (2022) showed that older adults' heightened susceptibility to *high-confidence* errors in tests of associative LTM occurred only in situations where they needed to remember specifically which scene was associated with a studied face (e.g., "was it this park, or a different park?") but not when they could rely upon general/gist representations (e.g., remembering that a face was paired with a nature scene as opposed to an indoor scene).

There is a paucity of research on older adults' confidence-accuracy relation on tests of WM. Most of the studies in this domain have assessed how accurate older adults are at predicting how much information they can hold in mind, with older adults often overestimating their capacity even more so than younger adults (Bunnell et al., 1999; Murphy et al., 1981). It is conceivable that older adults may exhibit *high-confidence* errors in tests of WM that resemble those in episodic LTM, given that metamemory monitoring can be resource-demanding (Stine-Morrow et al., 2006). As older adults have more limited attentional resource capacity (Craik & Bryd, 1982; Hasher & Zacks, 1988), the additional demands placed on their WM by evaluating the strengths of their memory representations in WM may contribute to *high-confidence* memory errors. Alternatively, older adults may be less prone to *high-confidence* memory errors in tests of

WM compared to LTM, as older adults are more capable of accessing specific details of a previous episode immediately after encoding, when the information can still reside within WM (Greene & Naveh-Benjamin, 2022).

The Present Study

As our review of the literature on the confidence-accuracy relation in young adulthood, childhood development, and adult aging indicates, several important gaps in our understanding of this critical relation remain to be filled. First, there have been few efforts to bridge these literatures with a lifespan approach. The existing studies that have done so have relied solely on tests of LTM (Fandakova et al., 2013; Shing et al., 2009), leaving unaddressed whether there are lifespan constancies or differences in the confidence-accuracy relation in tests of WM, the "gateway" through which new LTMs appear to be formed (Atkinson & Shiffrin, 1968; Forsberg et al., 2021b, 2022a, 2022b, 2023; Fukuda & Vogel, 2019). Second, there has been no unified approach for measuring the confidence-accuracy relation in tests of WM and LTM, as with a CAC analysis (Mickes, 2015) that overcomes limitations of the more widely used bivariate correlation approach, which remains the most commonly used method in studies of WM. Third, we have very limited understanding of how the confidence-accuracy relation is manifest in WM among children and young adults and virtually no understanding of this relation among older adults. This is an especially glaring limitation in light of developmental and age-related changes in the capacity of WM that underscore how much information young children and older adults, compared with young adults, can later access from LTM (e.g., Forsberg et al., 2022a, 2022b).

The present study sought to reconcile these limitations by applying a common method, CAC analysis (Mickes, 2015), to measure the confidence-accuracy relation across the lifespan using a novel paradigm (Forsberg et al., 2021b) that provides tests of both WM and LTM. The

data that informs our aims comes from two previously published studies using this paradigm (Forsberg et al., 2022a, 2022b), though this specific aspect of the data – the relation between participants' reported confidence in their WM and LTM recognition responses and their observed accuracy in those responses – has not been analyzed previously. Using this paradigm, we aimed to address two interrelated questions. First, are individuals of various ages (from childhood to older adulthood) aware of their memory strengths and weaknesses in both WM and LTM testing situations, such that, within an age group, individuals would be on average more accurate in their recognition memory when they express higher compared to lower retrospective confidence? Second, are individuals with generally poorer memory (e.g., young children or older adults) aware of their more impoverished memories, relative to those of other groups (e.g., young adults), such that they can calibrate their confidence ratings to ensure equal levels of accuracy at *high* confidence levels compared to individuals with stronger memories? In the ensuing, we lay out our hypotheses for each of these questions.

Hypotheses for the Confidence-Accuracy Relation Within an Age Group

Prior studies of LTM indicates that children and adolescents (Winsor et al., 2021), young adults (Wixted & Wells, 2017), and older adults (Colloff et al., 2017) all tend to be more accurate in their LTM recognition responses when they express higher compared to lower confidence (cf., Fandakova et al., 2013; Shing et al., 2009). Thus, we hypothesized that there would be a positive confidence-accuracy relation in tests of LTM within each age group of the present study.

In tests of WM, among young adults, we hypothesized that there would be a positive confidence-accuracy relation as well, in line with previous studies (Adam & Vogel, 2017; Bona & Silvanto, 2014; Vandenbroucke et al., 2014). However, because young adults are occasionally

prone to *high-confidence* errors (i.e., failing to detect a change when one occurs) when they overestimate their WM capacity (Cowan et al., 2016), it is possible that, in the present study, young adults would be more sensitive to *high-confidence* memory errors as the set size of encoded memoranda increases beyond their observed capacity limits. A similar, positive confidence-accuracy relation in WM may be found among children (e.g., Applin & Kibbe, 2021) and possibly older adults, given that the metamemory failures for WM observed in these groups (i.e., overestimating WM capacity; Bunnell et al., 1999; Flavell et al., 1970; Forsberg et al., 2021a; Murphy et al., 1981) mirror those observed in younger adults (albeit, these failures are more pronounced at the ends of the lifespan). Moreover, older adults can monitor their internal states during encoding about as well as younger adults (Hertzog et al., 2010), so they may be able to calibrate their confidence with their accuracy on tests immediately following encoding.

An additional point to consider is whether the *magnitude* of the confidence-accuracy relation (i.e., the difference in accuracy between lowest and highest confidence levels), within an age group, depends on whether the test occurs immediately after encoding (WM) or much later (LTM). A change in the magnitude of this relation may occur if, as memory signals become weaker or more contaminated by noise from WM to LTM, individuals are aware of this shifting quality of their mnemonic representations. If so, then participants may downregulate their LTM confidence relative to their WM confidence, such that they elicit *high* confidence ratings less often in tests of LTM. This does not mean that individuals would necessarily be more accurate when expressing *high* confidence in LTM than in WM tests, but rather, as a result of this potential downregulation in LTM that is less present in WM, there would be a more pronounced difference in recognition accuracy between lower and higher confidence levels in tests of LTM.

Hypotheses for Age Differences in the Confidence-Accuracy Mapping

We expect that certain groups (namely, young children and older adults) would have poorer overall memory performance than other groups, especially young adults. However, if individuals in groups with poorer memory are aware of their more impoverished memories and take this into account, then we would expect to find no age-related differences in recognition accuracy at *high* confidence levels, though such differences would likely be present at lower confidence levels. This would be so because these individuals would often downregulate their retrospective confidence, rarely expressing *high* confidence and only doing so in situations where they are likely to be accurate. As a result, the presence of an age difference at *low* confidence levels (e.g., young adults being more accurate than children or older adults) would provide misleading information as to whether there are age-related differences in confidence-accuracy calibration.[3]

Although studies of LTM have documented positive confidence-accuracy relations across the lifespan (e.g., Fandakova et al., 2013), young children (e.g., Winsor et al., 2021) and older adults (e.g., Dodson et al., 2007) are more prone to *high-confidence* memory errors than adolescents and young adults. Among older adults, these *high-confidence* LTM errors are characterized by overconfidence in erroneous recognition responses to new, unstudied information, particularly on tests of episodic rather than semantic content (i.e., tests requiring the retrieval of a specific prior episode; Dodson et al., 2007; Dodson, 2017; Greene et al., 2022). Because the LTM tests of the present study pertain to *episodic* memories (i.e., remembering specific items encountered in a specific time and place; Tulving, 1983), we hypothesized that older adults would be more prone to *high-confidence* LTM recognition errors than younger

[3] Relatedly, individuals who are *less accurate* at low confidence ratings may be said to have *better* metamemory (i.e., when they know their memory for a given item is unreliable, they express low confidence in whether they responded accurately on the basis of that memory), compared to individuals who are more accurate at low confidence ratings and thus may be unaware of the strengths of their memories.

adults. A similar tendency toward *high-confidence* LTM errors may also be found among young children in the present study, particularly those aged 7 and younger (e.g., Winsor et al., 2021) but potentially also for children up to age 10 (e.g., Brewer & Day, 2005; Howie & Roebers, 2007). It is conceivable that these age-related *high-confidence* memory errors may not appear in tests of WM, where young children or older adults may be better able to access their specific representations of recently encoded information (e.g., Greene & Naveh-Benjamin, 2022). Alternatively, because metamemory monitoring can be resource demanding (Stine-Morrow et al., 2006), it is possible that young children and older adults, with more limited attentional resource capacities (e.g., Cowan et al., 2006; cf., Craik & Bryd, 1982; Hasher & Zacks, 1988), may be as susceptible to *high-confidence* errors in WM as in LTM.

Method

Transparency and Openness

The two experiments on which the present study is based were pre-registered. However, the analyses for the present study were not pre-registered. Data and analysis scripts are available at https://osf.io/vfe9g/ (Greene et al., 2023).

Participants

Data from 320 participants from across the lifespan who participated in the experiments by Forsberg et al. (2022a, 2022b) were included in the analyses. This included 160 participants from Forsberg et al. (2022a), with 40 participants from each of four age groups: 1st-and-2nd grade children, 3rd-and-4th grade children, 5th-through-7th grade children, and young adults. In addition, 160 participants (80 young, 80 older adults) from Forsberg et al. (2022b) were included. Given the similarity in the procedures between the experiments, we combined the two samples of young adults, as both samples were similarly aged – *M*'s of 19.6 years (Forsberg et al., 2022a)

and 21.5 years (Forsberg et al., 2022b) – and were composed of a similar proportion of participants identifying as female or male (47.5% and 51.2% of young adult participants identified as female in Forsberg et al. (2022a) and Forsberg et al. (2022b), respectively). Demographic information is provided in Table 1.

Table 1. *Demographic Statistics of the Sample*

Age Group	*n*	Mean (SD) Age	Age Range	%Female, %Male
1st-2nd Graders	40	7.9 (0.72)	6-9	50.0%, 50.0%
3rd-4th Graders	40	9.8 (0.73)	8-11	47.5%, 52.5%
5th-7th Graders	40	11.9 (0.89)	10-13	47.5%, 52.5%
Young Adults	120	20.82 (2.38)	18-27	50.0%, 50.0%
Older Adults	80	69.0 (3.23)	65-77	67.5%, 32.5%

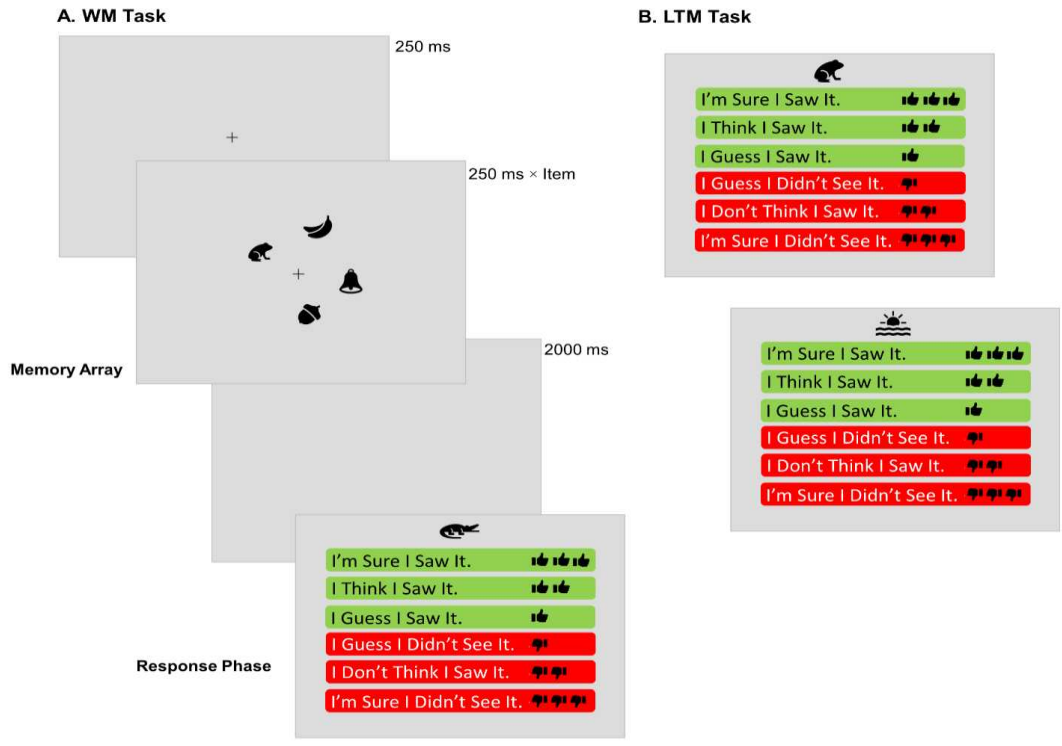
Note. Age is listed in years. %Female, %Male lists the percentage of participants who selected female or male as their preferred gender to a multiple-choice question; this question also included the options "Other" or "Prefer not to say," but no participants endorsed these options.

Materials, Design, and Procedure

Stimuli were 454 unique images from Microsoft Office Icons, which were automated for online experimentation via PsyToolkit (Stoet, 2010, 2017). All images were black and presented on a gray background. Participants studied a random 288 of these items; the remaining items were used as new test probes. There were two independent variables: the Set Size of the encoded memoranda (two, four, or six items presented concurrently at study) and the type of Memory Test (WM or LTM), both manipulated within-subject.

Both experiments used virtually identical procedures, with a few minor differences. All participants completed the experiment online, but participants in the study by Forsberg et al. (2022a) first met with the experimenter via Zoom, whereas participants in the study by Forsberg et al. (2022b) participated on the crowd-sourcing platform Prolific (Prolific, n.d.), without meeting the experimenter. Each experiment consisted of tasks presented in the following fixed order: (1) a WM probe-recognition task, (2) an interpolated activity unrelated to the WM task, and (3) a LTM probe-recognition test, assessing memory for items previously studied (but not tested) during the WM task (see Figure 1).

Figure 1. *Schematic Representation of the Procedure*



Note. An outline of a typical trial in working memory (WM; panel A) and long-term memory (LTM: panel B). In the WM task, items were presented around a central fixation cross in arrays of 2, 4, or 6 items

(Memory Array in panel A). The presentation duration of a study array was *n* x 250ms, where *n* denotes the number of items in the array. Following a 2000ms delay, a WM probe-recognition test occurred (Response Phase in panel A), with one item appearing at the top of the display that was either the same as an item from the previous array or was different. A 6-point rating scale appeared below the item, and participants selected one of the options corresponding to their confidence in whether the item was old or new. After completing all WM trials and following an additional 60 second period of interpolated activity, participants completed LTM probe-recognition tests (panel B), which consisted of a mix of old items from WM arrays which had not been tested on during the WM probe-recognition tests and new items. See online article for color version of figure.

Working Memory (WM) Task

Participants studied 288 unique items, with 96 items per encoding set size (i.e, the number of items presented concurrently in an array, which was either two (SS2), four (SS4), or six (SS6)). The set size manipulation produced conditions where the number of items to be encoded either fell within or outside a participant's actual WM capacity (the number of items that one can maintain in WM; see Cowan, 2001). The highest estimated capacity limits (*k*) for each group were as follows (see Forsberg et al., 2022a, 2022b): 1st-and-2nd grade children (*k* ~ 3.3 items), 3rd-and-4th grade children (*k* ~ 3.7 items), 5th-through-7th grade children (*k* ~ 4.0 items), young adults (*k* ~ 4.5 items), older adults (*k* ~ 3.2 items).

In each array, items were presented randomly in one of eight equidistant locations in an imaginary circle around a central fixation cross (see Figure 1A). Each trial began with a 250ms fixation period. The array was then presented for 250ms x *n*, where *n* represents the number of items in the array (e.g., in the SS6 condition, the memory array was presented for 1500ms). This was followed by a 2000ms blank delay. Then, a single probe appeared at the top of the display

that was either an old item from the just presented array or a new item that had not been presented at any point previously. There were an equal number of old and new test probes in the WM tests. Participants selected one of six response options appearing below the test probe, rating their confidence in whether they thought each test probe was old or new, from "I'm sure I saw it" (*high-confidence "old"* response) to "I'm sure I didn't see it" (*high-confidence "new"* response). Participants completed 88 WM trials (48, 24, and 16 trials at SS2, SS4, and SS6, respectively).

