



# Age-related differences in recall and recognition: a meta-analysis

Stephen Rhodes<sup>1,2</sup> · Nathaniel R. Greene<sup>2</sup> · Moshe Naveh-Benjamin<sup>2</sup>

Published online: 8 August 2019  
© The Psychonomic Society, Inc. 2019

## Abstract

Relative to younger adults, older adults tend to perform more poorly on tests of both free recall and item recognition memory. The age difference in performance is typically larger for recall tasks relative to those involving recognition. However, there have been reports of comparable age-related differences in free recall and item recognition performance. Further, a differential performance cost does not necessarily mean that processes involved in recall are specifically affected by age. Here we present a meta-analysis of 36 articles reporting 89 direct comparisons of free recall and item recognition in younger and older groups of participants. Standardized effect sizes reveal that age differences are larger for recall tasks (Hedges'  $g = 0.89$ , 95% confidence intervals [0.75, 1.03]) than for recognition tasks (0.54, [0.37, 0.72]). Further, Brinley analyses of the data suggest that distinct functions are needed to relate younger and older performance for the two tasks. These functions differ in intercept pointing to a disproportionate age difference in recall relative to recognition. This is in line with theories of memory and aging which posit specific deficits in processes related to search and retrieval from memory.

**Keywords** Recall · Recognition · Episodic memory · Cognitive aging

Age differences in memory tasks that require participants to recall items (e.g., words) from a studied list are typically larger than when the task requires recognition of studied items among unstudied lures. However, what this tells us about age-related changes in memory functioning is perhaps less clear. In other words, this age by task interaction may be consistent with a general explanation (e.g., decline in memory fidelity) rather than an explanation requiring a specific age-effect on an underlying process (e.g., search and retrieval from memory). There were two main aims of the present work: (1) to quantify age differences in free recall and item recognition in studies that have directly compared the two; (2) to use Brinley analysis to examine whether a single function can explain the relation between older and younger performance or if different functions are needed for recall and recognition, suggesting a specific deficit.

---

Stephen Rhodes and Nathaniel R. Greene contributed equally to this paper.

✉ Stephen Rhodes  
srhodes@research.baycrest.org

<sup>1</sup> Rotman Research Institute, Baycrest, 3560 Bathurst Street, Toronto, ON M6A 2E1, Canada

<sup>2</sup> University of Missouri, Columbia, MO 65211, USA

## Are age differences larger for recall than recognition?

Early work on the question of whether age differences are larger for free recall or item recognition proved to be rather mixed. Schonfield and Robertson (1966) found no age differences in a five-choice recognition task but a clear age-related effect on the recall of lists of 24 words. Harwood and Naylor (1969) also found greater age differences in recalling pictures of common objects relative to recognizing them, although in this case the latter age difference was also greater than zero (see also Craik & McDowd, 1987; Erber et al., 1980). On the other hand, White and Cunningham (1982) found that, when a correction for guessing was applied to recognition scores, only a main effect of age remained and no task interaction, suggesting similar age differences in recall and recognition (see also Botwinick & Storandt, 1980; Verhaeghen et al., 1998).

Since these initial studies, many more comparisons have been conducted on younger and older adults performing free recall or item recognition tasks. This has resulted in a fairly convincing case that (1) there are typically non-zero age differences in item recognition (see Fraundorf et al., 2019; Old & Naveh-Benjamin, 2008 for meta-analyses) and (2) age differences are larger for recall than they are for recognition (e.g., Danckert & Craik, 2013; Nyberg et al., 2003; Whiting et al., 1997). For example, in a recent series

of experiments comparing recall and recognition directly, Danckert and Craik (2013) aimed to rule out the criticism (attributed to Uttl et al., 2007) that different age trajectories of the two tasks could largely be explained by ceiling effects in recognition performance. In their experiment 1, for example, participants studied ten lists of 20 nouns and after each list, following a 30-s interval filled with backwards counting, were asked to recall as many items as they could. After all ten lists had been presented, each with its corresponding free recall test, a 5-min break was given before participants were administered a yes-no recognition test with the 200 words from all ten lists randomly intermixed with 200 unstudied lures. This large set size of items was constructed to prevent recognition scores from reaching ceiling. Importantly, even with the possibility of ceiling-level performance reduced in the recognition task, there was an age by task interaction, such that age differences in recall accuracy were larger than differences in corrected recognition scores (hits minus false alarms). Moreover, this interaction remained when the young adults who performed the highest (i.e., closest to ceiling) and the old adults who performed the lowest (i.e., closest to floor) were removed, confirming that disproportionate age differences in recall cannot be explained by ceiling effects in recognition alone.

While the literature might have essentially reached consensus on the idea that age differences in free recall are greater than item recognition, to our knowledge, no one has synthesized results of studies to measure just how big the discrepancy is. This was one aim of the present meta-analysis. After all, a detailed account of episodic memory and aging should aim to explain the magnitude of age differences rather than the mere presence of an interaction effect. A further aim was to examine study characteristics that might modulate age differences in recall and recognition performance.

## What does a differential age difference in recall relative to recognition tell us?

It could be that a specific process underlying recall is impaired with age or, given that age differences are found for both tasks, there could be a common underlying cause that appears differential given the transformations from underlying construct to observed performance levels (see Loftus, 1978; Rhodes et al., 2018; Salthouse, 2000). Both general and specific accounts seem viable. For example, in studies of item recognition memory and aging, it has often been suggested that older adults are impaired in recollecting specific events and instead rely more on a general feeling of familiarity (e.g., Howard et al., 2006; Jennings & Jacoby, 1993). Recollection is a process that is theoretically closely related to recall (Atkinson & Juola, 1974; Mandler, 1980) and, therefore, it is possible that both age differences in

recall and recognition are driven by a common problem in retrieving contextual details, which is far more important in recall tasks (cf. Yonelinas, 2002). However, the involvement of an additional recollection process in item recognition has been strongly contested (e.g., Dunn, 2004; Haaf et al., 2018; Pratte & Rouder, 2012). Therefore, if item recognition decisions are largely made on a single “familiarity” metric, the larger age-effect of recall could reflect a specific deficit of retrieving the contextual details (see Spencer & Raz, 1995 for a meta-analysis) that, according to many models of memory, are needed to guide retrieval (Raaijmakers & Shiffrin, 2002). Given the current debate regarding the involvement of recollection in item recognition, we are certainly unable to rule out the proposition that deficits in recollection contribute to *both* recognition and, to a greater extent, recall.

There are other ways in which a general age-related change could manifest itself as different age-effects for recall and recognition. For example, Craik (1983) arranged free recall and item recognition on a continuum in the extent to which they require observers to engage in self-initiated processes, with the former requiring much more effort. Assuming a general decline in “attentional resources” would, therefore, produce a larger age difference in tasks requiring more self-initiation, like recall. Further, the ability for the memory task to constrain participants’ responses to specific items from the encoding phase may be reduced in free recall relative to item recognition tasks, in line with a specificity principle of memory (Surprenant and Neath, 2009). Because free recall tasks require access to specific information in memory but provide participants with less specific cues to that information, whereas item recognition tasks provide more specific cues (e.g., the original study item), age differences in the two tasks could be related to a general underlying factor, namely a difficulty accessing the verbatim memory trace (Brainerd & Reyna, 2015). Relatedly, a general age-related decline in memory fidelity could reduce the overall strength of memories (e.g., Benjamin, 2010; Li et al., 2005). As recognition can proceed with even weak memories, this would also conceivably cause a larger age difference for recall.

In contrast to these general deficit accounts, the recent impressive modeling efforts of Healey and Kahana (2016) suggest that there may be specific deficits underlying age differences in recall. They identified four factors that contribute to age-related changes in recall and recognition, most of which apply to both tasks. To model their older adult data, Healey and Kahana (2016) had to vary parameters relating to the stability of attention during the encoding of items, the use of context to guide recall attempts (also used in recognition in an analogous fashion to recollection, see above), and the screening of memory for intrusions (which would prevent false alarms in recognition tasks).

While these three factors contribute in a similar fashion to age-related differences in recall and recognition, there was an additional set of age-related parameters that applied specifically to recall. Namely, assuming that retrieval computations are noisier and susceptible to competition in old age was found to improve model fit. Thus, while this model assumes that age differences in free recall and item recognition are largely due to common underlying processes, it also provides reason to suspect that the effect of age on free recall would be disproportionate.

The way the distinction between general and specific accounts of phenomena has typically been addressed in meta-analyses of the cognitive aging literature is by plotting the relationship between younger and older performance across the range of conditions and studies included (known as a “Brinley plot”, after Brinley, 1965). The rationale is that if a single function is sufficient to explain the entire collection of points—that is, if it is possible to predict the performance of older adults in one task given the performance of younger adults in another task—an explanation that proposes a specific deficit is needlessly complex. For example, if age differences in recall and recognition tasks vary on a continuum related to their demand for self-initiated processes (Craik, 1983), then we would expect a single Brinley function with a slope greater than one (i.e., age differences are greatest for tasks younger adults find more difficult and presumably demand more self initiation).

Alternatively, if different functions are required for the different tasks, this points to age-related deficits in specific processes implicated in one task but not the other (see Verhaeghen, 2013 for more detail on Brinley analysis). Brinley plots have been most widely applied to response times to test general slowing theories of cognitive aging but they have also been applied to accuracy (e.g., Bopp & Verhaeghen, 2005; Verhaeghen et al., 2003). For reaction time data, specific interpretations have been offered for the intercept and slope terms (e.g., Cerella, 1985; Myerson et al., 1990; Verhaeghen, 2013). While, to our knowledge, no such accounts exist for accuracy data, we may speculate on what possible outcomes would tell us about the underlying cognitive processes. Two Brinley functions differing in intercept and slope may point to age differences between the two tasks that scale with the level of difficulty (as indexed by younger adult accuracy). For example, in models of recall, it is often assumed that memories are reconstructed from a fragmented trace cued by a particular context and that the context used to probe memory evolves with each successful recall attempt (see, e.g., Raaijmakers & Shiffrin, 2002). An age-related deficit in the process of reconstructing memory traces could produce a cascading effect, such that age differences are magnified for more difficult recall tasks (i.e., those placing a greater demand on reconstruction; for a similar point, see Brainerd et al., 2009). This would produce

a Brinley for recall with a lower intercept than that for recognition but a steeper slope.

On the other hand, parallel Brinley functions that differ in intercept and not slope would imply a constant accuracy cost for older adults in free recall relative to item recognition. Such a constant cost may imply a different rule for the termination of memory search during recall (Raaijmakers & Shiffrin, 2002) or possibly a greater susceptibility to intrusions due to competition between recall candidates (Healey & Kahana, 2016). In any case, distinct functions for recall and recognition, combined with a larger overall age difference for the former, would implicate a specific age-related deficit in processes surrounding the search and retrieval of information from memory. Therefore, in addition to comparing overall age differences in free recall and item recognition, and exploring potential moderating factors, we also model the Brinley plot to explore the nature of age differences across these two tasks.

Here we have chosen to focus on articles that have reported *direct* comparisons of recall and recognition, to try and reduce the influence of other methodological factors that are not related to the mode of testing memory. Also, we have chosen to focus on measures of memory for individual *items* and not for associations between items, for which there is ample evidence of an age-related deficit relative to item memory (e.g., Ahmad et al., 2015; Naveh-Benjamin, 2000; Naveh-Benjamin et al., 2007; Old & Naveh-Benjamin, 2008).<sup>1</sup> Thus the present meta-analysis does not include tasks where the goal is to associate pairs of distinct items in memory (for example, associative recognition or cued recall; see Old & Naveh-Benjamin, 2008 for such a meta-analysis). Before outlining our approach in more detail, it is worth discussing some previous meta-analyses that are related to the present one.

## Previous meta-analyses

In their meta-analysis on aging and repetition priming, La Voie and Light (1994) also conducted a meta-analysis on recall and recognition tasks. They identified 18 recognition observations, with a standardized age difference of 0.497 [0.353, 0.641] (95% confidence intervals), and 18 recall observations producing an age difference of 0.968 [0.835, 1.101]. On closer inspection, it appears that 12 out of the 18 recall observations were cued recall tasks. As noted above, in the present work we have chosen to omit cued recall given its requirement to associate two distinct items, an operation

<sup>1</sup>Note that our definition of an “item” is quite broad and includes images, words, actions, and passages of text (e.g., sentences, stories). We describe the stimuli used in the reports we gathered in more detail below and consider this factor as a potential moderator in our analyses.

that is well known to produce a disproportionate age effect (Old & Naveh-Benjamin, 2008). Thus it is possible that some of the larger effect size for recall found by La Voie and Light (1994) could be attributable to the requirement to form associations between items, rather than the mode of testing per se. Light and Singh (1987) were the only study included in the meta-analysis of La Voie and Light (1994) to report data from both a free recall and an item recognition task (three experiments in total) and this study is included in the present meta-analysis.

As previously mentioned, Old and Naveh-Benjamin (2008) conducted a meta-analysis of studies assessing age differences in item and associative memory. One of the moderators they considered was the nature of the memory test, and their Table 4 presents estimates of standardized age differences in item memory for studies split into four categories: those where both item and associative tasks were recognition ( $n = 56$ ), where both were recall ( $n = 7$ ), where the former was recall and the latter recognition ( $n = 9$ ), and vice versa ( $n = 8$ ). For studies where the item task was recognition, the effect size estimates were fairly consistent (0.65 [0.58, 0.71] for studies also using associative recognition and 0.67 [0.50, 0.85] for studies using associative recall). However, when considering age differences in item recall, the effect size estimates were more variable (1.19 [1.03, 1.34] for studies using associative recall and 0.91 [0.72, 1.11] for studies using associative recognition). The variability in recall effect sizes is likely due to the small number of recall observations they identified (as their analysis was on a different topic) and also possibly due to variability between studies in the nature of the tasks used (i.e., they also included both free recall and cued recall paradigms). In the present meta-analysis, to avoid the potential for contamination of effect sizes through procedural differences between recognition and recall studies that are not reflective of differences between the modes of testing memory per se, we focused on studies that have directly compared the two tasks in the same groups of participants using the same general materials and procedures.

Fraundorf et al. (2019) have recently reported a meta-analysis of age differences in item recognition tasks from 232 experiments. Their analysis focused on signal detection theory measures of sensitivity and response criterion and not effect sizes (as they were unable to calculate measures of variability for their outcome measures). Consistent with the findings of Old and Naveh-Benjamin (2008), they report a sizable age difference in recognition sensitivity (0.46 [0.41, 0.51] in  $d'$  units, not taking into account moderators) and more modest, but reliable, differences in response criterion, such that older adults are more likely to respond “new”. In discussing their findings, Fraundorf et al. (2019) note that the robust age difference for item recognition raises important theoretical questions about the nature of age differences; they even go as far as to say

“[...] given that age deficits *do* exist in recognition, it is not necessarily clear that there is a theoretically meaningful division to be drawn between age-related effects on recall and recognition.” They also highlight the need for a meta-analysis directly comparing recall and recognition. Thus the goals of the present work were to synthesize the results of studies that have provided a direct comparison of item free recall and item recognition and to examine whether the extant data are consistent with general deficit explanations of memory and aging or those that suggest additional deficits associated with recall.

## Method

### Search and inclusion criteria

We searched the databases PsycINFO (plus PsychARTICLES), Google Scholar, and PubMed, along with relevant citations included in the sampled literature. Keywords in the searches included combinations of *recall*, *recognition*, *age*, *aging*, *young*, *old*, and *memory*. This search was carried out in September 2017. Our initial query yielded 238 results. The first two authors combed through these articles to determine whether they fit the scope of the meta-analysis and to find any additional sources not included in the initial search. The majority of these studies (132 or 55%) only included recall *or* recognition, but not both. Of the remaining 106 studies, only those which met the following criteria were considered: (a) The study, or experiments within, compared younger adults (with a mean age of 30 or younger) with older adults (with a mean age of 60 or older); (b) At least one experiment within the study included measures of both recall and recognition for individual items and the same or similar material; (c) The recall measure included in the study was free recall (three studies, Murphy et al., 1997; Schramke and Bauer, 1997; Spilich & Voss, 1983, also included measures referred to as cued recall, but their tasks did not require that participants form arbitrary associations between distinct items and only their free recall measure were included); (d) The data were reported in text or in a figure from which reasonable measures of average performance could be extracted (we discuss handling of missing variance estimates below); (e) The study procedures were clearly explained such that adequate information about the material on which participants were tested could be assessed; and (f) The article was written in English. Forty-four experiments from 37 articles satisfied these criteria. One study which reported the number of words recalled minus errors was removed, despite meeting the above criteria (Lalanne et al., 2013), given difficulty comparing this measure to recall accuracy. A table listing the studies included in the meta-analysis is given in the Appendix.

## Description of the included studies

From the resulting 36 articles included in the analyses, there were 89 unique conditions (i.e., 89 recall and 89 recognition observations for both age groups, or 178 of each in total). All recall observations were free recall. For item recognition, 75 observations used the standard “yes-no” (i.e., old-new) format, seven used a four-alternative forced choice format, four used a two-alternative forced choice format, and three used another forced choice format (no studies included in the meta-analysis used a remember-know procedure). On average, 32 young and 32 old adults completed each condition. The average (across conditions) mean age for young adults was 23.11 ( $SD = 3.09$ ) and for old adults, 70.36 ( $SD = 4.90$ ).

We also included information on the following variables to test for potential moderator effects:

1. The *learning instruction* used which assessed whether participants attended to the material under intentional or incidental learning instructions. Sixty-two of the 89 observations used intentional learning instructions, where participants were aware of a forthcoming memory test, and the remaining 27 used incidental learning instructions, where participants were unaware that they would later be tested on their memory. Informing participants of a future memory test can sometimes exacerbate or diminish age differences in performance depending on the nature of the task (e.g., Naveh-Benjamin et al., 2009).
2. The type of *stimuli* that participants had to remember. The majority of observations used words as stimuli ( $n = 52$ ). We categorized 15 observations as using passages of text (including scripts, sentences, statements, stories), seven as assessing memory for actions, 14 as using visual stimuli (pictures, visual matrices), and one study used different odors (which was omitted from the analysis of this moderator).
3. The *mode of presentation* of the to-be-remembered material, which is somewhat related to the above factor of stimuli. Seventy of the observations presented stimuli visually, 12 presented stimuli auditorily, six presented stimuli both visually and auditorily, and for the study assessing odor memory the presentation was olfactory.
4. The *list length* or the number of items that were studied for the subsequent memory test. This varied from 4 to 200, with the majority of observations ( $n = 60$ ) using list lengths of 24 or less. Thus we used log list length in our moderator analysis.
5. The *relatedness* of the studied items. In other words, whether the individual study items could be grouped in some way (e.g., via semantic relatedness) or whether

the individual items were selected to be unrelated. Twenty-nine used related study items and 60 used unrelated items.

6. The *order* in which the recall and recognition tests were administered. Eighty-one of the observations presented the recall test before the recognition test, seven counterbalanced the order, and only one presented the recognition test first. Related to this moderator, we also looked at whether tests of recall and recognition were based on the same study list or different lists. The vast majority of studies examined recall and recognition for the same lists (86 observations). Light and Anderson (1983) and Davis et al. (1990) were the only articles to use separate lists for each task (three observations).
7. To address the concern that smaller age differences in recognition relative to recall may be due to ceiling recognition performance in the younger group (see Danckert and Craik, 2013 for discussion) we considered the recognition accuracy of the younger group as a moderator.
8. Finally, we considered the age of the older group as a potential moderator of effect sizes (see Fraundorf et al., 2019 for the same approach). The average age of the older adult groups varied from 62 to 84. We also considered the difference in mean age between the younger and older groups (which ranged between 31.70 and 61.50) in a separate meta-regression.<sup>2</sup>

Each of these factors could conceivably modulate overall age differences and, more importantly, differences between recall and recognition. In particular, list length, item relatedness, and whether participants study items under intentional or incidental learning conditions could be reasonably expected to affect recall performance more than recognition. For example, assuming that free recall performance can be improved by forming associations between studied items, we may expect recall performance to particularly benefit from shorter list lengths or intentional study. We might also expect younger adults to benefit from this to a greater degree, as older adults appear to be less likely to spontaneously form associations between items (e.g., Naveh-Benjamin et al., 2007), which would increase age differences in recall. For item relatedness, there is evidence that older adults’ recall performance can be improved by presenting related items at study (e.g., Ahmad et al., 2015; Bastin et al., 2013; Naveh-Benjamin et al., 2005). Thus, we might expect related items to produce smaller age differences and possibly especially so for recall tasks, where forming relations between items is more beneficial.

<sup>2</sup>Two studies (Harwood & Naylor, 1969; Schonfield & Robertson, 1966) only reported age range so the mid-point of the range was used in place of the mean.

## Data extraction and processing

Data were extracted from reported estimates either within the text or in a figure or table. For data that were reported only in figures, we used the DataThief program to extract data points from the figures (available at <https://datathief.org>).