Long-Term Memory (LTM) Tests

After completing all 88 WM trials, participants completed an interpolated activity task for 60 seconds (see Forsberg et al., 2022a, 2022b for more details). Then they completed LTM probe-recognition tests, in which their memory for untested items from the earlier WM trials was assessed (see Figure 1B). Old LTM test probes were items from earlier studied arrays from the WM task; none of these items appeared as probes during the WM test trials. New LTM test probes had not appeared in either the WM study or test phases. The number of test items in the LTM task differed between Forsberg et al. (2022a) and Forsberg et al. (2022b). In the former, there were 168 LTM tests, composed of 46 new items, 36 old items sampled from SS2 arrays, 42 old items sampled from SS4 arrays, and 44 old items sampled from SS6 arrays. In Forsberg et al. (2022b), the number of new items was increased to 122, with no changes to the number of old items from each set size, resulting in 244 LTM test trials.

Analyses

The primary analysis of the present study was a CAC analysis (Mickes, 2015) designed to measure the relation between participants' reported confidence in their old/new recognition responses and their observed accuracy in their responses. First, however, we assessed the

memory discrimination abilities of individuals from the different age groups using receiver-operator-characteristic (ROC) analysis to provide a common framework for comparing performance on WM and LTM across the lifespan.

ROC Analyses

ROC analysis provides a bias-free metric of memory discrimination in old/new recognition tasks (Swets, 1988; Swets et al., 2000) and was applied to the present study using formulas for computing the ROC from confidence-ratings data (see Koen et al., 2016; Yonelinas & Parks, 2007). Memory discrimination was assessed with the area under the curve (AUC) metric, which ranges from 0 to 1, with values closer to 1 indicating near perfect memory discrimination, and values around 0.5 indicating chance-level discrimination.

CAC Analyses

We used CAC analyses to measure the confidence-accuracy relation (see Mickes, 2015; Wixted & Wells, 2017). In a CAC analysis, a participant's memory accuracy at a given confidence rating *i* is obtained by dividing the sum of their correct responses at the *i*th rating by the joint sum of their correct and incorrect responses at this rating. However, we relied instead on the proportion of responses at each rating for a given memory probe (e.g., what proportion of responses to old items from SS2 arrays in tests of LTM were erroneous *high-confidence "new"* responses). This was necessary to ensure that the CAC functions were placed on a common scale across age groups, given different number of new items in the LTM tests (see Procedure).[4] We split the 6-point confidence scale into two tripartite scales, one for "old" and one for "new"

[4] In the WM tests, whether the CAC functions relied on proportions of responses at each confidence rating or on the counts of responses at each rating, the analyses were on a common scale across age groups as there were no differences in the number of old and new test probes at each set size between experiments (Forsberg et al., 2022a, 2022b). However, in the LTM tests, relying on the counts of responses at each rating could bias age-related comparisons given that young and older adults in Forsberg et al. (2022b) responded to 122 new items, whereas all groups of children and the 40 young adults from Forsberg et al. (2022a) responded to only 46 new items.

recognition responses, each with *high*, *medium*, and *low* confidence ratings. We separately computed a CAC function for "old" and "new" recognition responses, using the formulas:

$$p(\text{Correct "old"})_i = p(\text{"old"}|\text{Old})_i \,/\, (p(\text{"old"}|\text{Old})_i + p(\text{"old"}|\text{New})_i)$$

$$p(\text{Correct "new"})_i = p(\text{"new"}|\text{New})_i \,/\, (p(\text{"new"}|\text{New})_i + p(\text{"new"}|\text{Old})_i)$$

That is, at the *i*th confidence level, the proportion of correct "old" responses was computed by dividing the proportion of hits ("old" responses to old items) by the summed proportion of hits and proportion of false alarms ("old" response to new items) at this confidence level. Similarly, the proportion of correct "new" responses at the *i*th confidence level was equal to the proportion of correct rejections ("new" responses to new items) divided by the summed proportion of correct rejections and proportion of misses ("new" responses to old items) at this confidence level. These proportions served as outcome variables in our statistical models.

We conducted a series of linear mixed effects models in the *lme4* package in R (Bates et al., 2015; R Core Team, 2023) to test our two core sets of hypotheses. Each model included a random intercept for each participant, as this was the maximal allowable random effects structure. Fixed effects varied across models, depending on the hypothesis being tested. To test our first set of hypotheses regarding within- age group relations between confidence and accuracy, we fit linear mixed effects models to the data from each age group separately for WM and LTM tests. For both types of memory tests, we tested for a main effect of Confidence Rating (low, medium, high). In tests of WM, we assessed whether the effect of Confidence Rating on accuracy depended on the Recognition Response ("old", "new") and/or the Set Size (SS2, SS4, SS6), within each age group. In tests of LTM, we combined these two factors into a single variable "Recognition Response by Probe Type," which included four levels created by combining the type of recognition response ("old" or "new") with the type of memory probe (old

items from SS2, SS4, or SS6; or new items). We also conducted analyses across memory tests, within each age group, to assess whether the change in memory accuracy from lowest to highest confidence levels differed between WM and LTM. This analysis was limited to "old" recognition responses and included fixed effects of Memory Test (WM, LTM) and Set Size.

For our second set of hypotheses, concerning age-related differences in *high-confidence* accuracy, we ran a separate linear mixed effects model for tests of WM and LTM. Each model included a fixed effect of Age Group (1st-and-2nd graders, 3rd-and-4th graders, 5th-through-7th graders, young adults, older adults). In WM, we tested whether this Age Group effect interacted with Recognition Response ("old", "new") and/or Set Size (SS2, SS4, SS6). In LTM, we tested whether Age Group interacted with the four-level "Recognition Response by Probe Type" factor defined previously.

To derive *F*-statistics and associated *p*-values for the fixed-effects terms in each model, we used the Satterthwaite approximation in the *lmerTest* (Kuznetsova et al., 2017) package to compute degrees of freedom based on Type III sum of squares, as recommended by Luke (2017). For significant main effects, we followed up with general linear hypothesis tests implemented via the *multcomp* package (Hothorn et al., 2008), using a Holm-correction for multiple comparisons. For significant interactions, we followed up with post-hoc comparisons in the *emmeans* package (Lenth, 2023), using Tukey adjustments for multiple comparisons.

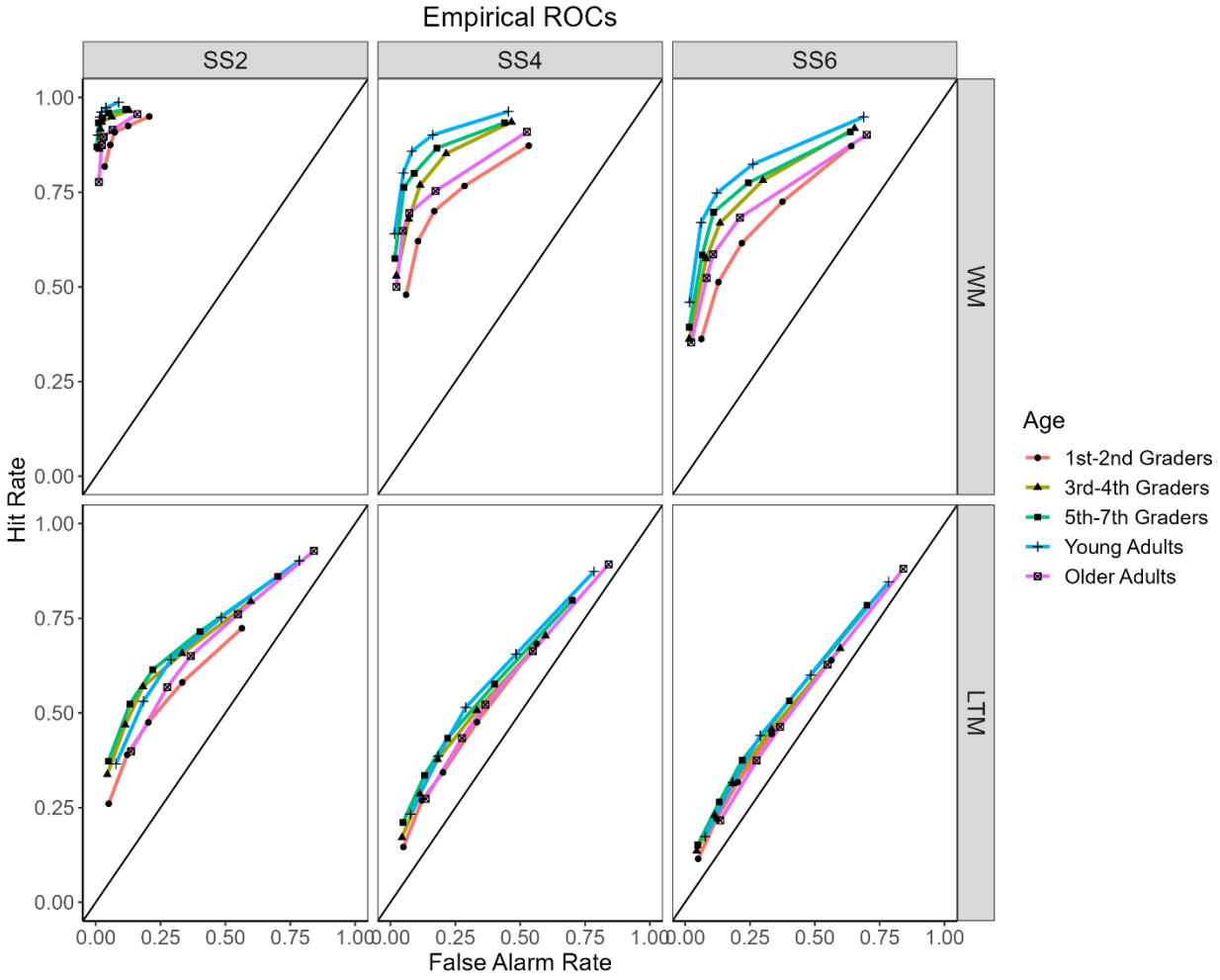
Results

Memory Discrimination on Tests of WM and LTM Across the Lifespan: ROC Results

Figure 2 depicts the average empirical ROC curves for each age group obtained under each encoding set size, separated by tests of WM (top panels) and LTM (bottom panels). ROC curves that quickly approach 1.0 on the Y-axis (cumulative hit rate) for values near 0 on the X-

axis (cumulative false alarm rate) indicate near-perfect discrimination, whereas ROC curves that lie closer to the diagonal indicate near-chance discrimination. A visual inspection of Figure 2 suggests that discrimination was generally better in tests of WM than LTM and for smaller rather than larger encoding set sizes. As a reminder, at the largest encoding set size (SS6), the number of items to be encoded exceeded the average WM capacity limits of all age groups, which ranged from k ~ 3.2 items (for older adults) to k ~ 4.5 items (for younger adults; see Method). Age-related differences in discrimination appeared to be more prominent in WM than in LTM.

Figure 2. *Empirical Receiver Operating Characteristic (ROC) Curves*



Note. Receiver operating characteristic (ROC) curves that lie closer to the solid diagonal line indicate near-chance discrimination of old and new items, whereas ROC curves that rapidly approach a cumulative hit rate of 1 for low cumulative false alarm rates (near 0) indicate near-perfect discrimination of old and new items. SS2 = set size 2; SS4 = set size 4; SS6 = set size 6; WM = working memory tests (top panels); LTM = long-term memory tests (bottom panels). Participant ages were as follows: 1st-2nd Graders (6-to-8 years old), 3rd-4th Graders (8-10 years old), 5th-7th Graders (10-13 years old), Young Adults (18-27 years old), Older Adults (65-77 years old). See online article for color version of figure.

To confirm these visual trends, we computed analyses on the AUC metric, which indexes memory discrimination, with higher AUC (near 1) corresponding to better discrimination. For each participant, a separate AUC was computed for each set size and in both WM and LTM. AUC metrics are listed in Table S1; all were reliably above chance (AUC of 0.50). The AUC metrics were submitted to a linear mixed effects model, with random by-subject intercepts, which detected significant main effects for each of the fixed effects of Age [$F(4, 315) = 14.07$, $p < .001$], Memory Test [$F(1, 1575) = 4679.82$, $p < .001$], and Set Size [$F(2, 1575) = 509.40$, $p < .001$]. However, because these main effects were qualified by significant two-way interactions involving each factor (all $p \leq .031$) and a significant Age x Memory Test x Set Size interaction [$F(8, 1575) = 2.52$, $p = .01$], for brevity, we concentrate on the post-hoc tests of the highest order interaction. Our aim was to measure age-related differences in mnemonic discrimination abilities, so we report post-hoc tests for assessing whether age differences depended on the Memory Test and Set Size.

On the WM tests, 1st-and-2nd grade children had lower AUC than young adults for SS2 arrays ($p_{Tukey} = .032$) and lower AUC than all groups (all $p_{Tukey} \leq .012$), except for older adults ($p_{Tukey} \geq .111$), for SS4 and SS6 arrays. Older adults had lower AUC than the 5th-through-7th

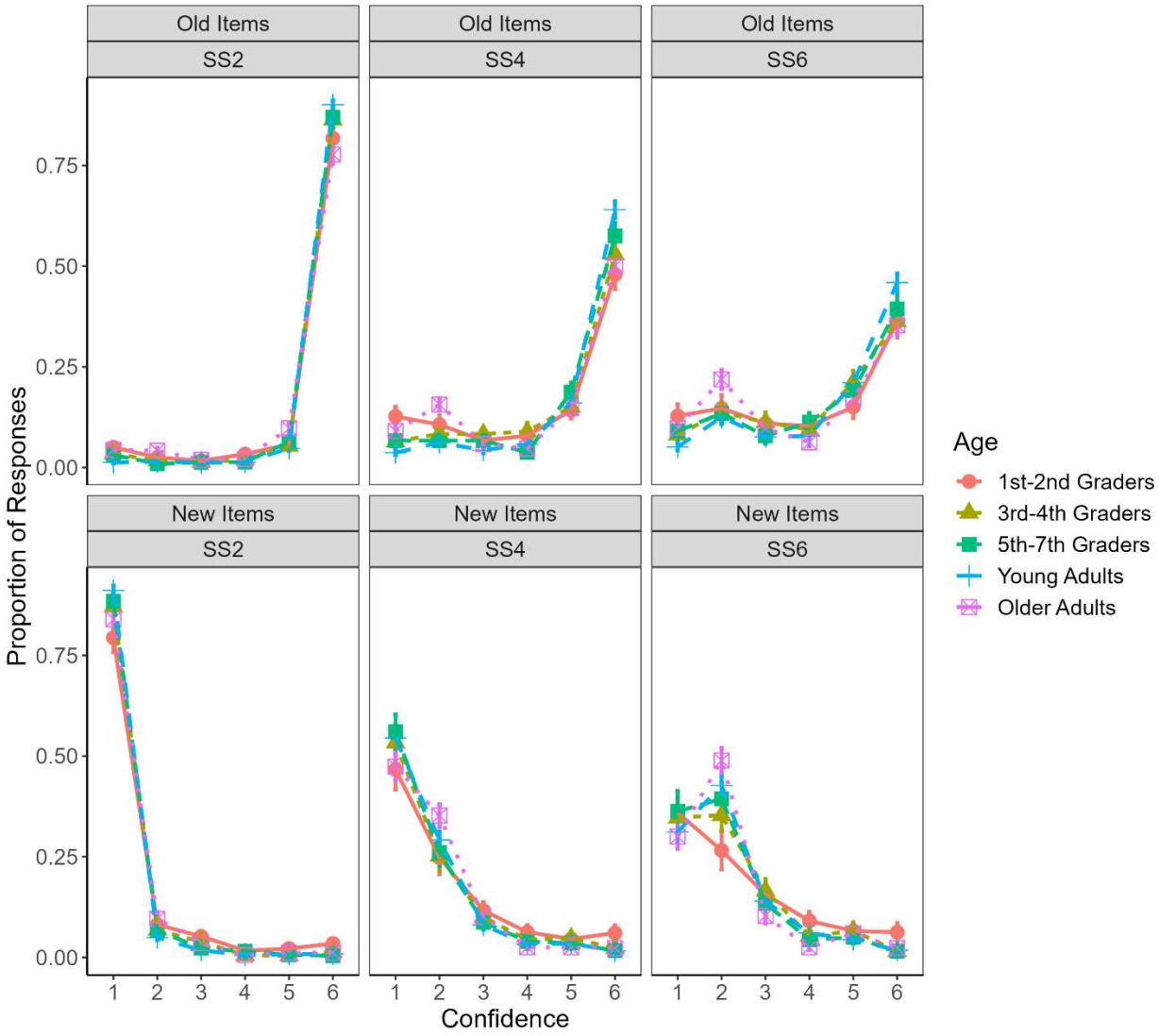
grade children (p_{Tukey} = .006) on WM tests at SS4 and lower AUC than both the 5th-through-7th graders and young adults (both p_{Tukey} < .001) on WM tests at SS6. Finally, the 3rd-and-4th grade children had lower AUC than the young adults (p_{Tukey} = .044) on WM tests at SS6. Thus, in WM, the youngest children generally had the poorest memory discrimination abilities at all set sizes, while the oldest children and the young adults had the best memory discrimination abilities. Older adults' memory discrimination abilities did not significantly differ from the youngest children but were worse than those of the oldest children and young adults at larger set sizes.