In the first step of processing, whenever a study reported multiple groups that fell within our predefined ranges for different age groups (for example, older groups aged 60–69, 70–79) we averaged scores and standard deviations weighted by the size of each group into either younger or older groups. Following this, we attempted to place all scores on a common scale: accuracy or proportion correct. This was fairly simple for recall which was often reported as either accuracy (including proportion or percent of list items recalled out of all items in the study list) or as the number correctly recalled out of a maximum number possible. For recognition, there was more variability in the reported measures requiring some transformation to the accuracy scale (which incorporates both hits and false alarms). Four articles (12 recognition observations) reported  $d'$  as their measure of recognition sensitivity. To transform this to accuracy, we used the following formula: accuracy =  $\Phi(d'/2)$ , where  $\Phi$  is the cumulative distribution function of a standard normal (see Macmillan & Creelman, 2005, p. 9). For standard deviation, as the transformation is non-linear, we took the largest and smallest deviation from mean accuracy implied by the reported  $d'$  standard deviations and averaged them.<sup>3</sup> This transformation makes the assumption of unbiased responding, which is unlikely to be strictly true (see Fraundorf et al., 2019). However, the effect sizes given in an analysis of raw  $d'$  scores (0.74, 95% confidence intervals [0.44, 1.04]) are comparable to those given by analysis of the transformed scores (0.75, [0.45, 1.05]). Three articles (13 recognition observations) reported the measure  $P_r$  (hits minus false alarms) which was converted to accuracy via: accuracy =  $\frac{1}{2}P_r + \frac{1}{2}$  (see Macmillan & Creelman, 2005, p. 7). Given this rescaling the associated standard deviations were divided by 2. One study (Rohling, Ellis, & Scogin, 1991) reported the measure  $P_a$  (McNicol, 1972) and three articles reported only hits (17 recognition observations). We did not perform any transformation on these measures or their associated standard deviations (i.e., they were treated as measures of accuracy).

Seven articles did not report standard deviations or enough information to calculate standard deviations.

<sup>3</sup>Specifically, we used the reported standard deviation on  $d'$  scale to find upper and lower deviations on accuracy scale. Upper =  $\phi(\frac{d'+s}{2}) - m$ , Lower =  $m - \phi(\frac{d'-s}{2})$ , where  $s$  is the standard deviation on  $d'$  scale and  $m$  is mean accuracy (transformed from  $d'$  using the formula in the main text). The average of the upper and lower deviations was used as our measure of standard deviation on accuracy scale.

These articles reported 16 overall observations constituting approximately 17.98% of the data set. To calculate effect sizes for these observations, we decided to interpolate standard deviations in a simple manner. Separately for each age group and each task (recall/recognition) we calculated the typical ratio of standard deviation to mean accuracy and used this ratio to produce standard deviations for observations where they were not reported. Crucially, unless otherwise noted, all of the results reported here hold when restricting analysis to only those articles that reported usable estimates of variance.

## Analysis

### Meta-analysis of age differences

Mean and standard deviation accuracy scores were used to calculate standardized mean differences (Hedges'  $g$ , Hedges, 1981). Effect sizes for age differences in accuracy were synthesized via multilevel mixed effects models implemented using the *metafor* package in R (R Core Team, 2018; Viechtbauer, 2010). These models account for the fact that there is clustering between observations from the same study (e.g., due to sampling, lab procedures). More specifically, effect size  $i$  from study  $j$  was modeled as coming from a normal distribution with known variance:

$$g_{ij} \sim \text{Normal}(\mu_{ij}, v_{ij}),$$

where  $g_{ij}$  is the effect size and  $v_{ij}$  is the sampling variance associated with the effect size. In the base model, the true effects,  $\mu_{ij}$  are modeled as follows:

$$\mu_{ij} = \beta_0 + u_j,$$

where  $u_j \sim \text{Normal}(0, \tau^2)$ . This allows for studies to randomly differ in their underlying effect sizes around the grand mean,  $\beta_0$ . The variability between studies in underlying effect sizes is estimated via the  $\tau$  parameter. This base model can be expanded with additional moderators (e.g., a  $\beta_1$  term) and random effects (in which case a covariance matrix in a multivariate normal distribution replaces  $\tau^2$ ).

### Brinley analysis

Brinley analysis assesses the relationship between the performance of younger adults and the performance of older adults, with the key question being whether different functions are needed for different tasks. As has been done previously (e.g., Verhaeghen et al., 2003), we performed a logit transformation of accuracy scores to try and ensure linearity. While the logic underlying Brinley analysis does not necessitate that the function relating young and old be linear (merely monotonically increasing or decreasing), the applied transformation appears to have had the desired

effect (see Fig. 1). The typical approach is to perform weighted hierarchical linear regression (following Sliwinski & Hall, 1998) where the performance of older adults is treated as the outcome,  $y$ , and the performance of younger adults is treated as the predictor,  $x$ . However, there are several issues with this standard analysis, which we have noted in previous work (Jaroslawska & Rhodes, 2019). Firstly, the implicit assumption is that older adults' scores are measured with unknown error, whereas the younger adult scores are error free. Secondly, while coefficients are weighted for sample size, this analysis does not make use of the information available, in particular the reported estimates of variability in performance. We can extend the basic model to account for this information by assuming that both younger adult (which forms the  $x$  axis) and older adult performance ( $y$  axis) are normally distributed with known error:

$$y_{ij} \sim \text{Normal}(\eta_{ij}, s_{yij}^2)$$

$$x_{ij} \sim \text{Normal}(\lambda_{ij}, s_{xij}^2)$$

where  $s_{yij}$  and  $s_{xij}$  are the reported standard errors. This approach estimates the “true” value of the predictor,  $\lambda_{ij}$ , and uses it in the model relating younger and older performance (i.e., to determine  $\eta_{ij}$ ). We can then consider three models that build incrementally. In model 1, the same intercept ( $\beta_0$ ) and slope ( $\beta_1$ ) terms are used regardless of task:

$$\mathcal{M}1 : \eta_{ij} = \beta_0 + \beta_1 \lambda_{ij} + b_{0j} + b_{1j}$$

where the  $b$  parameters allow for study level differences in both intercept and slope and are assumed to be drawn from a zero centered multivariate normal distribution. As above, this accounts for potential clustering of observations from the same study.

Model 2 adds a parameter,  $\beta_2$ , that allows for task differences in intercept:

$$\mathcal{M}2 : \eta_{ij} = \beta_0 + \beta_1 \lambda_{ij} + \beta_2 I_{ij} + b_{0j} + b_{1j}$$

where  $I_{ij}$  is an indicator to code whether the observation was from a recall or a recognition task. Model 3 additionally allows for task differences in Brinley slope:

$$\mathcal{M}3 : \eta_{ij} = \beta_0 + \beta_1 \lambda_{ij} + \beta_2 I_{ij} + \beta_3 I_{ij} \lambda_{ij} + b_{0j} + b_{1j}$$

These models are known as “errors-in-variables” regression models (Gillard, 2010; Riggs et al., 1978) and are easily implemented in a Bayesian framework with the R package `brms` (Bürkner, 2017, 2018), which serves as an interface to the sampling routines in `Stan` (Carpenter et al. 2016). To estimate this model we use mildly informative priors on the logit scale. Specifically, we used Cauchy distributions with a location of 0 and scale of 2.5 as priors on intercept and slope terms. For the standard deviations of random effects we used a half-Cauchy with the same location and scale as above. For the correlation matrix we used the LKJ prior in `Stan` (see Lewandowski et al., 2009 for details) with

the shape parameter set to 2. Briefly, this is a prior distribution on the correlation matrix for the study level effects (e.g., random intercept and slope terms). A value of 2 places greater prior probability on lower correlations (i.e., peaks at correlations of 0) but does not rule out strong correlations between parameters (a value of 1 gives a uniform distribution across the correlation matrix).

Posterior summaries for model parameters are based on 2000 samples, following 1000 warm-up samples, from four independent chains (i.e., 8000 samples total) with convergence monitored by the  $\hat{R}$  statistic described in Gelman et al. (2014) (pages 284–286). The data and analysis code for this article are available at <https://osf.io/5gx86/>.

## Results

### Age differences in recall and recognition

Our first analysis focused on overall age differences in performance. This model did not take into account the specific task (free recall or item recognition) that the effect size estimate came from. This revealed a clear age difference: 0.694 [0.581, 0.807] ( $z = 12.009, p < 0.01$ ) and substantial residual heterogeneity,  $Q(177) = 728.735, p < 0.01$ .

We then included task in the meta-analytic model with the tasks effects coded such that recall was coded -1 and recognition was coded +1. The intercept in this model was comparable to the overall age difference in the previous model, 0.717 [0.601, 0.834] ( $z = 12.097, p < 0.01$ ), and the coefficient for the effect of task was significantly different from zero, -0.174 [-0.283, -0.065] ( $z = -3.117, p < 0.01$ ).<sup>4</sup> The direction of this coefficient shows that age differences were smaller for recognition than for recall. The estimated age-related effect size (and 95% CI) for recognition is 0.544, [0.365, 0.722] and for recall is 0.891, [0.753, 1.029]. The inclusion of task in the model significantly reduced heterogeneity in effect sizes,  $Q(1) = 9.719, p < 0.01$ , but significant heterogeneity still remains,  $Q(176) = 586.579, p < 0.01$ .

In the above analysis, we not only assumed that studies differed in their overall effect sizes but also in the extent to which effect sizes differ between recall and recognition.

<sup>4</sup>A more principled analysis of the difference in age differences between recall and recognition would focus on the paired mean difference (i.e., the difference in performance between recall and recognition for each age group). For this analysis to be accurate, however, we would need the standard deviation of the mean difference, which relies on the correlation between the measures. This information is unfortunately not typically presented in experimental reports, and here we are effectively assuming that the correlation is 0. It is certainly the case that measures of recall and recognition will be positively correlated and, therefore, our estimate of the difference will be biased downwards. Thus the estimates reported here may be considered a lower bound on the difference in age-effects between recognition and recall.

To test whether there is variability in the difference between recall and recognition effect sizes between studies we fit an additional model which omitted the random study effect for task and compared it to the full model. The likelihood ratio test was significant,  $\chi^2(2) = 98.51$ ,  $p < 0.01$ , suggesting poorer fit for the reduced model ( $AIC = 365.47$ ) relative to the full model ( $AIC = 270.96$ ), supporting the notion of between study variability in the difference between recall and recognition effect sizes. Therefore, the random effect of task was retained in further moderator analyses.

## Bias

Publication bias is likely prevalent across the psychological sciences (see Ioannidis et al., 2014 for discussion). To assess the potential effects of publication bias in the present meta-analysis, we extended the hierarchical meta-analytic model from above to include the estimated standard error of each effect size as a predictor. This is akin to an Egger et al. (1997) test of funnel plot asymmetry (see also Jin et al., 2015), although we do not present the funnel plots themselves, as clustering between studies complicates their interpretation (but is accounted for in our hierarchical models). A significant relationship between effect sizes and the precision with which they are estimated (indexed in this case by the standard error) would indicate publication bias (i.e., an over-representation of small, low precision, studies reporting large effect sizes).