On the LTM tests, 1st-and-2nd grade children had lower AUC at SS2 than all groups (all $p_{Tukey} \leq .018$), except for older adults (p_{Tukey} = .188). However, at SS4, their AUC was only lower than that of young adults (p_{Tukey} = .024). Older adults' AUC on LTM tests at SS4 was marginally lower than that of younger adults (p_{Tukey} = .050). At SS6, there were no significant age-related differences in AUC on LTM tests (all $p_{Tukey} \geq .576$). Thus, age differences in memory discrimination abilities were less present in tests of LTM than in WM. The youngest children and occasionally the older adults had poorer LTM discrimination abilities than young adults. The youngest children also had poorer LTM discrimination abilities relative to the two older groups of children for items encoded under the smallest set size.

Confidence-Accuracy Relation in WM: CAC Results

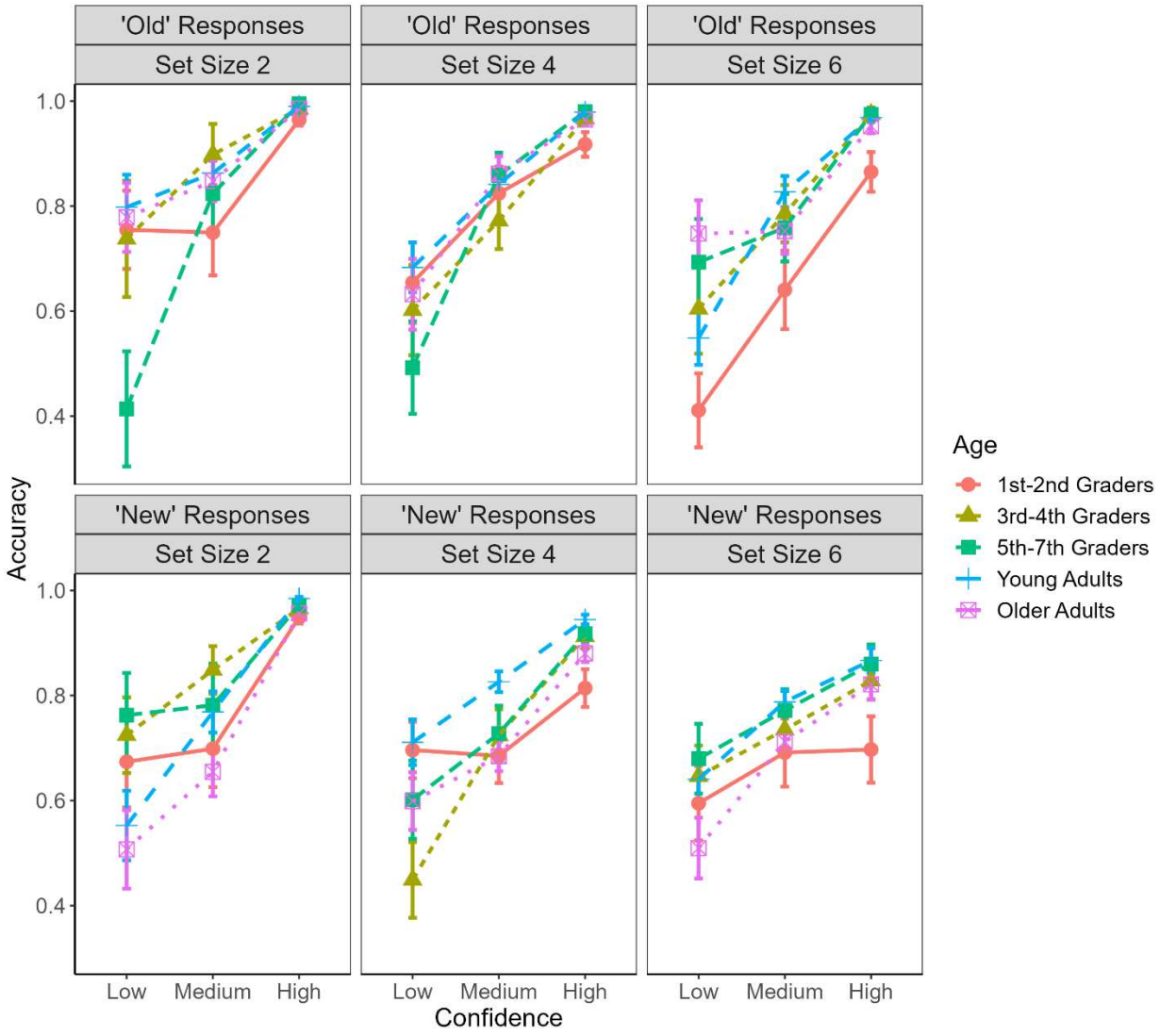
Figure 3 shows the proportion of responses at each confidence rating (ranging from 1 = *high-confidence "new"* responses to 6 = *high-confidence "old"* responses) to old and new items at each encoding set size in tests of WM, split by age group. Strikingly, participants of all ages responded with similar confidence patterns to old and new items at each set size. CAC curves based on these data are presented in Figure 4.

Figure 3. *Proportion of Responses at Each Confidence Rating for Each Probe in Working Memory Tests*



Note. Error bars represent +/- 1 standard error of the mean. Ratings are ordered as follows: 1 (*high-confidence "new"* responses), 2 (*medium-confidence "new"* responses), 3 (*low-confidence "new"* responses), 4 (low-confidence "old" responses), 5 (*medium-confidence "old"* responses), 6 *(high-confidence "old"* responses). SS2 = Set Size 2, SS4 = Set Size 4, SS6 = Set Size 6. See online article for color version of figure.

Figure 4. *Confidence-Accuracy Characteristic (CAC) Curves for Working Memory Tests*



Note. Error bars represent +/- 1 standard error of the mean. Lines between confidence ratings are intended to show the direction of the change in accuracy across adjacent confidence ratings, but the data are measured on a discrete scale. Set Size refers to the number of items (2, 4, or 6) studied concurrently in the study phase of the working memory trial. The confidence rating corresponds to how certain participants were when responding "old" or "new" to test probes in the recognition test. Memory accuracy at each confidence level was computed separately for "old" and "new" recognition responses. For "old" recognition responses, accuracy at each confidence level was calculated as the sum of hits (correct "old" responses to old items) divided by the joint sum of hits and false alarms (incorrect "old" responses to new

items) at the respective confidence level. For "new" recognition responses, accuracy at each confidence level was calculated as the sum of correct rejections (correct "new" responses to new items) divided by the joint sum of correct rejections and misses (incorrect "new" responses to old items) at the respective confidence level. See online article for color version of figure.

The CAC curves in Figure 4 generally depict increases in WM accuracy with increases in confidence, with highest accuracy observed under *high* confidence responses. This positive confidence-accuracy relation was manifest in each age group, with some variations. To address our first set of hypotheses about within-group relations between confidence and accuracy in tests of WM, we turn now to the results of our linear mixed effects models, fitted to the data from each age group separately. Each model tested the effect of Confidence Rating on WM accuracy and whether this effect depended on the Recognition Response ("old" versus "new") and/or on the Set Size (SS2, SS4, or SS6) of the WM array.[5]

WM Confidence-Accuracy Relation Across Childhood Development

Among each of the children groups, there were significant main effects of Confidence Rating on WM accuracy: 1st-and-2nd grade children [$F(2, 523.29) = 29.84$, $p < .001$], 3rd-and-4th grade children [$F(2, 527.48) = 63.21$, $p < .001$], and 5th-through-7th grade children [$F(2, 510.75) = 73.09$, $p < .001$]. WM accuracy was superior at *high* than both *medium* and *low* confidence among the 1st-and-2nd graders and the 5th-through-7th graders, all indicated post-hoc pairwise comparisons yielded $p_{Holm} < .010$. However, among the 3rd-and-4th graders, WM accuracy was only superior at *high* than *low* (but not *medium*) confidence, $p_{Holm} = .001$. Among all three groups, WM accuracy did not significantly differ between *low* and *medium* confidence, all p_{Holm}

[5] Throughout, we do not report on the main effects of Set Size or Recognition Response, as these were not germane to our hypotheses. We do, however, report on whether the main effect of Confidence Rating was qualified by significant interactions involving these factors.

≥ .220. Nevertheless, these results showed that there was some improvement in WM accuracy with increasing confidence, particularly between *low* and *high* confidence, in each group.

Notably, among the 3rd-and-4th grade children, the main effect of Confidence Rating was not qualified by any significant interactions with the remaining factors, all $p \geq .121$. Thus, this group of children had a similar confidence-accuracy relation for "old" and "new" responses in WM tests that did not differ by encoding set size. However, among the 1st-and-2nd grade children, there were marginally significant interactions of Confidence Rating x Recognition Response [$F(2, 502.51) = 2.55$, $p = .079$] and Confidence Rating x Set Size x Recognition Response [$F(4, 501.14) = 2.17$, $p = .072$]. Although neither interaction was significant, we followed-up on both, but we report only on the post-hoc tests of the highest-order (three-way) interaction. The main effect of Confidence Rating on 1st-and-2nd graders' WM accuracy reported earlier ((*low* = *medium*) < *high*) held for both "old" and "new" recognition responses at SS2. At SS4 and SS6, 1st-and-2nd graders' WM accuracy for "new" recognition responses did not differ across confidence ratings (all $p_{Tukey} \geq .175$), but their accuracy for "old" recognition responses was superior at *high* relative to *low* confidence at SS4 ($p_{Tukey} = .003$) and improved with each increase in confidence at SS6 (all $p_{Tukey} < .014$). Thus, the youngest children were somewhat more capable of calibrating their confidence with their memory accuracy for items in WM tests that they believed were old rather than new, especially as the encoding set size increased.

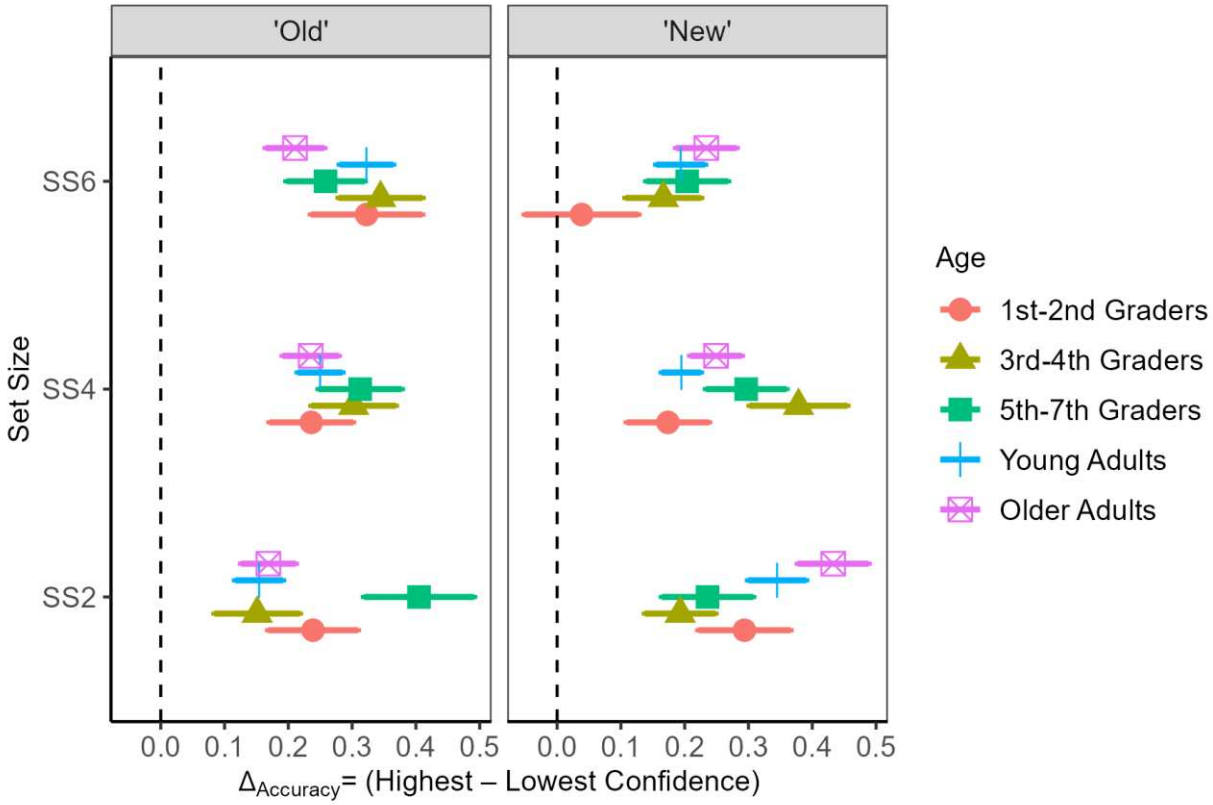
Among the 5th-through-7th grade children, there was a significant Confidence Rating x Recognition Response interaction [$F(2, 502.26) = 7.39$, $p < .001$] and marginally significant interactions of Confidence Rating x Set Size [$F(4, 501.80) = 2.16$, $p = .072$] and Confidence Rating x Set Size x Recognition Response [$F(4, 498.91) = 2.25$, $p = .062$]. We report first on the post-hoc tests of the significant two-way interaction before discussing those of the marginally

significant three-way interaction. For "new" recognition responses, the earlier-reported main effect on 5th-through-7th graders' WM accuracy held ((*low* = *medium*) < *high*), but for "old" recognition responses, there were improvements in WM accuracy with *each* increase in confidence, all $p_{Tukey} < .001$. Post-hoc tests of the three-way interaction indicated that this pattern for "new" recognition responses was obtained at SS2 and SS4, but at SS6, the only difference in "new" recognition accuracy was between *high* and *low* confidence ($p_{Tukey} = .015$). For "old" recognition responses, although 5th-through-7th graders' WM accuracy improved with each increase in confidence at SS2 (all $p_{Tukey} < .023$), their accuracy did not differ between *high* and *medium* confidence at SS4 ($p_{Tukey} = .100$) nor between *medium* and *low* confidence at SS6 ($p_{Tukey} = .617$). Thus, there were some fluctuations across recognition responses and encoding set sizes in the pattern of the oldest children's confidence-accuracy relation in WM, but they were always more accurate at *high* relative to *low* confidence.

These interactions should not obscure the fact that children ranging in age from 6 to 13 were usually most accurate in WM recognition at their highest expressed confidence levels and least accurate at their lowest expressed confidence levels. We computed the difference in WM accuracy (Δ) between the highest and lowest endorsed confidence levels for each participant, separately for "old" and "new" recognition responses at each set size. For most participants, this difference score was computed as *high-confidence accuracy* minus *low-confidence accuracy*. However, if a participant's highest or lowest confidence level was *medium*, the difference was computed as *medium-confidence accuracy* minus *low-confidence accuracy* or *high-confidence accuracy* minus *medium-confidence accuracy*, respectively. Difference scores are depicted in Figure 5. The only instance where there was no observed improvement in WM accuracy from

lowest to highest endorsed confidence levels was for "new" recognition responses in SS6 tests among the youngest children (1st-and-2nd graders), $t(37) = 0.43$, $p = .670$.

Figure 5. *Change in Working Memory Accuracy from Lowest to Highest Endorsed Confidence*



Note. Values depict the mean difference in working memory recognition accuracy for "old" (left side) and "new" (right side) responses between the highest and lowest endorsed confidence levels. Error bars represent +/- 1 standard error of the mean. Dashed line at 0.0 corresponds to a point null. SS denotes the set size of the studied array (2, 4, or 6 items). See online article for color version of figure.