There was indeed a significant relationship between standard error and age effect sizes ( $\beta = 0.424$  [0.296, 0.552] ( $z = 6.506$ ,  $p < 0.01$ )). Nevertheless, the estimated difference between recall and recognition effect sizes was largely unaffected,  $-0.170$  [−0.264, −0.076] ( $z = -3.535$ ,  $p < 0.01$ ). Further, the interaction between standard error and task suggested no significant relationship between the precision of the estimate and the size of the discrepancy between recall and recognition,  $-0.026$  [−0.107, 0.055] ( $z = -0.625$ ,  $p = 0.532$ ). Thus, publication bias appears to influence the overall estimate of age differences but not the estimate of the difference between tasks. This makes sense when considering that the absence of an age by task interaction is probably as of much interest to cognitive aging researchers as the presence of an interaction, and thus no more likely to enter the file drawer.

## Moderators

We considered several possible moderators of the age difference in free recall vs item recognition accuracy. To do this, we extended the meta-analytic model above, in which age differences are allowed to vary by task, to include two new terms: 1) the effect of the moderator on overall age differences and 2) the effect of the moderator on the

difference between tasks (i.e., the moderator by task interaction). We then assessed the reduction in heterogeneity in effect sizes achieved by including a moderator as well as the estimated interaction coefficients.

First, we assessed the effect of whether the to-be-remembered material was studied intentionally for a memory test or whether it was encountered incidentally for an unexpected test. Including this factor in the model did not significantly reduce heterogeneity,  $Q(2) = 0.048$ ,  $p = 0.977$ , and the instruction by task interaction coefficient was 0.009 [−0.073, 0.090] ( $z = 0.213$ ,  $p = 0.832$ ).

Stimulus type did not reduce heterogeneity in effect sizes  $Q(6) = 4.639$ ,  $p = 0.591$ . Words were chosen as the reference level (coded −1) and none of the interaction contrasts were significantly different from zero (actions: 0.006 [−0.332, 0.344] ( $z = 0.034$ ,  $p = 0.973$ ); visual: 0.168 [−0.058, 0.394] ( $z = 1.458$ ,  $p = 0.145$ ); passage: −0.113 [−0.352, 0.125] ( $z = -0.930$ ,  $p = 0.352$ )). Mode of presentation also did not significantly reduce heterogeneity,  $Q(4) = 1.560$ ,  $p = 0.816$ . Visual was chosen as the reference level and neither interaction contrast (for auditory −0.024 [−0.279, 0.230] ( $z = -0.186$ ,  $p = 0.852$ )) or auditory plus visual presentation (−0.079 [−0.504, 0.346] ( $z = -0.364$ ,  $p = 0.716$ )) was significant.

Next we considered log list length as a potential moderator of age differences. However, this did not significantly reduce variability in effect sizes,  $Q(2) = 4.793$ ,  $p = 0.091$ , and the task by list length interaction suggested a small non-significant increase in the discrepancy between recall and recognition with each  $SD$  increase in log list length, 0.036 [−0.047, 0.120] ( $z = 0.854$ ,  $p = 0.393$ ).

Our fifth potential moderator was whether the studied items were related or not. For the full data set (including those not reporting variance estimates) including this in the model did not reduce heterogeneity,  $Q(2) = 4.336$ ,  $p = 0.114$ , whereas there was a significant reduction for those studies reporting usable variance estimates,  $Q(2) = 6.869$ ,  $p < 0.05$ . In this case, however, the reduction in heterogeneity is mainly attributable to a main effect of item relatedness on age differences, 0.152 [0.019, 0.285] ( $z = 2.247$ ,  $p < 0.05$ ), such that overall age differences were *larger* for studies with related items. In both the analysis of the full data set (−0.029 [−0.146, 0.087] ( $z = -0.495$ ,  $p = 0.620$ )) and that restricted to those reporting variance (−0.039 [−0.150, 0.072] ( $z = -0.690$ ,  $p = 0.490$ )), there was no significant interaction between item relatedness and task.

The sixth moderator we looked at was the order of recall and recognition in the experiment. This factor also did not reduce heterogeneity,  $Q(4) = 4.567$ ,  $p = 0.335$ . Recall then recognition, which was by far the most prevalent order, was made the reference factor and neither of the interaction contrasts were significantly different from zero (recognition then recall: −0.046 [−0.507, 0.416] ( $z = -0.194$ ,  $p = 0.846$ ); counterbalanced: 0.023 [−0.240, 0.286] ( $z = 0.172$ ,  $p =$

0.863)). Whether or not the study used the same study list for both recall and recognition also did not significantly reduce heterogeneity,  $Q(2) = 1.867$ ,  $p = 0.393$ , although this analysis was limited by the small number of studies that used separate lists (only 3/89 observations).

For the full sample, the average performance of the younger group in the recognition task was a significant moderator,  $Q(2) = 9.363$ ,  $p < 0.01$ , and this was due to larger overall age differences for studies where younger adults performed well at recognition, 0.117 [0.041, 0.193] ( $z = 3.014$ ,  $p < 0.01$ ). When restricting analyses to only studies that reported estimates of variability, the moderator was not significant,  $Q(2) = 2.581$ ,  $p = 0.275$ , although the coefficient for overall age differences was in the same general direction, 0.074 [-0.017, 0.166] ( $z = 1.586$ ,  $p = 0.113$ ). Importantly, there was no evidence that younger adult recognition influenced the discrepancy between recall and recognition (-0.017 [-0.092, 0.057] ( $z = -0.455$ ,  $p = 0.649$ ) and (0.013 [-0.066, 0.092] ( $z = 0.330$ ,  $p = 0.742$ ) for the full and restricted analyses, respectively).

The final moderator analyses showed that the mean age of the older group did not significantly reduce heterogeneity in effect sizes,  $Q(2) = 2.170$ ,  $p = 0.338$ . The same was true for the difference in mean ages between the older and younger groups,  $Q(2) = 2.684$ ,  $p = 0.261$ .

As a final examination of the moderators we considered, we estimated a master model which included all of the moderators and their interaction with task. Even when all of the moderators were combined in one model, they did not significantly reduce heterogeneity in effect sizes,  $Q(24) = 34.144$ ,  $p = 0.082$ .

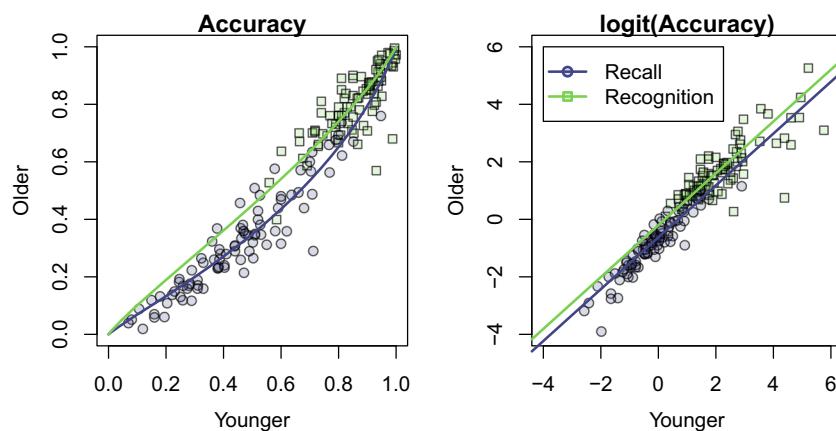
## Brinley

Figure 1 presents Brinley plots of this data set. One observation from Erber et al. (1980) was omitted due to an average

accuracy of 1 for younger adults in the recognition task. Three models were considered for logit transformed accuracy (see Analysis section above). The first was the baseline model, which assumes the same intercept and slope terms for recall and recognition. Model 2 allows for tasks to differ in their intercept term but share the same slope, whereas model 3 allows tasks to differ in both intercept and slope.

Table 1 presents the mean and 95% credible intervals for the population level parameters in each of these models. As this table shows, there appears to be a non-zero difference in intercept terms between recall and recognition. To compare these models, we used the `bridgesampling` package (Gronau & Singmann, 2017) to calculate marginal likelihoods for each model (see Gronau et al., 2017 for an introduction to bridge sampling). Assuming equal prior probabilities of these models (i.e., 1/3) the posterior probability of model 2 given the data is approximately 0.98. The Bayes factor for model 2 relative to model 1 is,  $B_{21} \approx 23364$ -to-1, and for model 3 relative to model 1 is,  $B_{31} \approx 459$ -to-1. For model 2 relative to model 3,  $B_{23} \approx 50$ -to-1. Thus the difference between tasks in Brinley intercepts is strongly supported by the data. These Brinley functions are displayed in Fig. 1. Note also that in both models 2 and 3 the overall intercept term is reliably negative and the slope term is smaller than 1 (although 1 is just included in the 95% credible interval for model 2). This is consistent with the overall age-effect on performance that we observed in the above standard meta-analysis.

The lower overall intercept for recall (-0.625, [-0.740, -0.517]), relative to recognition -0.204, [-0.366, -0.040]), supports a constant age-related deficit in this task, which, as we discussed in the Introduction section, may relate to different criteria for the termination of memory search (Raaijmakers & Shiffrin, 2002) or possibly a greater susceptibility to intrusions during recall (Healey & Kahana, 2016). However, we do note that our conclusions around



**Fig. 1** Brinley plots presenting the performance of older adults as a function of the performance of younger adults for recall and recognition tasks. The left panel plots performance on accuracy scale and the right panel plots log odd (logit) transformed accuracy. The lines are grand mean functions from the winning model in which recall and recognition differ in intercept

**Table 1** Results of Brinley analyses. Mean and 95% credible intervals are presented for population level parameters in the three models considered (see text for details)

	Model 1	Model 2	Model 3
Intercept	−0.551 [−0.657, −0.449]	−0.414 [−0.537, −0.295]	−0.397 [−0.531, −0.266]
Slope	1.081 [ 0.999, 1.163]	0.902 [ 0.804, 1.000]	0.897 [ 0.801, 0.994]
Task	—	0.210 [ 0.137, 0.280]	0.227 [ 0.133, 0.317]
Slope by Task	—	—	−0.019 [−0.080, 0.043]

potential differences between tasks in Brinley slopes are somewhat limited by the restricted range of recognition scores (see Fig. 1). Thus we cannot rule out the possibility that we would have also observed slope differences if there were greater variability in recognition accuracy. Nevertheless, the above results clearly favor distinct Brinley functions for recall and recognition relative to a single function across tasks. We also considered a version of model 1 with an additional quadratic trend to check for non-linearities in the Brinley plot (cf. Verhaeghen, 2013), however, this was not an improvement on model 2.