WM Confidence-Accuracy Relation in Young and Older Adulthood

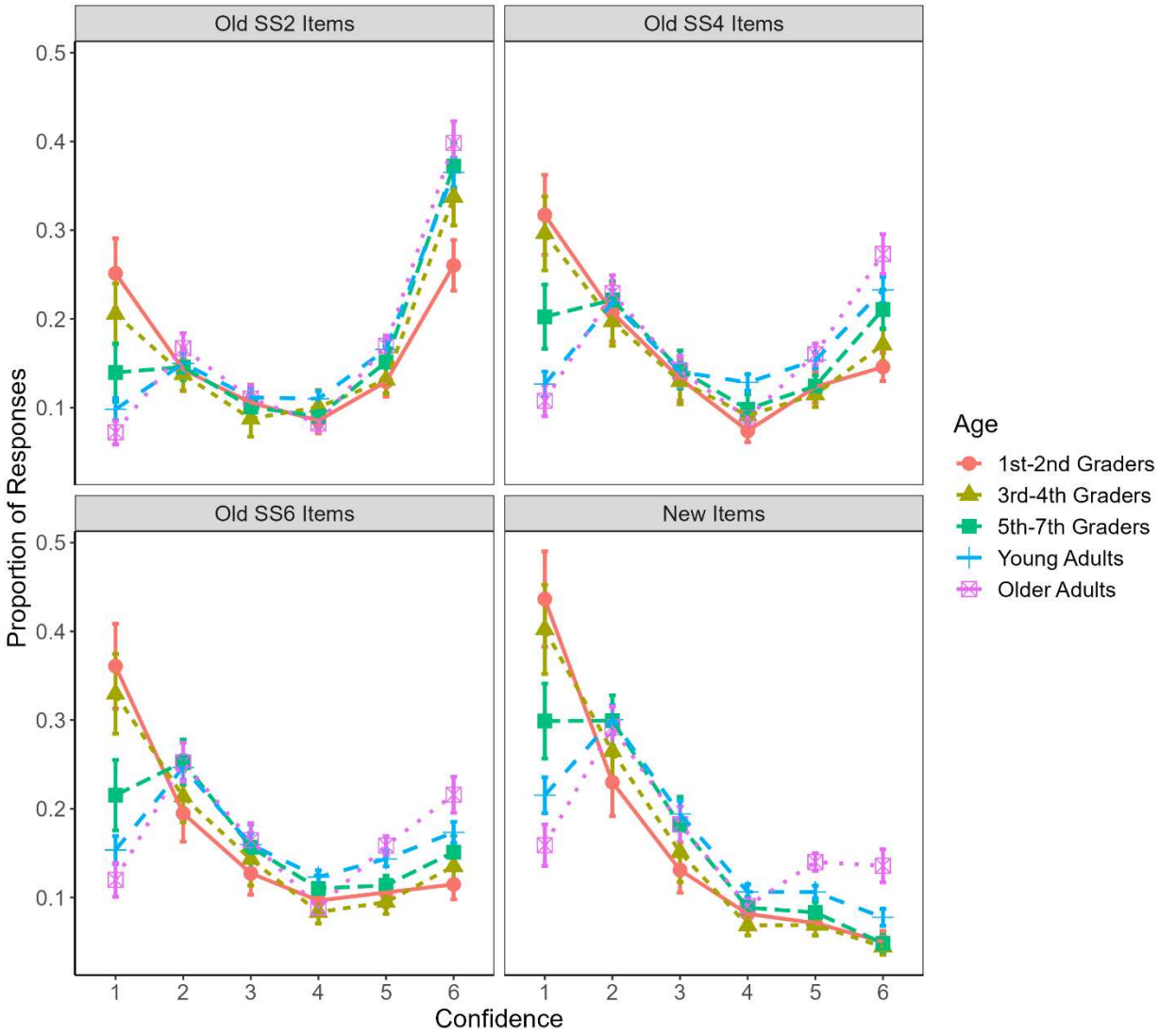
As with the children groups, the linear mixed effects models also detected significant main effects of Confidence Rating on WM accuracy among young adults [$F(2, 1548.3) = 165.17$,

$p < .001$] and older adults [$F(2, 1039.0) = 104.80$, $p < .001$]. In both groups, WM accuracy improved with each increase in confidence, all post-hoc pairwise comparisons yielded $p_{Holm} < .001$. However, these main effects were qualified by significant three-way Confidence Rating x Set Size x Recognition Response interactions among young adults [$F(4, 1506.1) = 6.51$, $p < .001$] and older adults [$F(4, 1005.1) = 4.60$, $p = .001$]. Among young adults, the same main effect of Confidence Rating (*low* < *medium* < *high*) was obtained in most conditions, except for "old" recognition accuracy at SS2, which did not differ between *low* and *medium* confidence, $p_{Tukey} = .465$. Older adults' WM accuracy did not significantly differ between *low* and *medium* confidence ratings for "old" recognition responses at both SS2 and SS6 and for "new" recognition responses at SS4 (all $p_{Tukey} \geq .168$). Nevertheless, among young and older adults alike, WM accuracy was always superior at *high* compared to *low* confidence (see Figure 5).

Confidence-Accuracy Relation in LTM: CAC Results

Next, we address whether a similar confidence-accuracy relation was found in tests of LTM, with increases in accuracy as participants' subjective confidence increased. Figure 6 shows the proportion of responses, split by age group, at each confidence rating (ranging from 1 = *high-confidence "new"* responses to 6 = *high-confidence "old"* responses) in tests of LTM to new items and to old items initially encoded under each set size.

Figure 6. *Proportion of Responses at Each Confidence Rating for Each Probe in Long-Term Memory Tests*



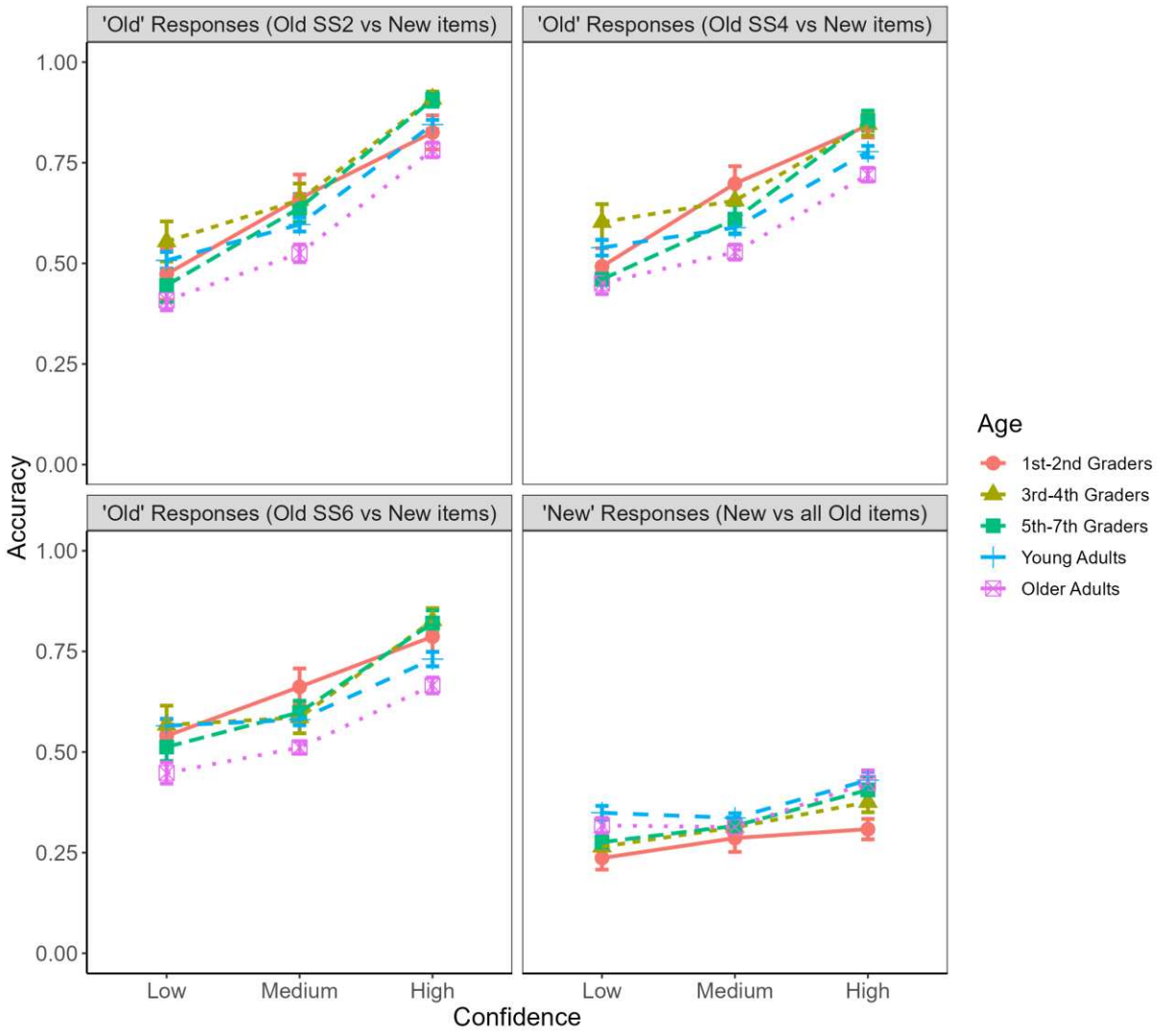
Note. Error bars represent +/- 1 standard error of the mean. Ratings are ordered as follows: 1 (*high-confidence "new"* responses), 2 (*medium-confidence "new"* responses), 3 (*low-confidence "new"* responses), 4 (*low-confidence "old"* responses), 5 (*medium-confidence "old"* responses), 6 (*high-confidence "old"* responses). SS indicates the set size (the number of items presented concurrently at encoding) under which the old items were encoded. SS2 = Set Size 2, SS4 = Set Size 4, SS6 = Set Size 6. See online article for color version of figure.

Unlike in the WM tests (see Figure 3), in the LTM tests, there were unique age-dependent patterns of responses. The two youngest groups of children (up to the 4th grade) responded often with *high-confidence "new"* (rating 1) responses to all probes. Meanwhile, the older groups, especially the older adults, responded more often with *high-confidence "old"* (rating 6) responses, particularly to old items. Indeed, collapsed across probe type, separate one-way ANOVAs detected significant main effects of Age Group on *high-confidence "new"* responses [$F(4, 315) = 13.90$, $p < .001$] and *high-confidence "old"* responses [$F(4, 315) = 5.81$, $p < .001$]. The 1st-through-4th grade children had a significantly higher proportion of *high-confidence "new"* responses than all other groups (all $p_{Tukey} \leq .033$). Meanwhile, young and older adults both had a significantly higher proportion of *high-confidence "old"* responses than 1st-and-2nd grade children (both $p_{Tukey} \leq .037$), and older adults additionally had a higher proportion of *high-confidence "old"* responses than 3rd-and-4th grade children ($p_{Tukey} = .011$). Thus, children up to about age 10 expressed higher confidence that items in the LTM tests were new, whereas adults (especially older adults) expressed higher confidence that items in the LTM tests were old, consistent with previous reports of age changes in bias (e.g., Cowan et al., 2006).

Despite these differences, the CAC curves depicted in Figure 7 show similar confidence-accuracy relationships (with increases in accuracy as confidence increases) in LTM across age groups, but the strength of this relation depended on the type of recognition response. For "old" recognition responses, there were greater improvements in accuracy with increases in confidence, relative to changes in accuracy across confidence levels to "new" recognition responses, in each age group. Indeed, across age groups, participants had low accuracy in "new" recognition responses, but "new" recognition accuracy still appeared to rise with increasing

confidence. Because in the LTM tests, new items did not map on to specific set sizes like old items did, our linear mixed effects models included a single fixed-effect factor combining Set Size and Recognition Response ("Recognition Response by Probe Type"), with four levels distinguishing old items from each set size (SS2, SS4, and SS6) and new items.

Figure 7. *Confidence-Accuracy Characteristic (CAC) Curves for Long-Term Memory Tests*



Note. Error bars represent +/- 1 standard error of the mean. Lines between confidence ratings are intended to show the direction of the change in accuracy across adjacent confidence ratings, but the data

are measured on a discrete scale. SS = Set Size, the number of items (2, 4, or 6) studied concurrently during an earlier study trial. The confidence rating corresponds to how certain participants were when responding "old" or "new" to test probes in the recognition test. Memory accuracy at each confidence level was computed separately for "old" and "new" recognition responses. For "old" recognition responses, accuracy was computed separately based on the set size from which old items were drawn. At each confidence level, accuracy to old items from a given set size was calculated as the proportion of hits (correct "old" responses to old items) at that confidence level divided by the sum of the proportion of hits and proportion of false alarms (incorrect "old" responses to new items) at that confidence level. For "new" recognition responses, accuracy at each confidence level was calculated as the proportion of correct rejections (correct "new" responses to new items) at that confidence level divided by the summed proportion of correct rejections and proportion of misses (incorrect "new" responses to old items, summed across old items from all encoding set sizes) at that confidence level. See online article for color version of figure.

LTM Confidence-Accuracy Relation Across Childhood Development

As in WM, there was significant a main effect of Confidence Rating on LTM accuracy among 1st-and-2nd grade children [$F(2, 383.41) = 38.92$, $p < .001$], 3rd-and-4th grade children [$F(2, 399.49) = 63.28$, $p < .001$], and 5th-through-7th grade children [$F(2, 404.69) = 140.45$, $p < .001$]. In all groups, LTM accuracy increased with each increase in confidence (all $p_{Holm} \leq .041$).

This main effect was qualified by a significant Confidence Rating x "Recognition Response by Probe Type" interaction among 1st-and-2nd grade children [$F(6, 379.91) = 2.79$, $p = .011$], 3rd-and-4th grade children [$F(6, 396.59) = 3.14$, $p = .005$], and 5th-through-7th grade children [$F(6, 401.93) = 6.63$, $p < .001$]. The youngest children's (1st-and-2nd graders) LTM accuracy for "old" recognition responses improved with *each* increase in confidence for responses to old items from SS2 and SS4 arrays (all $p_{Tukey} \leq .041$) but was only superior at *high*

relative to *low* confidence for responses to old items from SS6 arrays ($p_{Tukey} < .001$). Meanwhile, their LTM accuracy for "new" recognition responses did not significantly differ across confidence ratings (all $p_{Tukey} \geq .365$). Thus, as in WM, the youngest children were better at calibrating their confidence with their memory accuracy when they believed that items in the LTM tests were old rather than new.

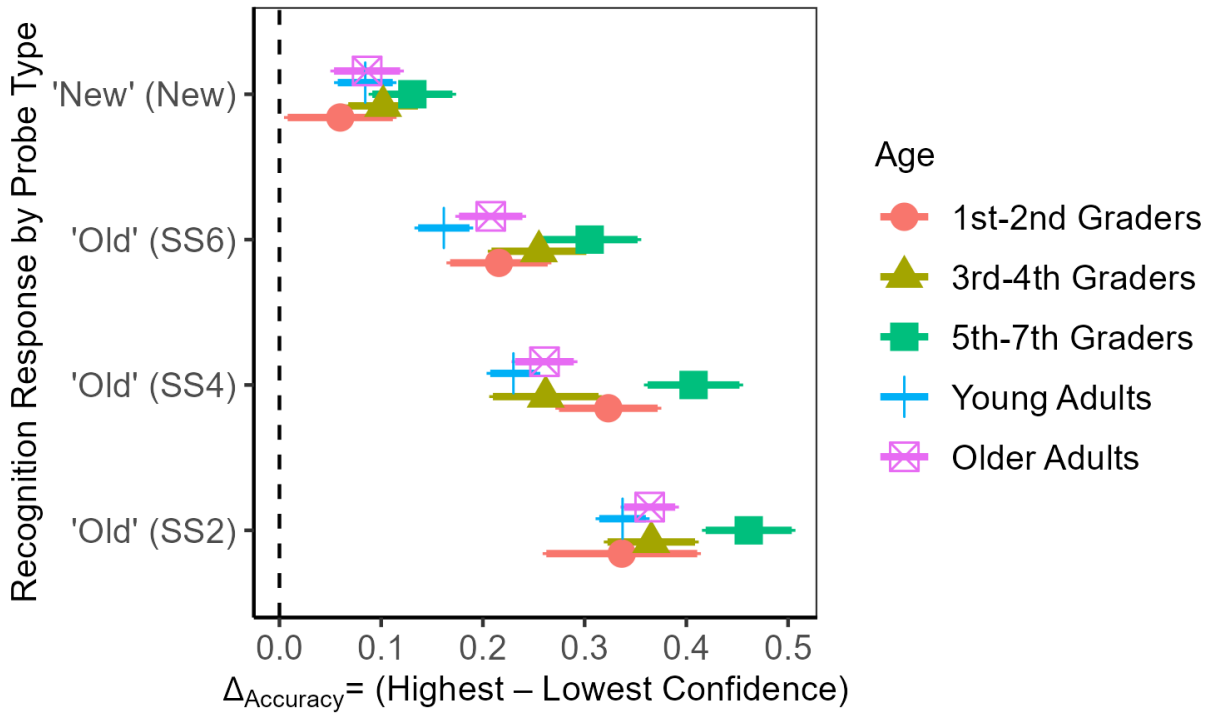
Among 3rd-and-4th grade children, LTM accuracy for "old" recognition responses similarly improved with each increase in confidence for responses to old items from SS2 arrays (all $p_{Tukey} \leq .042$) but was not significantly different between *medium* and *low* confidence for responses to old items from SS4 and SS6 arrays (both $p_{Tukey} \geq .479$). For "new" recognition responses, 3rd-and-4th grade children had superior LTM accuracy at *high* compared to *low* confidence ($p_{Tukey} = .041$). Thus, this group of children generally knew on which trials their LTM recognition, like their WM recognition, was most or least accurate, regardless of whether they responded "old" or "new."

Finally, the oldest children's (5th-through-7th graders) LTM accuracy for "old" recognition responses improved with each increase in confidence for responses to old items from SS2 and SS4 arrays (all $p_{Tukey} < .001$) but was not significantly different at *medium* and *low* confidence ratings for responses to old items from SS6 arrays ($p_{Tukey} = .091$). For "new" LTM recognition responses, 5th-through-7th grade children were more accurate at *high* compared to both *low* and *medium* confidence (both $p_{Tukey} \leq .044$). Thus, as in WM, the oldest children were generally aware of when their LTM recognition responses were most or least accurate.

These results show that children ranging in age from 6 to 13 were generally more in tune with which of their LTM recognition responses were the most accurate and which were the least accurate when they claimed to have previously seen an item, compared to when they claimed

that items were new (see Figure 8). In the latter case, the youngest children had the poorest insights into the reliability of their memories, as they were no more accurate in "new" LTM recognition responses at their highest relative to their lowest expressed confidence level. That is, their estimate of Δ (the change in LTM accuracy from lowest to highest expressed confidence) for "new" recognition responses did not differ significantly from 0, $t(37) = 1.16$, $p = .255$. This finding is consistent with the earlier-reported null effect of confidence ratings on 1st-and-2nd graders' WM accuracy to "new" recognition responses in the hardest encoding condition (i.e., SS6; see Figure 5). The older groups of children also had a weaker confidence-accuracy relation for "new" relative to "old" LTM recognition responses (Figure 8) but were nonetheless more accurate at their highest than lowest confidence levels.

Figure 8. *Change in Long-Term Memory Accuracy from Lowest to Highest Endorsed Confidence*



Note. Values depict the mean difference between highest and lowest endorsed confidence levels in long-term memory recognition accuracy for "old" responses to old items from each set size (SS2, SS4, or SS6) and "new" responses to new items. Error bars represent +/- 1 standard error of the mean. Dashed line at 0.0 corresponds to a point null. Formulas for calculating accuracy are listed in Figure 7 caption. See online article for color version of figure.

LTM Confidence-Accuracy Relation in Young and Older Adulthood

There were significant main effects of Confidence Rating on LTM accuracy among young adults [$F(2, 1276.0) = 201.51$, $p < .001$] and older adults [$F(2, 808.0) = 167.26$, $p < .001$] alike. Among both groups, LTM accuracy increased with each increase in confidence (all $p_{Holm} < .001$). This main effect was qualified by a significant Confidence Rating x "Recognition Response by Probe Type" interaction among young adults [$F(6, 1271.1) = 12.47$, $p < .001$] and older adults [$F(6, 791.63) = 7.65$, $p < .001$]. Among both adult groups, "old" recognition LTM accuracy improved with each increase in confidence for old items from SS2 and SS4 arrays (all $p_{Tukey} \leq .013$) but did not significantly differ between *low* and *medium* confidence responses to old items from SS6 arrays (both $p_{Tukey} \geq .052$). Similarly, young and older adults' "new" recognition LTM accuracy did not significantly differ between *low* and *medium* confidence ratings (both $p_{Tukey} \geq .558$). Nevertheless, much as in WM tests, young and older adults alike generally knew which of their LTM recognition responses were the most accurate and which were the least accurate. Indeed, as depicted in Figure 8, young and older adults were more accurate in their LTM recognition responses at their highest compared to their lowest endorsed confidence levels, though the magnitude of this effect was smaller for "new" recognition responses.

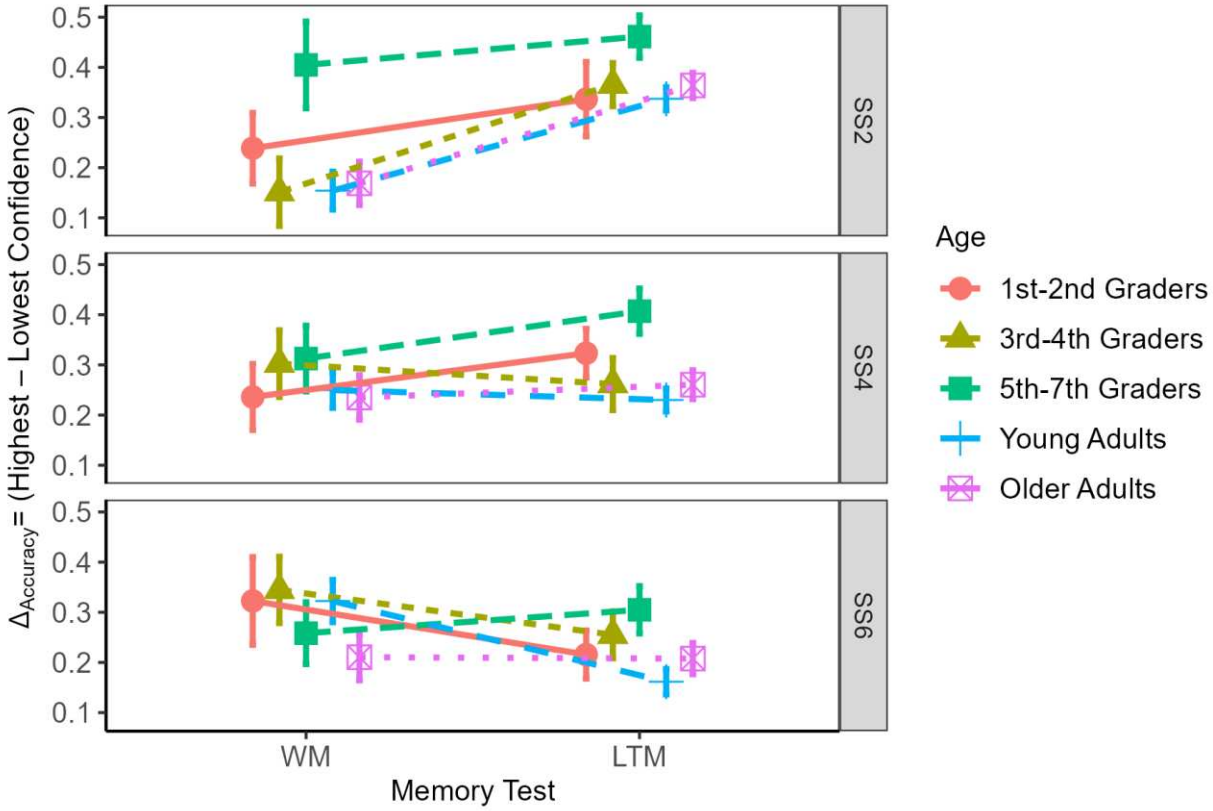
Confidence-Accuracy Relation: Comparing the Magnitude of the Relation between WM and LTM

We next assessed whether the magnitude of the confidence-accuracy relation (Δ, the accuracy difference between lowest and highest confidence levels) within an age group differed between tests of WM and LTM. We concentrated only on Δ for "old" recognition response accuracy because "new" recognition responses could not be split by encoding set size in LTM tests as they were in WM. Figure 9 shows within-age group changes between WM and LTM in Δ for "old" recognition responses.

We submitted the "old" recognition Δ scores to a separate linear mixed effects model in each age group. Each model included fixed-effects of Memory Test (WM vs LTM) and Set Size, plus their interaction. The only significant main effect of Memory Test was obtained among older adults [$F(1, 372.39) = 6.24$, $p = .013$], with a greater Δ in LTM than WM (for all other groups, $p \geq .166$). Among the youngest children (1st-and-2nd graders) and the oldest children (5th-through-7th graders), the Memory Text x Set Size interaction was also not significant (both $p \geq .247$), such that the magnitude of the confidence-accuracy relation for "old" recognition responses was comparable in tests of WM and LTM at all set sizes among these groups. However, there was a significant Memory Test x Set Size interaction among the 3rd-and-4th grade children [$F(2, 180.59) = 4.20$, $p = .016$], the young adults [$F(2, 522.65) = 16.84$, $p < .001$], and the older adults [$F(2, 359.39) = 4.40$, $p = .013$]. In each of these age groups, the magnitude of the confidence-accuracy relation was greater in tests of LTM than tests of WM only for recognition of items encoded in SS2 arrays (all $p_{Tukey} \leq .009$). That is, under the most favorable encoding conditions (i.e., in SS2, which fell within all groups' average capacity limits), participants in these age groups exhibited a smaller difference in "old" recognition accuracy between their

lowest and highest confidence levels when their memory was tested immediately after encoding (WM tests) compared to much later (LTM tests). Young adults, however, had a weaker confidence-accuracy relation (i.e., smaller Δ) in tests of LTM than tests of WM for old items from SS6 arrays, $p_{Tukey} < .001$.

Figure 9. *Change in Memory Accuracy for "Old" Recognition Responses from Lowest to Highest Endorsed Confidence in Tests of Working Memory versus Long-Term Memory*



Note. Values depict the mean difference in accuracy (Delta, or Δ) for "old" recognition responses between highest and lowest endorsed level of confidence, separately for tests of working memory (WM) and long-term memory (LTM), within each age group. Error bars represent +/- 1 standard error of the mean. Lines are intended to convey direction of change in Δ between tests of WM and LTM within an age

group. SS indicates the encoding set size of the items, with 2, 4, or 6 items per set. See online article for color version of figure.

We also assessed whether, at the individual level, "old" recognition Δ significantly correlated between tests of WM and LTM. In most groups, we did not detect significant relationships between WM Δ and LTM Δ (see Figure S1 in the online supplement), with one exception. For "old" recognition responses to SS4 items, WM Δ was significantly positively correlated with LTM Δ among the 1st-and-2nd grade children ($r = 0.38$, $p = .033$).

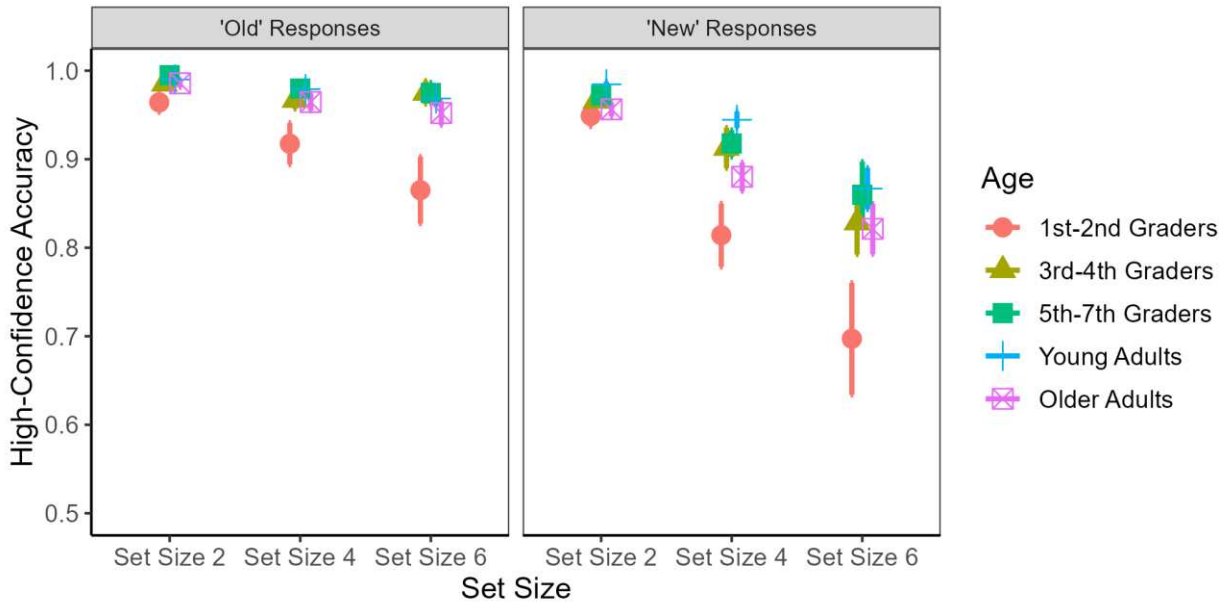
Age-Related Differences in Memory Accuracy at High Confidence Levels

The preceding CAC analyses showed that participants of all ages had some insights into the strengths and weaknesses of their memories in both WM and LTM tests, as recognition accuracy was generally highest at each participant's highest endorsed confidence level (see Figures 5 and 8). This positive confidence-accuracy relation was obtained across groups that varied in their observed memory discrimination abilities, which was generally poorest for the youngest children and older adults and best for the oldest children and young adults (see ROC results). Were individuals in age groups with poorer memory discrimination abilities able to take their weaker memories signals into account, "downregulating" their confidence ratings to obtain an equivalent level of accuracy at *high* confidence levels relative to individuals in age groups with superior memory discrimination abilities? To address this second core question of our study, we turn now to the results of our linear mixed effects analysis of age differences in *high-confidence* accuracy, with a separate analysis in tests of WM and LTM.

Age-Related Differences in High-Confidence Accuracy in Working Memory

Figure 10 depicts age-related differences in accuracy at *high* confidence ratings in tests of WM, separately for "old" and "new" recognition responses at each encoding set size. There was a significant main effect of Age Group, $F(4, 325.72) = 15.45$, $p < .001$, but post-hoc tests failed to detect any significant age-related differences (all $p_{Holm} \geq .524$). There was also a significant Age Group x Set Size interaction $F(8, 1500.51) = 3.11$, $p = .002$. At SS2, there were no significant age-related differences in *high-confidence* WM accuracy (all $p_{Tukey} \geq .369$), which was high (>90% correct) among all groups. At SS4 and SS6, however, the youngest children (1st-and-2nd graders) had lower *high-confidence* WM accuracy than all other groups (all $p_{Tukey} \leq .022$). Meanwhile, older adults were less accurate at *high* confidence levels than younger adults at SS4 ($p_{Tukey} = .037$) but not at SS6 ($p_{Tukey} = .238$). Thus, participants in the age groups with the worst memory discrimination abilities in WM (namely, the youngest children and occasionally the older adults) were somewhat more error-prone at *high* confidence levels than other groups.

Figure 10. *Age-Related Differences in High-Confidence Accuracy in Working Memory*

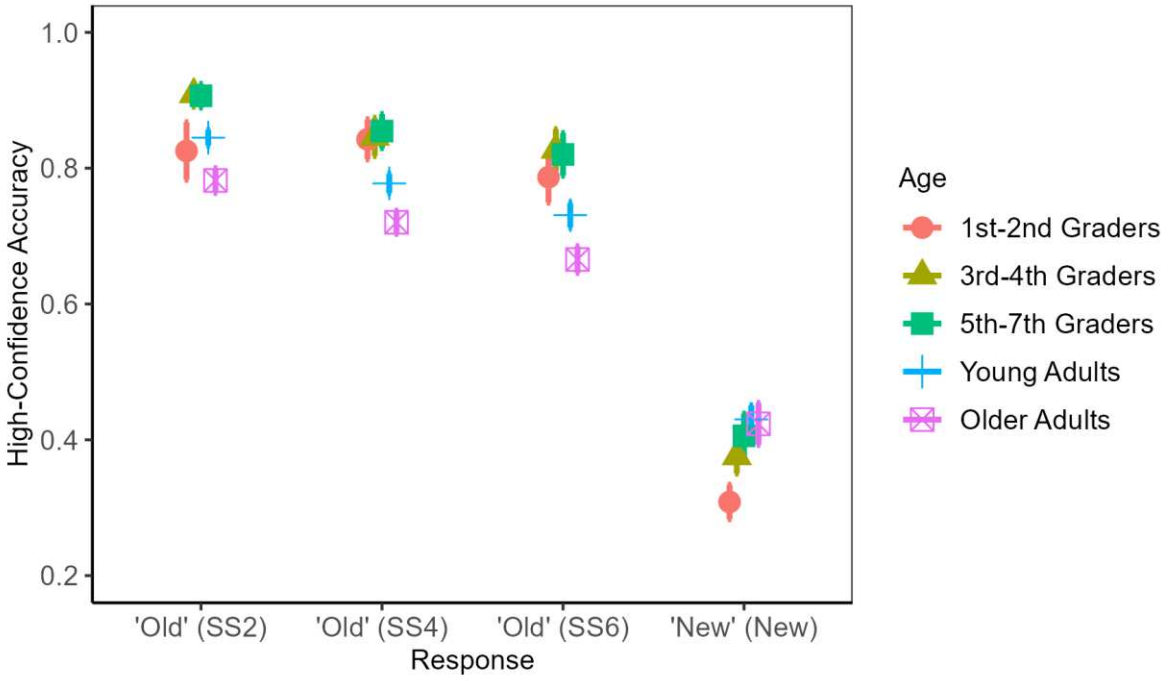


Note. Error bars represent +/- 1 standard error of the mean. Figure depicts how accurate participants were at the highest endorsed confidence rating when responding "old" (left panel) or "new" (right panel) in the working memory tests at each set size (x-axis). See online article for color version of figure.