Given the evidence for separate intercepts by task, we went on to consider each of the potential moderators outlined above and their modulation of the Brinley intercept. For each moderator an extension of model 2 was created that included a main effect of the moderator as well as an interaction term with task. For every moderator we considered, but one, neither their main effects nor their interactions with task were significantly different from zero (i.e., all coefficients included zero within their 95% credible intervals). The exception to this was the mean difference between age groups included in the study. While this factor had no influence on the Brinley intercept overall (−0.017, [−0.100, 0.059] per SD increase in mean age difference) its interaction with task was non-zero (−0.062, [−0.110, −0.016]) and suggests that the difference in Brinley intercepts between tasks gets *smaller* with increasing differences in mean age. This is certainly unexpected and we offer no explanation of this result. Indeed when we calculated a Bayes' factor for this moderator model relative to the original Model 2 the latter was favored by well over 100-to-1, suggesting that adding this moderator did not improve the likelihood of the model given the data.

## Discussion

While there have been some studies that have found approximately equivalent age differences in free recall and item recognition performance (Botwinick & Storandt, 1980; Verhaeghen et al., 1998; White & Cunningham, 1982), the literature has essentially reached the conclusion that there are non-zero age differences in item recognition

(Fraundorf et al., 2019; Old & Naveh-Benjamin, 2008) and that the age effect is probably larger for free recall tasks (Danckert & Craik, 2013; Erber et al., 1980; Harwood & Naylor, 1969; Nyberg et al., 2003; Schonfield & Robertson, 1966; Whiting et al., 1997). The present meta-analysis aimed to build on this previous work in two ways: First, we wanted to estimate the magnitude of age differences in these tasks by combining the findings of a range of studies that have directly compared recall and recognition performance in younger and older adults using similar materials and study procedures. Second, using Brinley analysis, we wanted to address the question of to what extent any larger age effect for recall relative to recognition can be considered *disproportionate*. If the magnitude of age difference in recall cannot be effectively predicted from age differences in recognition this might suggest a specific deficit. For both of these sets of analyses we considered several characteristics of the studies that could conceivably moderate recall and recognition performance.

Examining standardized effect sizes, we do find that age differences in recognition performance are significantly larger than zero, 0.544, [0.365, 0.722]. This is in line with the recent meta-analysis of Fraundorf et al. (2019), who report an age difference of around 0.46 in  $d'$  units, and with the meta-analysis of Old and Naveh-Benjamin (2008), who report an age effect size for item recognition of around 0.65. While the age difference in item recognition is clearly greater than zero, we found that effect sizes for recall were reliably larger, 0.891, [0.753, 1.029], with an estimated difference in effect sizes of 0.347 [0.129, 0.566]. This suggests that there is indeed a theoretically meaningful division to be drawn between age-related effects on recall and recognition. The Brinley plot (Fig. 1) gives us further insight into the nature of this difference. Specifically, it appears that two Brinley functions are required to relate the performance of younger and older groups of participants; one for item recognition and another, that is shifted downwards, for recall (see Table 1). This implies a constant cost to older adults' recall performance across the range of performance levels, although our conclusion in this regard is somewhat limited by the range of recognition scores. Nevertheless, the support for distinct Brinley functions for recall and recognition tasks over a single function was clear.

These findings are in line with accounts of memory and aging that posit a specific deficit to processes related to recall. For example, as discussed in the introduction, there is good evidence for general declines in the fidelity of memory representations with age (e.g. Benjamin, 2010; Li et al., 2005). This would affect recognition performance but, to the extent that item recognition decisions can be made on a single familiarity metric (Dunn, 2004; Haaf et al., 2018; Pratte and Rouder, 2012) and do not particularly rely on recall-like processes, such as recollection, a specific deficit in searching and recalling from memory would produce the overall age differences and the Brinley findings we present here. This kind of account was implemented by Healey and Kahana (2016) whose model of memory and aging contains a number of parameters that apply generally in explaining age-deficits in memory performance but in addition assumes that recall computations are noisier for older adults, which would produce both the general decline we see in performance (i.e., in recognition memory) as well as the disproportionate effect of age on recall.

Our findings are more difficult to reconcile with the idea that age differences in these tasks can solely be explained by placing the tasks on a continuum of self-initiated processing, demand for environmental support, or related concepts. However, it may be the case that tasks that are more of a mix of recall and recognition, such as associative recognition (Old & Naveh-Benjamin, 2008; Rotello & Heit, 2000) or cued recall (Craik & McDowd, 1987), may produce such a continuum between the extremes of item recognition and free recall. Note, however, that our estimate of the age difference in free recall was comparable to the estimated difference reported in a smaller set of recall tasks by La Voie and Light (1994) (0.968 [0.835, 1.101]) and that their set included mostly cued recall conditions. It is conceivable that the presence of a few free recall studies in the analysis by La Voie and Light (1994) may have made the age difference larger in their study than if cued recall tests alone were used, so the similar effects obtained with free recall only in the present study do not necessarily rule out the potential for a continuum. Nevertheless, another prediction of this account was not realized; specifically an account in which the “difficulty” of the task determines the magnitude of age differences would predict a Brinley function with a slope greater than 1. While a slope of 1 is just contained in the 95% most probable values of our best fitting Brinley model (see Table 1), we can be fairly confident that the Brinley slope is not greater than 1. This is similar to the finding of Fraundorf et al. (2019) who plotted the recognition  $d'$  of older adults as a function of the  $d'$  of younger adults and fit a linear function with a slope less than 1.

We considered a range of study characteristics that could conceivably moderate age differences in these tasks. In the set of studies that reported estimates of variance we found

that item relatedness played a role in overall age differences, such that the gap between younger and older adults' performance was larger when items were related. This role of item relatedness in modulating age differences did not appear to differ between recall and recognition. This is perhaps surprising given reports of the benefits of item relatedness for older adults' performance, particularly in tasks requiring explicit memory for association between items where it has been shown to reduce the difference between age groups (e.g., Ahmad et al., 2015; Bastin et al., 2013; Delhaye et al., 2019; Naveh-Benjamin et al., 2005) and we expected that presenting related items would reduce the age difference in performance. However, the literature does suggest a potential interpretation of this result. Older adults appear to rely more heavily on gist, rather than verbatim, information in memory tasks (see Brainerd and Reyna, 2015 for a review) and may be able to capitalize on their intact memory for gist information to support memory for lists of related items. However, when considering tasks requiring memory for specific items, as we have focused on here, greater reliance on gist would also likely result in a greater incidence of false recognition (Delhaye et al., 2019; Koutstaal & Schacter, 1997; Koutstaal et al., 1999) and false recall (Brainerd & Reyna, 2015; Brainerd et al., 2009) for older adults. However, this interpretation is highly speculative, given the rather small influence of item relatedness on age differences in the present meta-analysis and the fact that its influence was significant only for a subset of the studies included.

We had initially expected that other factors that influence the ease with which items can be associated (i.e., list length, study instructions) might also modulate the age differences in recall and recognition, but this did not appear to be the case in the studies we examined. Indeed none of the moderators we considered was found to modulate the difference between recall and recognition. This is a limitation of the present work as there is substantial remaining heterogeneity in effect sizes even after accounting for the role of task (i.e., recall vs recognition). It is likely that there are other characteristics of the experiments that systematically vary and contribute to the residual heterogeneity but we did not identify them.

It is also possible that there are other sources of variability that are harder to glean from experimental reports. For example, work with groups of younger participants has shown that there are considerable item effects on memory performance (Cox et al., 2018; Freeman et al., 2010; Rouder & Lu, 2005); that is, items vary in how recallable or recognizable they are (in ways that go beyond broad categorizations, such as high- or low-frequency words). It is reasonable to assume that item effects interact with age differences, such that the age difference found with one randomly sampled set of items will differ with another random set. Thus, a source of residual variability in the effect sizes

found here may stem from considering age differences in memory as a fixed effect across items (like the language-as-a-fixed-effect fallacy; Clark, 1973). Future work on age differences in memory should aim to take into account item variability when estimating performance differences (for example, using mixed effects models; see, Baayen et al., 2008 for an introduction). In addition, factors like the time of day that participants are tested also influence the magnitude of age differences (see May et al., 1993). Factors such as this, which rarely make it into experimental reports, undoubtedly contribute to between study heterogeneity, although the magnitude of their influence is unknown. Unfortunately in the present meta-analysis we have been unable to account for this residual heterogeneity in effect sizes. We now turn to other possible limitations of the present work.

## Limitations and future directions

The present meta-analysis shares the inherent limitations of all meta-analyses. Our selection criteria, while we think they are reasonable, may have introduced a particular bias by excluding a section of the literature (we discuss one effect of our selection criteria below). Publication bias can also distort estimates of effect size, and we did find that smaller, less precise, studies tended to report larger age differences overall. Interestingly, we did not find evidence of bias for the contrast between recall and recognition. While this does not allow us to claim that there is *no* publication bias in this regard we do speculate that, given the field's interest in the presence or absence of age group interactions, a null finding would be as likely to be published as a rejection of the null. Further, it is clear that in aggregating over a range of different studies we can make no claims about causality, as it is possible some other variable is driving the age differences we observe. Some of this is inherent to cross-sectional studies of aging, which are correlational in nature. Nevertheless, we tried to minimize the potential confounding of recall and recognition tasks with other methodological variables by focusing on studies that directly compared the two with similar materials and study procedures. The results of the analyses presented here should therefore be considered a converging, but not cast-iron, source of evidence along with careful tests of age by task interactions (e.g., Danckert & Craik, 2013) and detailed computational models of age-related change to episodic memory (e.g., Healey & Kahana, 2016; Li et al., 2005).

As we have noted several times, we chose to focus on studies that reported direct comparisons of free recall and item recognition tasks in the same groups of younger and older adults. This was to try to minimize differences between tasks other than the mode of testing. However, this resulted in a sample of studies in which the vast

majority tested recall before recognition, as opposed to the opposite ordering or counterbalanced presentation of the two. Further, the majority of studies used the exact same study list for both tasks. It may be that previous recall attempts somehow contaminate recognition. For example, tests of retrieval, even without feedback, are well known to improve subsequent retention (known as retrieval practice or the testing effect; Carrier & Pashler, 1992; Roediger & Karpicke, 2006). Indeed, Fraundorf et al. (2019) found that age differences in recognition sensitivity were somewhat smaller (by 0.196 [0.030, 0.363] in  $d'$  units) when a free recall task was performed prior to the recognition task. This is concerning as it might be that including a majority of studies in which recall was tested prior to recognition for the same list has exacerbated the discrepancy between recall and recognition tasks. Nevertheless, our 95% confidence intervals for the age effect size for item recognition ([0.37, 0.72]) overlap with those reported by Old and Naveh-Benjamin (2008) ([0.58, 0.71]), who included a broader range of recognition studies. Further, the confidence intervals for our estimate of the age difference in recall ([0.75, 1.03]) do not overlap with either item recognition estimate. Future experimental studies should attempt to compare age differences in recall and recognition tasks where the order of the tasks is counterbalanced and different study lists are used to reduce the possible influence of testing effects (although the extant literature suggests that the testing effect benefits retention regardless of age; Coane, 2013; Meyer & Logan, 2013; Pastötter & Bäuml, 2019). In addition, future meta-analyses could relax the need for direct comparisons of tasks, which would also allow them to compare a broader range of tasks (e.g., cued recall, associative recognition).