Age-Related Differences in High Confidence Accuracy in Long-Term Memory

Figure 11 shows *high-confidence* accuracy in tests of LTM as a function of age group, separately for "old" responses to old items from each encoding set size and for "new" responses to new items. There was a significant main effect of Age Group, $F(4, 311.41) = 5.09$, $p < .001$, but as depicted in Figure 11, age-related differences in *high-confidence* LTM accuracy were dependent on the type of recognition response (i.e., Age Group x "Recognition Response by Probe Type" interaction), $F(12, 902.32) = 6.09$, $p < .001$.

Figure 11. *High-Confidence Accuracy in Long-Term Memory*



Note. Error bars represent +/- 1 standard error of the mean. Figure depicts how accurate participants were at the highest endorsed confidence rating when responding "old" or "new" in tests of long-term memory. For each type of recognition response, the value in parentheses corresponds to the type of recognition probe for which the response was correct, with "old" recognition responses split by the encoding set size under which old items were studied (SS2, SS4, SS6). See online article for color version of figure.

For "old" recognition responses across all set sizes, there were no significant differences in *high-confidence* accuracy among the three groups of children (all $p_{Tukey} \geq .243$) nor between young and older adults (all $p_{Tukey} \geq .115$). However, older adults were less accurate at *high* confidence levels for "old" recognition responses than the 3rd-and-4th grade children and the 5th-through-7th grade children at each set size (all $p_{Tukey} < .004$), and they were also less accurate than the 1st-and-2nd grade children at SS4 and SS6 (both $p_{Tukey} < .006$) but not at SS2 ($p_{Tukey} = .730$). Young adults were significantly less accurate than the 3rd-and-4th grade children ($p_{Tukey} = .032$) and marginally less accurate than the 5th-through-7th grade children ($p_{Tukey} = .058$) in *high-confidence* "old" recognition accuracy only at SS6. Meanwhile, for "new" recognition responses, there were no significant differences in *high-confidence* accuracy among the children (all $p_{Tukey} \geq .083$) nor between young and older adults ($p_{Tukey} = 1$). However, young and older adults alike were more accurate than the youngest children (1st-and-2nd graders) when rating their "new" recognition responses with *high* confidence (both $p_{Tukey} < .006$).

It is worth acknowledging here that we did not detect a significant adult age-related difference in LTM *high-confidence* accuracy when such differences are usually obtained in cross-sectional studies comparing young and older adults (e.g., Dodson et al., 2007; Dodson & Krueger, 2006; Fandakova et al., 2013; Greene & Naveh-Benjamin, 2022; Kelley & Sahakyan, 2003). Typically, these studies have found that older adults are more sensitive to *high-confidence*

false alarms (i.e., erroneously endorsing new items as "old" with *high* confidence). A visual inspection of Figure 11 suggests that older adults may have had a somewhat greater tendency than young adults toward *high-confidence* false alarms in the present data as well, as reflected by their generally lower accuracy in "old" recognition responses in tests of LTM. It is conceivable that any true adult age-related difference in *high-confidence* accuracy for "old" recognition responses may have been obscured by the large number of post-hoc comparisons that made tests of such differences more conservative. There is, however, *a priori* grounds to consider a more focused analysis limited to adult age-related differences in *high-confidence* "old" recognition accuracy, given the extensive prior literature on this specific analysis. Thus, we conducted a more constrained analysis, examining differences in *high-confidence* "old" recognition accuracy between young and older adults only in a 2 (Age Group) x 3 (Set Size) linear mixed effects analysis. There was a significant main effect of Age Group, $F(1, 194.46) = 8.86$, $p = .003$, with young adults outperforming older adults, but no significant interaction with the encoding set size, $F(2, 386.85) = 0.01$, $p = .986$. This more focused analysis replicated the traditional findings of adult age-related differences in *high-confidence* accuracy in tests of episodic LTM, but it is important to interpret these results with caution, given the failure to detect such differences under more conservative testing situations involving comparisons across *all* age groups.

Discussion

Results of the present study revealed strikingly universal relations between observed memory accuracy and subjective confidence in one's memory recognition across the lifespan. Children aged 6 to 13, young adults aged 18 to 27, and older adults aged 65 to 77 were almost always on average more accurate in their recognition responses when they expressed higher compared to lower confidence in those responses. Positive confidence-accuracy relations within

each age group were obtained both immediately after encoding, when items could still be maintained within capacity-constrained working memory (WM), and following a much longer sequence of events, when information had to be retrieved from long-term memory (LTM). Our results replicate those of earlier findings in childhood development (Winsor et al., 2021), young adulthood (Wixted & Wells, 2017), and older adulthood (Colloff et al., 2017) documenting positive confidence-accuracy relations in LTM, albeit with variations in the strength of these relations across the lifespan (cf. Fandakova et al., 2013; Shing et al., 2009). Extending on these earlier studies, we found that similar confidence-accuracy relations arise in WM, the "gateway" through which new LTMs are formed (Forsberg et al., 2021b, 2022a, 2022b, 2023; Fukuda & Vogel, 2019; cf., Atkinson & Shiffrin, 1968, Cowan, 1988, 2019; Cowan et al., 2024). Moreover, the magnitude of the confidence-accuracy relation within an age group (i.e., the difference in recognition accuracy between lowest and highest confidence levels) was usually comparable in WM and in LTM, with rare exceptions.

The pervasiveness of the confidence-accuracy relation across the lifespan and across memory tests that varied in their difficulty (in terms of number of items to be encoded) and timing (relative to encoding) speaks to a robust ability of individuals to evaluate the contents of their memories to determine when they feel more or less certain that they have previously experienced an episode. Yet, age-related differences in confidence-accuracy calibration were present in some conditions. Specifically, a few important differences in *high-confidence* accuracy emerged among the groups with generally poorest memory discrimination abilities (children aged 6 to 8, and older adults) relative to the groups with the best memory discrimination abilities (children aged 10 to 13 and young adults), but there were also some surprising cases in tests of LTM where children outperformed young and older adults. In tests of

WM, the youngest children (6-to-8-year-olds) were more error-prone at *high* confidence levels than most other groups (with the exception of older adults) when the encoding set sizes exceeded their WM capacity of approximately 3.3 items (i.e., in the SS4 and SS6 conditions). Older adults were also occasionally less accurate than young adults at *high* confidence levels in tests of WM, but only in cases where the encoding set size (SS4) exceeded older adults' WM capacity (k ~ 3.2 items) but was within the capacity limits of younger adults (k ~ 4.5 items). When the encoding set size was within (SS2) or exceeded (SS6) the capacity limits of *both* young and older adults, adult age-related differences in *high-confidence* WM accuracy were not present.

Meanwhile, in tests of LTM, older adults appeared to be somewhat more prone to *high-confidence* false alarms (endorsing new items as "old") than younger adults were, replicating prior work (Dodson, 2017; Dodson et al., 2007; Fandakova et al., 2013; Greene et al., 2022). It is important to acknowledge, however, that these adult age differences were only detected under very focused comparisons between young and older adults but not under more conservative statistical tests involving comparisons of all age groups. Older adults were also generally less accurate than all groups of children when endorsing *high* confidence in their "old" recognition responses in tests of LTM. Strikingly, even the young adults were occasionally overconfident in their "old" recognition responses in tests of LTM, relative to children aged 8 to 13, but only in discriminating new items from old items encoded under the most extreme conditions (i.e., in the SS6 condition). Participants of all ages had rather low accuracy in their "new" recognition responses in tests of LTM, but "new" recognition accuracy did improve with increasing confidence across age groups, except among the 6-to-8-year-old children. At *high* confidence levels, young and older adults were more accurate in their "new" LTM recognition responses than the youngest children were.

These results provide new insights into age-related differences in metamemory across the lifespan that have important implications for understanding the source of memory failures for young children and older adults, relative to young adults with generally more optimal memory capabilities (e.g., Cowan et al., 2006; Naveh-Benjamin & Cowan, 2023). Before discussing these implications, we first consider *why* such a pervasive confidence-accuracy relation was obtained across the lifespan, putting aside for now age differences in the strength of this relation.

On the Pervasiveness of the Confidence-Accuracy Relation Across the Lifespan

The present study adds to a growing body of evidence showing that confidence does, in fact, track memory accuracy among children (Winsor et al., 2021), young adults (Wixted & Wells, 2017), and older adults (Colloff et al., 2017), even though conventional views have often suggested that this was not the case among individuals with poorer memory capabilities (e.g., children and older adults; Dodson & Krueger, 2006; Newcombe & Bransgrove, 2007; Keast et al., 2007; Knutsson & Allwood, 2014; Powell et al., 2013). Much of the conventional wisdom has been predicated on empirical measures of the confidence-accuracy relation that can under- or overestimate this relation (for extensive critiques, see Busey et al., 2000; Juslin et al., 1996; Mickes, 2015). Here, we used CAC analyses that overcome these limitations (Mickes, 2015; Wixted & Wells, 2017), allowing us to measure the confidence-accuracy relation in a common way across tests of WM and episodic LTM and among individuals varying in age from 6 to 77. We observed an almost universal relation between retrospective confidence and memory accuracy, in WM and in LTM testing situations; for both “old” and “new” recognition responses; across encoding set sizes that fell within (SS2), just outside (SS4), or greatly beyond (SS6) most individuals’ WM capacities; and across age groups, from childhood to older adulthood. That is, memory accuracy was almost always superior at each individual’s highest relative to their lowest

endorsed confidence in their recognition responses (see Figures 5 and 8). Why was such a pervasive confidence-accuracy relation obtained?

Theorists have long sought to explain the confidence-accuracy relation, dating back to the seminal ROC analysis of recognition memory by Egan (1958) and the ensuing development of signal detection theory (Green & Swets, 1966; for a review, see Yonelinas & Parks, 2007). From a signal detection theory perspective, confidence ratings represent different criteria/thresholds placed along a memory strength dimension, sometimes referred to as familiarity, such that when a signal (i.e., the internal representation generated by a recognition probe) exceeds a given criterion, participants become increasingly confident that the signal is "old." These criteria may be fixed in such a way that their relative placement along the memory strength dimension does not change as a function of the degree to which signal and noise distributions overlap (e.g., Glanzer & Bowles, 1976; Hintzman, 1988). Such a fixed-criterion model would yield decreasing accuracy at a given confidence rating as the memory discrimination process becomes increasingly impaired. Alternatively, individuals may adjust the relative placements of their ratings' criteria along the strength dimension to optimize relatively constant accuracy at a given confidence level across conditions of better or worse memory. This is in line with the predictions of several formal models of recognition memory, including some global matching models of memory (e.g., Shiffrin & Steyvers, 1997) and likelihood-ratio models (e.g., McClelland & Chappell, 1998; Osth et al., 2017; Stretch & Wixted, 1998). However, one's ability to adjust their criteria may depend on acquiring error feedback about when one's memories are accurate or not, and such feedback is more likely to be developed over the course of one's life (Mickes et al., 2011). Consequently, although both young and older adults appear to be able to adjust their confidence ratings with task difficulty to some shared extent (Colloff et al., 2017; Semmler et al.,

2018), children younger than 10 are less adept at doing so (Winsor et al., 2021). Indeed, the youngest children in the present study (particularly those aged 6 to 8) were less capable than all other groups of adjusting their confidence ratings with their memory signals when they believed that items were new, rather than old, in tests of WM and LTM.

There is a very telling feature of our data that speaks more strongly to the criterion-shift than the fixed-criterion account of the confidence-accuracy relation, across all age groups, at least in WM. An examination of the observed proportions of responses at each confidence level in tests of WM (depicted in Figure 3) shows dramatic shifts in *high-confidence* responses within each age group as a function of the difficulty of the memory test (i.e., based on the set size). The precipitous drop in *high-confidence* "old" responses to old items with increasing set size could be explained either by a criterion-shift model (i.e., individuals may shift their *high-confidence* criteria to more conservative positions for more difficult tests, like those occurring at larger set sizes) or by a fixed-criterion model (i.e., with a fixed criterion for *high-confidence "old"* ratings, far fewer old items would pass this criterion when the signal distribution is itself shifted closer to the noise distribution). However, the mirrored drop in *high-confidence* "new" responses to new items with increasing set size cannot be reconciled by a fixed-criterion model because the noise distribution from which new items are sampled should be fixed to the same position across set sizes. That is, only the signal distribution (but not the noise distribution) would change across set sizes.[6] Thus, assuming a fixed-criterion model, we should have observed comparable rates of *high-confidence* "new" responses to new items, regardless of set size. That these rates clearly

[6] This assumption could be challenged if one were to develop a signal detection model in which the variance of the noise distribution (which is typically modeled as a normal distribution with mean = 0 and SD = 1) was allowed to vary as a function of the set size, but doing so would require modifications to calculating metrics of discrimination or response bias that may in turn alter the interpretation of those metrics, many of which (like *d´*) are based on *z*-statistics from a standard normal distribution. We are aware that in the unequal variance model, one can directly model the variance of the signal distribution, but in this case, the variance of the noise distribution is still fixed at 1.

change as a function of the set size indicates that participants of all ages appear to shift their ratings' criteria for *high-confidence* responses based on the difficulty of the test, at least in WM. In tests of LTM, we cannot clearly disentangle the fixed-criterion and criterion-shift models based on the present data because new items were different from *all* old items from each set size. Thus, the LTM discrimination process was not specific to a given set size, so individuals probably had only one set of criteria for judging whether an item was new or old. Nevertheless, in LTM as in WM, the most parsimonious explanation for why a positive confidence-accuracy relation exists at all is that both confidence and accuracy are mapped to the same latent memory strength continuum, as predicted by signal detection theory (Stretch & Wixted, 1998; Yonelinas & Parks, 2007; cf. McClelland & Chappell, 1998; Osth et al., 2017; Shiffrin & Steyvers, 1997; but for an alternative perspective, see Busey et al., 2000; Koriat, 1993).

On Age-Related Differences in High-Confidence Accuracy

Our lifespan approach was particularly useful because it allowed us to consider whether the source of age differences in memory are the same for young children and older adults, both of whom have generally poorer memories than young adults (Brockmole & Logie, 2013; Cowan et al., 2006; Fandakova et al., 2013; Shing et al., 2009, 2010). Indeed, in the present study, an ROC analysis revealed that the youngest children (6-to-8-year-olds) had the poorest memory discrimination abilities in WM and LTM, and older adults occasionally (i.e., at SS4) had worse memory discrimination abilities than young adults in both WM and LTM. The *high-confidence* errors in WM and LTM tests of the youngest children and older adults, respectively, speak to potential different sources of their memory errors, relative to young adults. Young children appear to be overconfident about the process of encoding information into memory, whereas

older adults appear to be overconfident about what information they have retained in the long-term.