Our Brinley analyses suggested that two distinct functions are needed to relate younger and older accuracy for recall and recognition tasks. We made the assumption of linearity<sup>5</sup> in our analysis and, of course, it is possible that one could find a single monotonic function for the Brinley plot that applies to both tasks, which would help a continuum type account of age differences in episodic memory tasks (although see the discussion of Brinley slopes above). We cannot completely rule this out, however, we note that we performed a transformation on the accuracy data to try and ensure linearity (see Fig. 1) and, in addition, considered a simple quadratic model to relax the assumption of linearity.

<sup>5</sup>Verhaeghen and Marcoen (1993) reported Brinley plots of recall tasks and found some evidence of non-linearity. However, they fit their ordinary least squares regressions to the raw accuracy scale, rather than a transformation like the present analysis. The left panel of Fig. 1 presents our results transformed to the accuracy scale and demonstrate a similar non-linearity on this scale.

## Conclusions

In summary, we identified 36 articles reporting 89 direct comparisons of free recall and item recognition performance in the same groups of younger and older adults. Synthesizing the results of these articles confirms that age differences are larger for recall than recognition, but differences are clearly larger than zero for the latter. Despite these clear mean differences, and our consideration of several possible moderators, substantial variability in effect sizes between studies still remains to be explained.

When plotting the performance of older adults as a function of the performance of younger adults we find that separate lines for recall and recognition, differing in intercept, are an improvement over a single line. This is in line with a conclusion that the age difference in recall is disproportionate to that for recognition and supports theories of memory and aging which posit specific deficits in processes related to retrieval.

**Author Note** The data and analysis code for this article are available at <https://osf.io/5gx86/>.

## Appendix

**Table 2** Sample of studies included in the present meta-analysis

	Study/Condition	N young	N old	$M_{age}$ young	$M_{age}$ old	g recall	g recognition
1	Adamowicz and Hudson (1978).	12	12	19.00	71.70	0.50 [−0.31, 1.31]	3.20 [1.99, 4.41]
2	Bohn et al. (2016).	25	25	22.48	67.56	0.66 [0.09, 1.23]	−0.21 [−0.77, 0.34]
3	Burack and Lachman (1996).	24	24	23.67	70.21	0.87 [0.28, 1.46]	0.73 [0.15, 1.32]
4	Charles et al. (2003). Experiment 2 Negatively-Valenced Images	32	32	23.53	74.06	1.47 [0.92, 2.02]	0.44 [−0.05, 0.94]
5	Charles et al. (2003). Experiment 2 Neutral Images					1.25 [0.71, 1.79]	0.08 [−0.41, 0.57]
6	Charles et al. (2003). Experiment 2 Positively-Valenced Images					1.13 [0.60, 1.66]	0.32 [−0.17, 0.81]
7	Charles et al. (2003). Experiment I Negatively-Valenced Images	48	48	24.56	70.98	1.79 [1.31, 2.26]	0.07 [−0.33, 0.47]
8	Charles et al. (2003). Experiment I Neutral Images					1.39 [0.94, 1.83]	−0.30 [−0.70, 0.11]
9	Charles et al. (2003). Experiment I Positively-Valenced Images					0.47 [0.07, 0.88]	−0.37 [−0.77, 0.04]
10	Danckert and Craik (2013). Experiment 1	24	24	20.71	69.54	1.47 [0.84, 2.11]	0.36 [−0.21, 0.93]
11	Danckert and Craik (2013). Experiment 2 Deep Encoding	24	24	20.90	70.00	1.16 [0.55, 1.77]	0.00 [−0.57, 0.57]
12	Danckert and Craik (2013). Experiment 2 Learning Condition					0.67 [0.09, 1.26]	−0.16 [−0.73, 0.40]
13	Danckert and Craik (2013). Experiment 3 Grouping Condition	20	19	20.70	72.80	1.47 [0.77, 2.18]	−0.36 [−0.99, 0.28]
14	Danckert and Craik (2013). Experiment 3 Individual Condition	20	20	20.70	72.80	0.77 [0.12, 1.41]	−0.68 [−1.31, −0.04]
15	Davis et al. (1990). Experiment 1 Delayed Test following 5 learning trials	45	21	27.50	75.50	1.39 [0.82, 1.96]	−0.09 [−0.61, 0.42]
16	Davis et al. (1990). Experiment 1 Immediate Test following Trial 1					0.25 [−0.27, 0.77]	0.28 [−0.24, 0.80]
17	Davis et al. (1990). Experiment 2	47	62	26.03	74.24	0.31 [−0.07, 0.69]	0.43 [0.05, 0.81]
18	Erber (1974).	38	38	23.30	70.00	1.19 [0.70, 1.68]	0.71 [0.25, 1.18]
19	Erber et al. (1980). Intentional Control Condition	12	12	23.90	69.40	1.29 [0.41, 2.17]	0.53 [−0.29, 1.34]
20	Erber et al. (1980). Non-semantic + intentional learning condition					1.72 [0.78, 2.65]	1.11 [0.25, 1.97]

**Table 2** (continued)

	Study/Condition	<i>N</i> young	<i>N</i> old	<i>M<sub>age</sub></i> young	<i>M<sub>age</sub></i> old	<i>g</i> recall	<i>g</i> recognition
21	Erber et al. (1980). Non-semantic orienting condition					0.32 [−0.48, 1.13]	0.68 [−0.15, 1.50]
22	Erber et al. (1980). Semantic + Intentional Learning Condition					1.39 [0.50, 2.28]	0.43 [−0.38, 1.24]
23	Erber et al. (1980). Semantic Orienting Task (Pleasant/Unpleasant)					1.00 [0.16, 1.85]	0.84 [0.01, 1.67]
24	Gordon and Clark (1974). Delayed	22	22	24.73	71.23	1.99 [1.27, 2.72]	0.79 [0.18, 1.41]
25	Gordon and Clark (1974). Immediate			24.73	71.23	0.80 [0.19, 1.42]	0.79 [0.18, 1.41]
26	Haaland et al. (2003). Logical Memory Delayed	552	551	24.47	75.28	1.33 [1.20, 1.46]	0.87 [0.75, 0.99]
27	Harwood and Naylor (1969). Delayed Test	58	59	70.00	30.00	0.74 [0.36, 1.11]	1.56 [1.15, 1.97]
28	Howard (1983). Experiment 1	20	20	31.10	69.50	0.65 [0.02, 1.29]	0.67 [0.03, 1.31]
29	Kalenzaga et al. (2016). No processing instruction; Negative Words Delayed	38	39	24.39	71.48	0.33 [−0.12, 0.78]	1.32 [0.83, 1.82]
30	Kalenzaga et al. (2016). No processing instruction; Neutral Words Delayed					0.53 [0.08, 0.99]	0.96 [0.49, 1.44]
31	Kalenzaga et al. (2016). No processing instruction; Positive Words Delayed					0.43 [−0.02, 0.88]	0.62 [0.16, 1.08]
32	Kalenzaga et al. (2016). Semantic processing instruction; Negative Words Delayed					0.57 [0.11, 1.03]	0.47 [0.02, 0.93]
33	Kalenzaga et al. (2016). Semantic processing instruction; Neutral Words Delayed					0.60 [0.14, 1.06]	0.51 [0.06, 0.97]
34	Kalenzaga et al. (2016). Semantic processing instruction; Positive Words Delayed					1.40 [0.90, 1.90]	0.29 [−0.16, 0.74]
35	Karrasch et al. (2010).	22	22	21.70	70.60	0.32 [−0.27, 0.92]	−0.01 [−0.60, 0.58]
36	Knopf and Neidhardt (1989). Experiment 1 Motor Encoding Condition, High Familiarity, Delayed	30	30	26.40	67.20	1.36 [0.80, 1.92]	0.28 [−0.23, 0.79]
37	Knopf and Neidhardt (1989). Experiment 1 Motor Encoding Condition, Low Familiarity, Delayed					1.18 [0.64, 1.73]	0.39 [−0.12, 0.90]
38	Knopf and Neidhardt (1989). Experiment 1 Motor Encoding Condition, Medium Familiarity, Delayed					1.04 [0.50, 1.58]	0.39 [−0.12, 0.90]
39	Knopf and Neidhardt (1989). Experiment 1 Verbal Encoding Condition, High Familiarity, Delayed					0.86 [0.33, 1.39]	0.29 [−0.22, 0.79]
40	Kramer et al. (1989).	16	16	33.00	64.70	0.70 [−0.01, 1.42]	0.15 [−0.55, 0.84]
41	Leigland et al. (2004). Delayed	25	25	23.90	72.30	1.67 [1.03, 2.32]	1.51 [0.88, 2.14]
42	Light and Anderson (1983). Experiment 2	24	24	24.58	72.63	0.59 [0.01, 1.17]	0.69 [0.10, 1.27]
43	Light and Singh (1987). Experiment 1 Pleasantness Condition	16	16	23.50	67.70	1.63 [0.83, 2.43]	0.40 [−0.30, 1.10]
44	Light and Singh (1987). Experiment 1 Vowel Comparison Condition					1.28 [0.52, 2.04]	−0.57 [−1.27, 0.14]