Previous studies have shown that children younger than 8 exaggerate their WM capacities to a far greater extent than young adults do (Flavell et al., 1970; Forsberg et al., 2021a). Our findings that 6-to-8-year-old children were less accurate than young adults (and, indeed, than most other groups) in their *high-confidence* WM recognition responses when the encoding set size exceeded their capacity (i.e., in the SS4 and SS6 conditions) are well-aligned with these earlier findings. Moreover, the fact that young children's accuracy in "new" WM recognition responses at SS6 did not differ between their lowest and highest confidence levels suggests that they may have been overconfident in their abilities to encode information. That is, they appeared to be overconfident that they would have remembered seeing an item if it had been presented. Collectively, these findings suggest that young children's memory errors, relative to young adults, may be primarily attributable to failures during encoding in accurately self-monitoring how much information they have acquired (and, complementarily, what information they failed to acquire). Speculatively, such failures in self-monitoring during encoding may be related to young children's less developed frontal lobes (Giedd et al., 1999; Sowell et al., 2001), the brain region implicated in numerous executive functions (e.g., Alvarez & Emory, 2006; Stuss, 2011) that may be necessary for accurate self-monitoring of one's memories.

Although older adults may also exaggerate their WM capacities (Bunnell et al., 1999; Murphy et al., 1981), their ability to self-monitor during encoding is often comparable to that of younger adults (Hertzog et al., 2010). This is so even though the frontal lobes also undergo changes with adult aging (Cabeza & Dennis, 2012; West, 1996), but many of these changes may be compensatory (Cabeza et al., 2018). In the present study, the only condition in which older

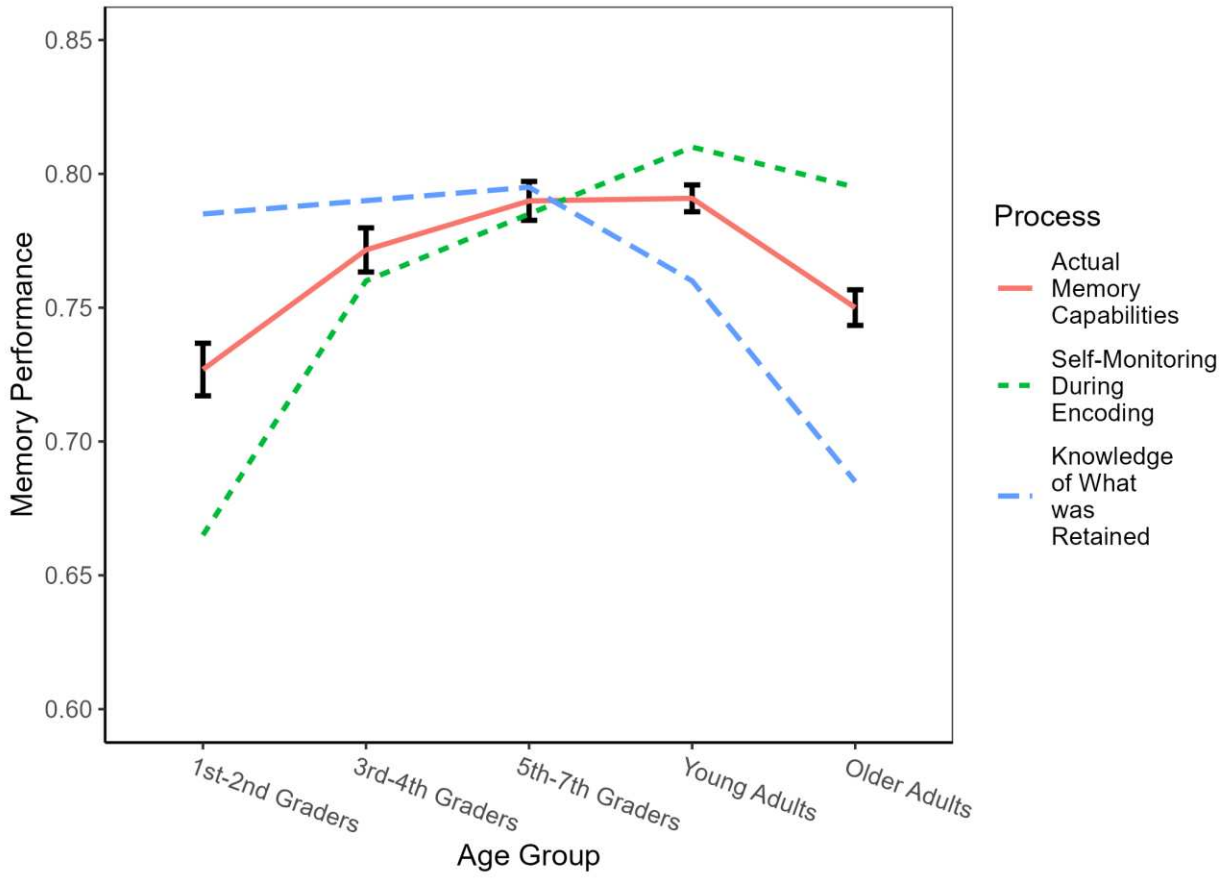
adults were more prone than younger adults to *high-confidence* memory errors in WM tests was when the encoding set size just barely exceeded their capacity limits of about 3.2 items (i.e., in the SS4 condition). When the encoding set size greatly exceeded their capacity limits (i.e, in the SS6 condition), older adults appeared to have some awareness that their WM was severely overburdened, leading them to downregulate their confidence in their recognition responses to obtain equivalent accuracy levels at *high* confidence ratings as younger adults. However, by the time of LTM testing, older adults were somewhat more likely than young adults to commit *high-confidence* false alarms (i.e., to endorse new items as "old" with great certainty; cf., Dodson et al., 2007; Fandakova et al., 2013; Greene et al., 2022; Kelley & Sahakyan, 2003). Thus, the source of older adults' memory errors may be related more to temporal lobe-based processes (e.g., those involving the hippocampus) implicated in retrieving the details of what was previously learned (Robin & Moscovitch, 2017; Yassa et al., 2011). Older adults may engage in less elaborate or effortful retrieval needed to reintegrate specific details of past episodes (Jacoby et al., 2005; cf. Luo & Craik, 2009). This may lead them to conflate with high confidence that new items or events are old on the basis of a vague sense of familiarity to the items (e.g., from a long history of encountering similar items) in the absence of recollection of the context in which those items were encountered previously (Jennings & Jacoby, 1993, 1997).

Notably, even the young adults, like the older adults, were more prone to *high-confidence* "old" recognition errors in tests of LTM, relative to children aged 8 to 13. Young and older adults alike were more likely to express *high* confidence that items in the LTM tests were "old," whereas children of all ages rarely did so and more often expressed *high* confidence that items in the LTM tests were "new." There was a somewhat gradual change in these extreme confidence ratings across childhood development (as shown in Figure 6), with the confidence ratings of

older children more closely matching those of young adults. Thus, with aging, individuals become more certain about what they have retained in the long-term, such that age-related changes in metamemory monitoring of long-term retention may be a more gradual, lifespan process. Occasionally, and especially with advanced aging, this increased certainty comes at the expense of greater susceptibility to falsely remembering things that never happened.

Figure 12 provides a theoretical sketchpad for the mapping between metamemory processes and actual memory capabilities from young childhood to older adulthood based on the data from the present study. The red curve in Figure 12 corresponds to the observed memory discrimination abilities, the average AUC across tests of WM and LTM, of participants from each age group. The green line represents the ability to self-monitor one's memory during encoding, and the blue line represents the ability to accurately assess what or how much information one can retain in the long-term. The figure depicts increases in encoding self-monitoring from early childhood through adolescence and into young adulthood, with relative stability in this process into older adulthood. In contrast, the ability to accurately assess how much information one retained in the long-term declines from childhood to young adulthood, with further, more marked declines in older adulthood.

Figure 12. *Theoretical Relationship between Memory and Metamemory Monitoring Processes at Encoding or Retrieval Across the Lifespan*



Note. The actual memory capabilities represent the average area-under-the-curve (AUC) metric from the tests of working memory and long-term memory of the present study; error bars represent +/- 1 standard error of the mean. See online article for color version of figure.

Conclusions

In a lifespan sample, we found that there was a generally universal relation between individuals' retrospective confidence ratings and their accuracy in recognition, both in WM and episodic LTM. That is, children aged 6 to 13, young adults aged 18 to 27, and older adults aged 65 to 77 all exhibited similar confidence-accuracy relations, whereby item recognition accuracy

was almost always highest when participants expressed *high* confidence in their responses and lowest when they expressed *low* confidence. Despite this largely invariant confidence-accuracy relation that was found in both WM and LTM, individuals of different age groups were not equally adept at adjusting their confidence ratings in accord with their memory signals. Young children were the least accurate when expressing *high* confidence in recognition tests of WM, and older adults were most prone to *high-confidence* errors in recognition tests of LTM when judging items to be "old." These results provide novel insights into the *memory-metamemory relation* and its progression across the lifespan and also may point to different sources of memory errors for children and older adults, relative to younger adults. Compared with young adults, young children appear to be impaired in memory self-monitoring during encoding, whereas older adults appear to be impaired in monitoring their retention of information in the long-term.

Constraints on Generality

Although the present study informs our understanding of lifespan similarities and differences in the confidence-accuracy relation in recognition memory, whether these findings extend to other types of stimuli or settings remains to be determined in future research. For instance, we cannot ascertain from the present study whether the same findings would be obtained with stimuli from a different modality (e.g., remembering what was said by different people) or if participants were tasked with learning information that they may deem more important to retain (e.g., remembering *who* did *what*). In addition, although our lifespan approach included participants from three major periods of the lifespan (childhood, young adulthood, and older adulthood), we did not have participants from middle adulthood (e.g., age 40 to 60). Thus, there is an important gap in our current understanding of the confidence-

accuracy relation in WM and LTM, as it is unknown how this relation is manifest during middle age. Finally, based on our data it is not possible to disentangle several different reasons for the declines in performance between WM and LTM: they could be attributable to interference across trials, temporal decay, or retrieval-induced forgetting if the items tested in LTM were weakened by other items having been tested in WM. Nevertheless, not knowing why memory declines between WM and LTM does not impact our ability to measure the relation between accuracy and confidence in the two memory phases.

References

Adam, K. C. S., & Vogel, E. K. (2017). Confident failures: Lapses of working memory reveal a metacognitive blindsight. *Attention, Perception, & Psychophysics, 79*(5), 1506-1523. https://doi.org/10.3758/s13414-017-1331-8

Allwood, C. M., Jonsson, A.-C., & Granhag, P. A. (2005). The Effects of Source and Type of Feedback on Child Witnesses' Metamemory Accuracy. *Applied Cognitive Psychology, 19*(3), 331–344. https://doi.org/10.1002/acp.1071

Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsycholigal Review, 16*, 17–42. https://doi.org/10.1007/s11065-006-9002-x

Applin, J. B., & Kibbe, M. M. (2021). Young children monitor the fidelity of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(5), 808-819. https://doi.org/10.1037/xlm0000971

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation, 2*(4), 89-195. https://doi.org/10.1016/S0079-7421(08)60422-3

Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science, 11*(5), 750-760. https://doi.org/10.1111/j.1467-7687.2008.00725.x

Barber, S. J., & Mather, M. (2014). Stereotype threat in older adults: When and why does it occur, and who is most affected? In P. Verhaegen & C. Hertzog (Eds.), *The Oxford handbook of emotion, social cognition, and problem solving during adulthood* (pp. 302–320). Oxford, UK: Oxford University Press.

Bartsch, L. M., Loaiza, V. M., & Oberauer, K. (2019). Does limited working memory capacity underlie age differences in associative long-term memory? *Psychology and Aging, 34*(2), 268–281. https://doi.org/10.1037/pag0000317

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bona, S., & Silvanto, J. (2014). Accuracy and confidence of visual short-term memory do not go hand-in-hand: Behavioral and neural dissociations. *PloS One*, *9*(3), e90808. https://doi.org/10.1371/journal.pone.0090808

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*(4), 691–695. https://doi.org/10.1037/0021-9010.72.4.691

Brainerd, C. J., & Reyna, V. F. (2015). Fuzzy-trace theory and lifespan cognitive development. *Developmental Review, 38*, 89-121. https://doi.org/10.1016/j.dr.2015.07.006

Brainerd, C. J., Reyna, V. F., & Forrest, T. J. (2002). Are young children susceptible to the false-memory illusion? *Child Development, 73*(5), 1363–1377. https://doi.org/10.1111/1467-8624.00477

Brewer, N., & Day, K. (2005). The confidence-accuracy and decision latency-accuracy relationships in children's eyewitness identification. *Psychiatry, Psychology and Law, 12*(1), 119–128. https://doi.org/10.1375/pplt.2005.12.1.119

Brockmole, J. R., & Logie, R. H. (2013). Age-related change in visual working memory: A study of 55,753 participants aged 8 –75. *Frontiers in Psychology, 4*, 12. https://doi.org/10.3389/fpsyg.2013.00012

Brubaker, M. S., & Naveh-Benjamin, M. (2018). The effects of stereotype threat on the associative memory deficit of older adults. *Psychology and Aging, 33*(1), 17–29. https://doi.org/10.1037/pag0000194

Bunnell, J. K., Baken, D. M., & Richards-Ward, L. A. (1999). The effect of age on metamemory for working memory. *New Zealand Journal of Psychology, 28*(1), 23-29.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48. https://doi.org/10.3758/BF03210724

Cabeza, R., Albert, M., Belleville, S., Craik, F. I. M., Duarte, A., Grady, C. L., Lindenberger, U., Nyberg, L., Park, D. C., Reuter-Lorenz, P. A., Rugg, M. D., Steffener, J., & Rajah, M. N. (2018). Maintenance, reserve and compensation: the cognitive neuroscience of healthy ageing. *Nature Reviews: Neuroscience*, *19*(11), 701–710. https://doi.org/10.1038/s41583-018-0068-2

Cabeza, R., & Dennis, N. A. (2012). Frontal lobes and aging. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (2nd ed., pp. 628 – 653). New York, NY: Oxford University Press.

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32*(3), 243-258. https://doi.org/10.1037/pag0000168

Courage, M.L., & Cowan, N. (eds) (2022). *The development of memory in infancy and childhood*. 2nd Edition. Routledge.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114. https://doi.org/10.1017/S0140525X01003922

Cowan N. (2016). Working Memory Maturation: Can We Get at the Essence of Cognitive Growth? *Perspectives on Psychological Science, 11*(2), 239–264. https://doi.org/10.1177/1745691615621279

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

Cowan, N. (2019). Short-term memory based on activated long-term memory: A review in response to Norris (2017). *Psychological Bulletin, 145*(8), 822–847. https://doi.org/10.1037/bul0000199

Cowan, N., & Alloway, T. P. (2008). The development of working memory in childhood. In M. Courage & N. Cowan (Eds.), *Development of memory in infancy and childhood* (2nd Edition, pp. 303-342). New York, NY: Routledge.

Cowan, N., Bao, C., Bishop-Chrzanowski, B. M., Costa, A. N., Greene, N. R., Guitard, D., Li, C., Musich, M. L., & Unal, Z. E. (2024). The relation between attention and memory. *Annual Review of Psychology, 75*(1). Advance online publication. https://doi.org/10.1146/annurev-psych-040723-012736

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory

and cognitive aptitudes. *Cognitive Psychology, 51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Cowan, N., Hardman, K., Saults, J. S., Blume, C. L., Clark, K. M., & Sunday, M. A. (2016). Detection of the number of changes in a display in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(2), 169–185. https://doi.org/10.1037/xlm0000163

Cowan, N., Li, Y., Glass, B. A., & Saults, J. (2018). Development of the ability to combine visual and acoustic information in working memory. *Developmental Science, 21*, e12635. http://doi.org/10.1111/desc.12635

Cowan, N., Naveh-Benjamin, M., Kilb, A., & Saults, J. S. (2006). Life-span development of visual working memory: When is feature binding difficult? *Developmental Psychology, 42*(6), 1089. https://doi.org/10.1037/0012-1649.42.6.1089

Craik, F. I. M., & Byrd, M. (1982). Aging and cognitive deficits: The role of attentional resources. In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes* (pp. 191–211). Plenum Press.

Dodson, C. S. (2017). Aging and memory. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp. 403–421). Oxford, UK: Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21053-5

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22*(1), 122-133. https://doi.org/10.1037/0882-7974.22.1.122

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30*(1), 113-125. https://doi.org/10.1002/acp.3178

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychonomic Bulletin & Review, 13*(5), 770–775. https://doi.org/10.3758/bf03193995

Dunlosky, J., & Tauber, S. K. (2016). *The Oxford Handbook of Metamemory.* Oxford University Press: New York, NY.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Technical Note AFCRC-TN-58-51). Bloomington: Indiana University, Hearing and Communication Laboratory.