**Table 2** (continued)

	Study/Condition	<i>N</i> young	<i>N</i> old	<i>M<sub>age</sub></i> young	<i>M<sub>age</sub></i> old	<i>g</i> recall	<i>g</i> recognition
45	Light and Singh (1987). Experiment 3 Pleasantness Condition	16	16	21.30	69.10	1.15 [0.40, 1.90]	0.43 [-0.28, 1.13]
46	Light and Singh (1987). Experiment 3 Vowel Comparison Condition					0.51 [-0.20, 1.21]	-0.05 [-0.74, 0.64]
47	Madden (1986).	24	24	19.30	67.50	1.55 [0.91, 2.20]	1.44 [0.81, 2.08]
48	Murphy et al. (1997). Odor Condition Delayed	20	20	22.80	68.80	1.37 [0.69, 2.06]	2.57 [1.73, 3.41]
49	Murphy et al. (1997). Word Condition Delayed					0.99 [0.33, 1.64]	1.13 [0.46, 1.80]
50	Neely et al. (2017).	32	33	21.53	69.27	0.01 [-0.47, 0.50]	-0.19 [-0.68, 0.29]
51	Perlmutter (1978). Incidental Learning Condition	32	32	23.00	62.00	0.88 [0.36, 1.39]	0.77 [0.26, 1.28]
52	Perlmutter (1978). Intentional Learning Condition					0.44 [-0.05, 0.94]	0.31 [-0.19, 0.80]
53	Perlmutter (1979). Experiment I Incidental Learning Condition	48	48	20.00	63.00	0.78 [0.36, 1.19]	0.16 [-0.24, 0.56]
54	Perlmutter (1979). Experiment I Intentional Learning Condition					0.78 [0.36, 1.19]	0.57 [0.16, 0.98]
55	Rohling et al. (1991).	30	30	25.40	73.60	0.08 [-0.42, 0.59]	0.12 [-0.39, 0.62]
56	Ruiz-Gallego-Largo et al. (2012). Source ("I"), Referent ("About Me") Condition	30	30	23.47	70.30	1.04 [0.50, 1.58]	0.57 [0.06, 1.09]
57	Ruiz-Gallego-Largo et al. (2012). Source ("I"), Referent ("About You") Condition					0.47 [-0.05, 0.98]	0.38 [-0.13, 0.89]
58	Ruiz-Gallego-Largo et al. (2012). Source ("You"), Referent ("About Me") Condition					0.52 [0.00, 1.03]	0.27 [-0.24, 0.78]
59	Ruiz-Gallego-Largo et al. (2012). Source ("You"), Referent ("About You") Condition					0.40 [-0.11, 0.91]	-0.17 [-0.68, 0.33]
60	Russo and Parkin (1993). Experiment 1 Full attention condition	24	24	26.30	73.70	1.26 [0.64, 1.88]	0.29 [-0.28, 0.86]
61	Schmitter-Edgecombe and Simpson (2001). Incidental Condition	24	48	20.11	70.86	0.63 [0.13, 1.13]	0.78 [0.27, 1.28]
62	Schmitter-Edgecombe and Simpson (2001). Intentional Learning - Content + Temporal Order Condition					0.69 [0.19, 1.20]	0.82 [0.31, 1.33]
63	Schmitter-Edgecombe and Simpson (2001). Intentional Learning - Content Condition					0.43 [-0.07, 0.92]	0.62 [0.12, 1.12]
64	Schonfield and Robertson (1966).	59	21	67.50	29.50	1.25 [0.72, 1.79]	-0.10 [-0.60, 0.40]
65	Schonfield et al. (1983). Nonswitch Condition	16	16	22.80	68.80	0.67 [-0.05, 1.38]	0.72 [0.01, 1.44]
66	Schonfield et al. (1983). Switch Condition					0.36 [-0.34, 1.06]	0.50 [-0.21, 1.20]
67	Schramke and Bauer (1997). Exercise- Exercise condition	12	12	19.75	68.08	0.50 [-0.31, 1.32]	0.60 [-0.21, 1.42]
68	Schramke and Bauer (1997). Exercise- Rest condition	12	12	19.00	71.91	0.82 [-0.01, 1.66]	0.43 [-0.38, 1.24]
69	Schramke and Bauer (1997). Rest-Exercise condition	12	12	19.67	70.00	1.31 [0.43, 2.20]	0.41 [-0.40, 1.22]
70	Schramke and Bauer (1997). Rest-Rest condition	12	12	20.00	67.42	0.72 [-0.11, 1.54]	1.01 [0.16, 1.85]
71	Sommers and Huff (2003). Experiment 1 Recall-plus-recognition condition	24	24	19.30	70.90	2.17 [1.46, 2.88]	0.42 [-0.15, 0.99]
72	Sommers and Huff (2003). Experiment 2	24	24	19.10	71.20	1.25 [0.63, 1.87]	0.53 [-0.05, 1.10]
73	Spilich and Voss (1983).	12	12	19.30	75.80	2.39 [1.34, 3.43]	0.13 [-0.67, 0.94]
74	Trelle et al. (2015). Episodic Condition; Self-Referential Domain; Binary Judgment	24	24	21.50	70.46	1.18 [0.57, 1.79]	0.39 [-0.19, 0.96]

**Table 2** (continued)

	Study/Condition	<i>N</i> young	<i>N</i> old	<i>M<sub>age</sub></i> young	<i>M<sub>age</sub></i> old	<i>g</i> recall	<i>g</i> recognition
75	Trelle et al. (2015). Episodic Condition; Self-Referential Domain; Narrative Response					2.19 [1.47, 2.90]	0.10 [-0.46, 0.67]
76	Trelle et al. (2015). Episodic Condition; Semantic Domain; Binary Judgment					1.47 [0.84, 2.11]	0.62 [0.04, 1.20]
77	Trelle et al. (2015). Episodic Condition; Semantic Domain; Narrative Response					1.28 [0.66, 1.90]	-0.52 [-1.10, 0.05]
78	Trelle et al. (2015). Semantic Condition; Self-Referential Domain; Binary Judgment	24	24	22.29	69.21	0.95 [0.35, 1.54]	0.45 [-0.12, 1.02]
79	Trelle et al. (2015). Semantic Condition; Self-Referential Domain; Narrative Response					1.05 [0.45, 1.65]	-0.33 [-0.90, 0.24]
80	Trelle et al. (2015). Semantic Condition; Semantic Domain; Binary Judgment					0.74 [0.16, 1.33]	0.58 [0.01, 1.16]
81	Trelle et al. (2015). Semantic Condition; Semantic Domain; Narrative Response					0.97 [0.37, 1.57]	-0.31 [-0.88, 0.26]
82	Vakil et al. (1996). 1 second exposure condition	20	15	22.50	84.00	0.00 [-0.67, 0.67]	1.61 [0.84, 2.37]
83	Vakil et al. (1996). 6 second exposure condition	15	15	22.50	84.00	0.00 [-0.72, 0.72]	1.35 [0.56, 2.15]
84	Zelinski and Miura (1988). Experiment 1 Delayed test; no theme condition	17	17	24.93	69.24	1.21 [0.48, 1.94]	0.28 [-0.39, 0.96]
85	Zelinski and Miura (1988). Experiment 1 Delayed test; theme condition					1.28 [0.54, 2.01]	0.34 [-0.33, 1.02]
86	Zelinski and Miura (1988). Experiment 1 Immediate test; no theme condition					0.91 [0.21, 1.62]	0.38 [-0.30, 1.06]
87	Zelinski and Miura (1988). Experiment 1 Immediate test; theme condition					0.96 [0.25, 1.67]	0.06 [-0.61, 0.74]
88	Zelinski and Miura (1988). Experiment 2 Delayed test; no theme condition	17	18	22.12	69.18	0.65 [-0.03, 1.33]	0.21 [-0.46, 0.87]
89	Zelinski and Miura (1988). Experiment 2 Delayed test; theme condition	16	16	22.12	69.18	0.57 [-0.13, 1.28]	0.26 [-0.44, 0.95]

As some studies included multiple conditions tested on the same groups of participants, *N* and mean age (where available) are presented for each new sample. Additional information can be found at <https://osf.io/5gx86/>

## References

- Adamowicz, J., & Hudson, B. (1978). Visual short-term memory, response delay, and age. *Perceptual and Motor Skills*, 46(1), 267–270.
- Ahmad, F. N., Fernandes, M., & Hockley, W. E. (2015). Improving associative memory in older adults with unitization. *Aging, Neuropsychology, and Cognition*, 22(4), 452–472.
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In Krantz, D. H., Atkinson, R. C., Luce, R. D., & Suppes, P. (Eds.) *Contemporary developments in mathematical psychology* (vol. 1): *Learning, memory, and thinking*, (pp. 239–289). San Francisco: Freeman.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bastin, C., Diana, R. A., Simon, J., Collette, F., Yonelinas, A. P., & Salmon, E. (2013). Associative memory in aging: The effect of unitization on source memory. *Psychology and Aging*, 28(1), 275.
- Benjamin, A. S. (2010). Representational explanations of “process” dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review*, 117(4), 1055–1079.
- Bohn, L., See, S. T. K., & Fung, H. H. (2016). Time perspective and positivity effects in Alzheimer’s disease. *Psychology and Aging*, 31(6), 574–582.
- Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 60(5), P223–P233.
- Botwinick, J., & Storandt, M. (1980). Recall and recognition of old information in relation to age and sex. *Journal of Gerontology*, 35(1), 70–76.
- Brainerd, C., & Reyna, V. (2015). Fuzzy-trace theory and lifespan cognitive development. *Developmental Review*, 38, 89–121.
- Brainerd, C., Reyna, V., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: A unified theory. *Psychological Review*, 116(4), 783.
- Brinley, J. F. (1965). Cognitive sets, speed and accuracy of performance in the elderly. *Behavior, Aging and the Nervous System*, 114–149.

- Burack, O. R., & Lachman, M. E. (1996). The effects of list-making on recall in young and elderly adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 51(4), P226–P233.
- Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M. . . . , et al. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20(2), 1–37.
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642.
- Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin*, 98(1), 67.
- Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology: General*, 132(2), 310.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, 2(2), 95–100.
- Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, 147(4), 545.
- Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London B*, 302(1110), 341–359.
- Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3), 474.
- Danckert, S. L., & Craik, F. I. M. (2013). Does aging affect recall more than recognition memory? *Psychology and Aging*, 28(4), 902–909.
- Davis, H. P., Cohen, A., Gandy, M., Colombo, P., VanDusseldorp, G., Simolke, N., & Romano, J. (1990). Lexical priming deficits as a function of age. *Behavioral Neuroscience*, 104(2), 288.
- Delhaye, E., Folville, A., & Bastin, C. (2019). How to induce an age-related benefit of semantic relatedness in associative memory: It's all in the design. *Psychology and Aging*, 34(4), 572.
- Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, 111(2), 524–542.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.
- Erber, J. (1974). Age differences in recognition memory. *Journal of Gerontology*, 29(2), 177–181.
- Erber, J., Herman, T. G., & Botwinick, J. (1980). Age differences in memory as a function of depth of processing. *Experimental Aging Research*, 6(4), 341–348.
- Fraundorf, S. H., Hourihan, K. L., Peters, R. A., & Benjamin, A. S. (2019). Aging and recognition memory: A meta-analysis. *Psychological Bulletin*, online ahead of print.
- Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, 62(1), 1–18.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* Vol. 3. Boca Raton: Chapman & Hall/CRC.
- Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *REVSTAT-Statistical Journal*, 8(1), 57–80.
- Gordon, S. K., & Clark, W. C. (1974). Application of signal detection theory to prose recall and recognition in elderly and young adults. *Journal of Gerontology*, 29(1), 64–72.
- Gronau, Q. F., & Singmann, H. (2017). Bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors. Retrieved from <https://CRAN.R-project.org/package=bridgesampling>
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., & Steingrover, H. . . . (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97.
- Haaf, J. M., Rhodes, S., Sun, T. K., Snyder, H. K., Naveh-Benjamin, M., & Rouder, J. (2018). Revisiting the remember-know task: Replications of Gardiner and Java (1990). Retrieved from <https://psyarxiv.com/2dyn7/>
- Haaland, K. Y., Price, L., & Larue, A. (2003). What does the WMS-III tell us about memory changes with normal aging? *Journal of the International Neuropsychological Society*, 9(1), 89–96.
- Harwood, E., & Naylor, G. (1969). Recall and recognition in elderly and young subjects. *Australian Journal of Psychology*, 21(3), 251–257.
- Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, 123(1), 23–69.
- Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128.
- Howard, D. V. (1983). The effects of aging and degree of association on the semantic priming of lexical decisions. *Experimental Aging Research*, 9(3), 145–151.
- Howard, M. W., Bessette-Symons, B., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging*, 21(1), 96.
- Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5), 235–241.
- Jaroslawska, A. J., & Rhodes, S. (2019). Adult age differences in the effects of processing on storage in working memory: A meta-analysis. *Psychology and Aging*, 34(4), 512.
- Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: Aging, attention, and control. *Psychology and Aging*, 8(2), 283–293.
- Jin, Z.-C., Zhou, X.-H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, 34(2), 343–360.
- Kalenzaga, S., Lamidey, V., Ergis, A.-M., Clarys, D., & Piolino, P. (2016). The positivity bias in aging: Motivation or degradation? *Emotion*, 16(5), 602–610.
- Karrasch, M., Myllyniemi, A., Latvala, L., Söderholm, C., Ellfolk, U., & Laine, M. (2010). The diagnostic accuracy of an incidental memory modification of the Boston Naming Test (memo-BNT) in differentiating between normal aging and mild Alzheimer's disease. *The Clinical Neuropsychologist*, 24(8), 1355–1364.
- Knopf, M., & Neidhardt, E. (1989). Aging and memory for action events: The role of familiarity. *Developmental Psychology*, 25(5), 780.
- Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, 37(4), 555–583.
- Koutstaal, W., Schacter, D. L., Galluccio, L., & Stofer, K. A. (1999). Reducing gist-based false recognition in older adults: Encoding and retrieval manipulations. *Psychology and Aging*, 14(2), 220.
- Kramer, J. H., Blusewicz, M. J., & Preston, K. A. (1989). The premature aging hypothesis: Old before its time? *Journal of Consulting and Clinical Psychology*, 57(2), 257.

- La Voie, D., & Light, L. L. (1994). Adult age differences in repetition priming: A meta-analysis. *Psychology and Aging, 9*(4), 539–553.
- Lalanne, J., Rozenberg, J., Grolleau, P., & Piolino, P. (2013). The self-reference effect on episodic memory recollection in young and older adults and Alzheimer's disease. *Current Alzheimer Research, 10*(10), 1107–1117.
- Leigland, L. A., Schulz, L. E., & Janowsky, J. S. (2004). Age-related changes in emotional memory. *Neurobiology of Aging, 25*(8), 1117–1124.
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001.
- Li, S. C., Naveh-Benjamin, M., & Lindenberger, U. (2005). Aging neuromodulation impairs associative binding: A neurocomputational account. *Psychological Science, 16*(6), 445–450.
- Light, L. L., & Anderson, P. A. (1983). Memory for scripts in young and older adults. *Memory & Cognition, 11*(5), 435–444.
- Light, L. L., & Singh, A. (1987). Implicit and explicit memory in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 531.
- Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*(3), 312–319.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*, (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Madden, D. J. (1986). Adult age differences in visual word recognition: Semantic encoding and episodic retention. *Experimental Aging Research, 12*(2), 71–78.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252–271.
- May, C. P., Hasher, L., & Stoltzfus, E. R. (1993). Optimal time of day and the magnitude of age differences in memory. *Psychological Science, 4*(5), 326–330.
- McNicol, D. (1972). *A primer of signal detection theory*. London: George Allen & Unwin.
- Meyer, A. N., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*(1), 142.
- Murphy, C., Nordin, S., & Acosta, L. (1997). Odor learning, recall, and recognition memory in young and elderly adults. *Neuropsychology, 11*(1), 126.
- Myerson, J., Hale, S., Wagstaff, D., Poon, L. W., & Smith, G. A. (1990). The information-loss model: A mathematical theory of age-related cognitive slowing. *Psychological Review, 97*(4), 475.
- 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.
- Naveh-Benjamin, M., Brav, T. K., & Levy, O. (2007). The associative memory deficit of older adults: The role of strategy utilization. *Psychology and Aging, 22*(1), 202–208.
- Naveh-Benjamin, M., Craik, F. I. M., Guez, J., & Kreuger, S. (2005). Divided attention in younger and older adults: Effects of strategy and relatedness on memory performance and secondary task costs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 520.
- Naveh-Benjamin, M., Shing, Y. L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S.-C. (2009). Adult age differences in memory for name–face associations: The effects of intentional and incidental learning. *Memory, 17*(2), 220–232.
- Neely, K. A., Samimy, S., Blouch, S. L., Wang, P., Chennavasin, A., Diaz, M. T., & Dennis, N. A. (2017). Memory-guided force control in healthy younger and older adults. *Experimental Brain Research, 235*(8), 2473–2482.
- Nyberg, L., Maitland, S. B., Rönnlund, M., Bäckman, L., Dixon, R. A., Wahlin, Å., & Nilsson, L.-G. (2003). Selective adult age differences in an age-invariant multifactor model of declarative memory. *Psychology and Aging, 18*(1), 149–160.
- Old, S. R., & 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.
- Pastötter, B., & Bäuml, K.-H. T. (2019). Testing enhances subsequent learning in older adults. *Psychology and Aging, 34*(2), 242.
- Perlmutter, M. (1978). What is memory aging the aging of? *Developmental Psychology, 14*(4), 330.
- Perlmutter, M. (1979). Age differences in adults' free recall, cued recall, and recognition. *Journal of Gerontology, 34*(4), 533–539.
- Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1591–1607.
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Raijmakers, J. G., & Shiffrin, R. M. (2002). Models of memory. In Pashler, H., & Medin, D. (Eds.) *Stevens' handbook of experimental psychology: Memory and cognitive processes*. (3rd ed., pp. 43–76). New York: Wiley.
- Rhodes, S., Cowan, N., Parra, M. A., & Logie, R. H. (2018). Interaction effects on common measures of sensitivity: Choice of measure, type I error, and power. *Behavior Research Methods, online ahead of print*.
- Riggs, D. S., Guarneri, J. A., & Addelman, S. (1978). Fitting straight lines when both variables are subject to error. *Life Sciences, 22*(13–15), 1305–1360.
- Roediger, H. L. III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.
- Rohling, M. L., Ellis, N. R., & Scogin, F. (1991). Automatic and effortful memory processes in elderly persons with organic brain pathology. *Journal of Gerontology, 46*(4), P137–P143.
- Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28*(6), 907–922.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.
- Ruiz-Gallego-Largo, T., Simón, T., & Suengas, A. G. (2012). Age differences in source monitoring and referent discrimination. *The Spanish Journal of Psychology, 15*(1), 20–28.
- Russo, R., & Parkin, A. J. (1993). Age differences in implicit memory: More apparent than real. *Memory & Cognition, 21*(1), 73–80.
- Salthouse, T. A. (2000). Methodological assumptions in cognitive aging research. In Craik, F. I. M., & Salthouse, T. A. (Eds.) *The handbook of cognitive aging*. (2nd ed., pp. 467–498). Mahwah, NJ: Erlbaum.
- Schmitter-Edgecombe, M., & Simpson, A. L. (2001). Effects of age and intentionality on content memory and temporal memory for performed activities. *Aging, Neuropsychology, and Cognition, 8*(2), 81–97.
- Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology, 20*(2), 228–236.
- Schonfield, D., Davidson, H., & Jones, H. (1983). An example of age-associated interference in memorizing. *Journal of Gerontology, 38*(2), 204–210.
- Schramke, C. J., & Bauer, R. M. (1997). State-dependent learning in older and younger adults. *Psychology and Aging, 12*(2), 255.
- Sliwinski, M. J., & Hall, C. B. (1998). Constraints on general slowing: A meta-analysis using hierarchical linear models with random coefficients. *Psychology and Aging, 13*(1), 164.

- Sommers, M. S., & Huff, L. M. (2003). The effects of age and dementia of the Alzheimer's type on phonological false memories. *Psychology and Aging, 18*(4), 791.
- Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: A meta-analysis. *Psychology and Aging, 10*(4), 527.
- Spilich, G. J., & Voss, J. F. (1983). Contextual effects upon text memory for young, aged-normal, and aged memory-impaired individuals. *Experimental Aging Research, 9*(1), 45–49.
- Surprenant, A. M., & Neath, I. (2009). Principles of memory. *Psychology Press*, Taylor & Francis.
- Trelle, A. N., Henson, R. N., & Simons, J. S. (2015). Identifying age-invariant and age-limited mechanisms for enhanced memory performance: Insights from self-referential processing in younger and older adults. *Psychology and Aging, 30*(2), 324.
- Uttl, B., Henry, M., & Baltimore, K. (2007). Are smaller age differences on old/new recognition versus free recall tests artifacts of easy memory tests? *Canadian Journal of Psychology, 61*, 374.
- Vakil, E., Melamed, M.-D., & Even, N. (1996). Direct and indirect measures of contextual information: Older versus young adult subjects. *Aging, Neuropsychology, and Cognition, 3*(1), 30–36.
- Verhaeghen, P. (2013). *The elements of cognitive aging: Meta-analyses of age-related differences in processing speed and their consequences*. London: Oxford University Press.
- Verhaeghen, P., & Marcoen, A. (1993). Memory aging as a general phenomenon: Episodic recall of older adults is a function of episodic recall of young adults. *Psychology and Aging, 8*(3), 380–388.
- Verhaeghen, P., Steitz, D. W., Sliwinski, M. J., & Cerella, J. (2003). Aging and dual-task performance: A meta-analysis. *Psychology and Aging, 18*(3), 443–460.
- Verhaeghen, P., Vandenbroucke, A., & Dierckx, V. (1998). Growing slower and less accurate: Adult age differences in time-accuracy functions for recall and recognition from episodic memory. *Experimental Aging Research, 24*, 3–19.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. Retrieved from <http://www.jstatsoft.org/v36/i03/>
- White, N., & Cunningham, W. R. (1982). What is the evidence for retrieval problems in the elderly? *Experimental Aging Research, 8*(3), 169–171.
- Whiting, I., Wythe, L., & Smith, A. D. (1997). Differential age-related processing limitations in recall and recognition tasks. *Psychology and Aging, 12*(2), 216–224.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.
- Zelinski, E. M., & Miura, S. A. (1988). Effects of thematic information on script memory in young and old adults. *Psychology and Aging, 3*(3), 292.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.