Fandakova, Y., Shing, Y. L., & Lindenberger, U. (2013). Differences in binding and monitoring mechanisms contribute to lifespan age differences in false memory. *Developmental Psychology, 49*(10), 1822–1832. https://doi.org/10.1037/a0031361

Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the life span: A meta-analysis of age differences. *Psychological Bulletin, 141*(6), 1228–1265. https://doi.org/10.1037/bul0000013

Flavell, J. H. (1971). Stage-related properties of cognitive development. *Cognitive Psychology, 2*(4), 421-453. https://doi.org/10.1016/0010-0285(71)90025-9

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology, 1*, 324-340. https://doi.org/10.1016/0010-0285(70)90019-8

Flavell, J. H., Miller, P., & Miller, S. (1993). *Cognitive development*. Englewood Cliffs, NJ: Prentice Hall.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Forsberg, A., Blume, C., & Cowan, N. (2021a). The development of metacognitive accuracy in working memory across childhood. *Developmental Psychology, 57*(8), 1297-1317. https://doi.org/10.1037/dev0001213

Forsberg, A., Guitard, D., Adams, E. J., Pattanakul, D., & Cowan, N. (2022a). Children's long-term retention is directly constrained by their working memory capacity limitations. *Developmental Science, 25*(2), e13164. https://doi.org/10.1111/desc.13164

Forsberg, A., Guitard, D., Adams, E. J., Pattanakul, D., & Cowan N. (2023). Working memory constrains long-term memory in children and adults: Memory of objects and bindings. *Journal of Intelligence, 11*(5): 94. https://doi.org/10.3390/jintelligence11050094

Forsberg, A., Guitard, D., & Cowan, N. (2021b). Working memory limits severely constrain long-term retention. *Psychonomic Bulletin & Review, 28*(2), 537-547. https://doi.org/10.3758/s13423-020-01847-z

Forsberg, A., Guitard, D., Greene, N. R., Naveh-Benjamin, M., & Cowan, N. (2022b). The proportion of working memory items recoverable from long-term memory remains fixed despite adult aging. *Psychology and Aging, 37*(7), 777–786. https://doi.org/10.1037/pag0000703

Fraundorf, S. H., Hourihan, K. L., Peters, R. A., & Benjamin, A. S. (2019). Aging and recognition memory: A meta-analysis. *Psychological Bulletin, 145*(4), 339–371. https://doi.org/10.1037/bul0000185

Fukuda, K., & Vogel, E. K. (2019). Visual short-term memory capacity predicts the "bandwidth" of visual long-term memory encoding. *Memory & Cognition, 47*(8), 1481-1497. https://doi.org/10.3758/s13421-019-00954-0

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology, 18*(1), 1–16. https://doi.org/10.1002/acp.934

Ghetti, S., Lyons, K. E., Lazzarin, F., & Cornoldi, C. (2008). The development of metamemory monitoring during retrieval: The case of memory strength and memory absence. *Journal of Experimental Child Psychology, 99*(3), 157–181. https://doi.org/10.1016/j.jecp.2007.11.001

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., Paus, T., Evans, A. C., & Rapoport, J. L. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2*(10), 861–863. https://doi.org/10.1038/13158

Gilchrist, A. L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller chunks in older adults. *Memory, 16*(7), 773–787. https://doi.org/10.1080/09658210802261124

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2*(1), 21–31. https://doi.org/10.1037/0278-7393.2.1.21

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition, ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition, 8*(2), 233–243. https://doi.org/10.1037/h0101835

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greene, N. R., Chism, S., & Naveh-Benjamin, M. (2022). Levels of specificity in episodic memory: Insights from response accuracy and subjective confidence ratings in older adults and in younger adults under full or divided attention. *Journal of Experimental Psychology: General, 151*(4), 804-819. https://doi.org/10.1037/xge0001113

Greene, N. R., Forsberg, A., Guitard, D., Cowan, N., & Naveh-Benjamin, M. (2023, September 28). WM-LTM metamemory. Retrieved from osf.io/vfe9g

Greene, N. R., & Naveh-Benjamin, M. (2022). Adult age differences in specific and gist associative episodic memory across short- and long-term retention intervals. *Psychology and Aging, 37*(6), 681–697. https://doi.org/10.1037/pag0000701

Greene, N. R., & Naveh-Benjamin, M. (2023). Adult age-related changes in the specificity of episodic memory representations: A review and theoretical framework. *Psychology and Aging, 38*(2), 67–86. https://doi.org/10.1037/pag0000724

Greene, N. R., Naveh-Benjamin, M., & Cowan, N. (2020). Adult age differences in working memory capacity: Spared central storage but deficits in ability to maximize peripheral storage. *Psychology and Aging, 35*(6), 866-880. https://doi.org/10.1037/pag0000476

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in*

research and theory (Vol. 22, pp. 193–225). Academic Press. https://doi.org/10.1016/S0079-7421(08)60041-9

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review, 123*(1), 23–69. https://doi.org/10.1037/rev0000015

Henkel, L. A., Johnson, M. K., & De Leonardis, D. M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General, 127*(3), 251–268. https://doi.org/10.1037/0096-3445.127.3.251

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 417–466). Mahwah, NJ: Lawrence Erlbaum Associates.

Hertzog, C., Sinclair, S. M., & Dunlosky, J. (2010). Age differences in the monitoring of learning: Cross-sectional evidence of spared resolution across the adult life span. *Developmental Psychology, 46*(4), 939–948. https://doi.org/10.1037/a0019812

Hiller, R. M., & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory & Cognition, 2*(3), 185–191. https://doi.org/10 .1016/j.jarmac.2013.07.001

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*(7), 871–893. https://doi.org/10.1002/acp.1302

Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language, 52*, 494-504. https://doi.org/10.1016/j.jml.2005.01.007

Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: Aging, attention, and control. *Psychology and Aging, 8*(2), 283-294. https://doi.org/10.1037/0882-7974.8.2.283

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging, 12*(2), 352-361. https://doi.org/10.1037/0882-7974.12.2.352

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304-1316. https://doi.org/10.1037/0278-7393.22.5.1304

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology, 97*(4), 286–314. https://doi.org/10.1016/j.jecp.2007 .01.007

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48*(4), 704-721. https://doi.org/10.1016/S0749-596X(02)00504-1

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young children's sensitivity to their own ignorance in informing others. *PLoS ONE, 11*(3), e0152595. https://doi.org/10.1371/journal.pone.0152595

Knutsson, J., & Allwood, C. M. (2014). Opinions of legal professionals: Comparing child and adult witnesses' memory report capabilities. *The European Journal of Psychology Applied to Legal Context, 6*(2), 79-89. https://doi.org/10.1016/j.ejpal.2014.06.001

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.609

Koriat, A., Goldsmith, M., Schneider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*(4), 405–437. https://doi.org/10.1006/jecp.2000.2612

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models." *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lachman, J. L., Lachman, R., & Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology, 15*(5), 543–551. https://doi.org/10.1037/0012-1649.15.5.543

Lenth, R. (2023). emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.8. Retrieved from https://CRAN.R-project.org/package=emmeans

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology, 71*(6), 1092-1107. https://doi.org/10.1037/0022-3514.71.6.1092

Light, 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

Light, L. L., & Anderson, P. A. (1985). Working-memory capacity, age, and memory for discourse. *Journal of Gerontology, 40*(6), 737-747. https://doi.org/10.1093/geronj/40.6.737

Lindsay, D. S., Johnson, M. K., & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology, 52*(3), 297–318. https://doi.org/10.1016/0022-0965(91)90065-Z

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9*(3), 215–218. https://doi.org/10.1111/1467-9280.00041

Lindsay, R. C. L., Wells, G., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*(1), 79-89. https://doi.org/10.1037/0021-9010.66.1.79

Liu, Y., Su, Y., Xu, G., & Pei, M. (2018). When do you know what you know? The emergence of memory monitoring. *Journal of Experimental Child Psychology, 166,* 34-48. https://doi.org/10.1016/j.jecp.2017.06.014

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Luo, L., & Craik, F. I. M. (2009). Age differences in recollection: Specificity effects at retrieval. *Journal of Memory and Language, 60*(4), 421-436. https://doi.org/10.1016/j.jml.2009.01.005

Marquié, J. C., & Huet, N. (2000). Age differences in feeling-of-knowing and confidence judgments as a function of knowledge domain. *Psychology and Aging, 15*(3), 451–461. https://doi.org/10.1037/0882-7974.15.3.451

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

Metcalfe, J., & Dunlosky, J. (2008). Metamemory. In J. H. Bryne (Ed.), *Learning and Memory: A Comprehensive Reference*, Vol. 2 (pp. 349-362). New York: Academic Press. https://doi.org/10.1016/B978-012370509-9.00159-5

Mickes, L., (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93-102. https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140*(2), 239-257. https://doi.org/10.1037/a0023007

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology*, *36*(2), 185-193. https://doi.org/10.1093/geronj/36.2.185

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1170–1187. https://doi.org/10.1037/0278-7393.26.5.1170

Naveh-Benjamin, M. & Cowan, N. (2023). Age-related changes in working memory: Roles of attention, executive function, and knowledge. *Nature Reviews Psychology, 2*, 151-165. https://doi.org/10.1038/s44159-023-00149-0.

Naveh-Benjamin, M., & Old, S. R. (2008). Aging and memory. In J. H. Byrne, H. Eichenbaum, R. Menzel, H. L. Roediger, & D. Sweatt (Eds.), *Learning and memory: A comprehensive Reference* (pp. 787-808). Oxford, UK: Elsevier.

Nelson, T.O., & Narens, L. (1990) Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.) *The Psychology of Learning and Motivation*, Vol. 26, (pp. 125-173). New York: Academic Press.

Newcombe, P. A., & Bransgrove, J. (2007). Perceptions of witness credibility: Variations across age. *Journal of Applied Developmental Psychology, 28*(4), 318–331. https://doi.org/10.1016/j.appdev.2007.04.003

Old, S., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23*(1), 104-118. https://doi.org/10.1037/0882-7974.23.1.104

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92,* 101-126. https://doi.org/10.1016/j.cogpsych.2016.11.007

Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval,

and divided attention. *Journal of Experimental Psychology: Applied, 19*(1), 55–71. https://doi.org/10.1037/a0031602

Perlmutter, M. (1978). What is memory aging the aging of? *Developmental Psychology, 14*(4), 330–345. https://doi.org/10.1037/0012-1649.14.4.330

Pliske, R. M., & Mutter, S. A. (1996). Age differences in the accuracy of confidence judgments. *Experimental Aging Research, 22*(2), 199-216. https://doi.org/10.1080/03610739608254007

Powell, M. B., Garry, M., & Brewer, N. (2013). Eyewitness testimony. In I. Freckelton & H. Selby (Eds.), *Expert evidence: Law, practice, procedure and advocacy* (5th ed.). Thomson Reuters.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*(1), 96-111. https://doi.org/10.1016/0022-0965(87)90053-1

Riggs, K. J., McTaggart, J., Simpson, A., & Freeman, R. P. (2006). Changes in the capacity of visual working memory in 5-to 10-year-olds. *Journal of Experimental Child Psychology, 95*(1), 18–26. https://doi.org/10.1016/j.jecp.2006.03.009

Rhodes, S., & Cowan, N. (2018). Attention in working memory: attention is needed but it yearns to be free. *Annals of the New York Academy of Sciences*, *1424*(1), 52–63. https://doi.org/10.1111/nyas.13652

Rhodes, S., Greene, N. R., & Naveh-Benjamin, M. (2019). Age-related differences in recall and recognition: A meta-analysis. *Psychonomic Bulletin & Review, 26*(5), 1529-1547. https://doi.org/10.3758/s13423-019-01649-y

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal-neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences, 17*, 114-123. https://doi.org/10.1016/j.cobeha.2017.07.016

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*(6), 1052-1067. https://doi.org/10.1037/0012-1649.38.6.1052

Roebers, C. M., & Howie, P. (2003). Confidence judgments in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85*(4), 352-371. https://doi.org/10.1016/S0022-0965(03)00076-6

Salthouse, T. A. (2014). Correlates of cognitive change. *Journal of Experimental Psychology: General, 143*(3), 1026-1048. https://doi.org/10.1037/a0034847

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27*(5), 763– 776. http://doi.org/10.1037/0012-1649.27.5.763

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. https://doi.org/10.3758/BF03209391

Schneider, W. (1985). Developmental trends in the metamemory-memory behavior relationship: An integrative review. In D. L. Forrest-Pressley, G. E. MacKinnon, & T. G. Waller (Eds.), *Cognition, Metacognition, and Human Performance* (Vol. 1, pp. 57- 109). San Diego, CA: Academic Press.

Schneider, W., & Pressley, M. (2013). *Memory development between two and twenty.* New York, NY: Psychology Press.

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400–415. https://doi.org/10.1037/xap0000157

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S. C., & Lindenberger, U. (2010). Episodic memory across the lifespan: the contributions of associative and strategic components. *Neuroscience and Biobehavioral Reviews*, *34*(7), 1080–1091. https://doi.org/10.1016/j.neubiorev.2009.11.002

Shing, Y. L., Werkle-Bergner, M., Li, S. C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*(2), 169-179. https://doi.org/10.1080/09658210802190596

Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology, 111*(4), 695–707. https://doi.org/10.1016/j.jecp.2011.10.007

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage model. *Memory & Cognition, 33*, 1116-1129. https://doi.org/10.3758/BF03193217

Sowell, E. R., Delis, D., Stiles, J., & Jernigan, T. L. (2001). Improved memory functioning and frontal lobe maturation between childhood and adolescence: A structural MRI study. *Journal of the International Neuropsychological Society, 7*(3), 312–322. https://doi.org/10.1017/S135561770173305X

Stark, S. M., Yassa, M. A., Lacy, J. W., & Stark, C. E. L. (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia, 51*(12), 2442–2449. https://doi.org/10.1016/j.neuropsychologia.2012.12.014

Stine-Morrow, E. A. L., Shake, M. C., Miles, J. R., & Noh, S. R. (2006). Adult age differences in the effects of goals on self-regulated sentence processing. *Psychology and Aging, 21*(4), 790-803. https://doi.org/10.1037/0882-7974.21.4.790

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379–1410. https://doi.org/10.1037/0278-7393.24.6.1379

Stuss, D. (2011). Functions of the frontal lobes: Relation to executive functions. *Journal of the International Neuropsychological Society, 17*(5), 759-765. https://doi.org/10.1017/S1355617711000695

Suchow, J. W., Fougnie, D., & Alvarez, G. A. (2017). Looking inward and back: Real-time monitoring of visual working memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 660–668. https://doi.org/10.1037/xlm0000320

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293. https://doi.org/10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26. https://doi.org/10.1111/1529-1006.001

Taylor, J. E. (1958). *Selected writings of John Hughlings Jackson* (Vol. 2). London, UK: Staples.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*, 49. https://doi.org/10.1186/s41235-017-0086-z

Tulving, E. (1983). *Elements of episodic memory.* Oxford, England: Clarendon Press.

Vandenbroucke, A. R., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science, 25*(4), 861-873. https://doi.org/10.1177/0956797613516146

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin, 120*(2), 272–292. https://doi.org/10.1037/0033-2909.120.2.272

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. https://doi.org/10.1167/4.12.11

Wingfield, A., Stine, E. A., Lahar, C. J., & Aberdeen, J. S. (1988). Does the capacity of working memory change with age? *Experimental Aging Research, 14*, 103–107. https://doi.org/10.1080/03610738808259731

Winsor, A. A., Flowe, H. D., Seale-Carlisle, T. M., Killeen, I. M., Hett, D., Jores, T., Ingham, M., Lee, B. P., Stevens, L. M., & Colloff, M. F. (2021). Child witness expressions of certainty are informative. *Journal of Experimental Psychology: General, 150*(11), 2387–2407. https://doi.org/10.1037/xge0001049

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235-269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Wixted, J. T., & Mickes, L. (2022). Eyewitness memory is reliable, but the criminal justice system is not. *Memory, 30*(1), 67-72. https://doi.org/10.1080/09658211.2021.1974485

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515–526. https://doi.org/10.1037/a0039510

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Yassa, M. A., Mattfield, A. T., Stark, S. M., & Stark, C. E. (2011). Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America, 108*(21), 8873-8878. https://doi.org/10.1073/pnas.1101567108

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

Zacks, R. T., Hasher, L., & Li, K. Z. H. (2000). Human memory. In T. A. Salthouse & F. I. M. Craik (Eds.), *Handbook of aging and cognition*, 2nd edition (pp. 293-357). Mahwah, NJ: Lawrence Erlbaum.