Abstract: The present meta-analysis investigated the influence of age on face recognition. A total of 19 studies with 79 comparisons of younger and older participants were included. Analyses revealed small to moderate effects for hits, and large effects for false alarms and signal detection theory (SDT) measures. Younger participants outperformed older participants on most face recognition measures. Younger participants made more hits (gu = 0.31) and fewer false alarms (gu = 0.95), thus had better SDT recognition performance (gu = 1.01) than older participants. These effects were largest for young faces, smaller for mixed-age faces, and smallest for older faces. Furthermore, older participants used a more liberal response criterion, that is, they were more likely to choose a face than younger participants (gu = 0.54). Meta-regression analyses revealed that young faces (vs. mixed-age faces) and longer retention intervals were associated with greater differences between the age groups for hits but not for false alarms. Funnel plot and trim-and-fill analyses indicated the presence of a publication bias. Theoretical implications for future research and for older people as eyewitnesses are outlined.

DOI: http://dx.doi.org/10.1037/pag0000282
Memory for Faces in Old Age:
A Meta-Analysis
Natalie Martschuk and Siegfried L. Sporer
Justus-Liebig University of Giessen,
Germany

Corresponding Author:
Natalie Martschuk,
Faculty of Business, Justice, and Behavioural Sciences,
Charles Sturt University,
Manly, Australia
Email: nmartschuk@csu.edu.au;

Siegfried L. Sporer,
Department of Psychology,
Faculty 06 Psychology and Sports Science,
Justus-Liebig University of Giessen,
Giessen, Germany,
Email: Siegfried.L.Sporer@psychol.uni-giessen.de.

Acknowledgements
We are grateful to Dr. Kerstin Kocab for her contribution to the early stages of the meta-analysis including literature research, coding scheme development, and coding of the primary studies. We thank Christian Meissner for commenting on a section of the manuscript and to Stephen Doherty for proof-reading the manuscript. This research is a portion of the first author’s Ph.D. thesis. Natalie Martschuk is now a Research Associate at Charles Sturt University, Manly, Australia.
Memory for Faces in Old Age: A Meta-Analysis

Abstract

The present meta-analysis investigated the influence of age on face recognition. A total of 19 studies with 79 comparisons of younger and older participants were included. Analyses revealed small to moderate effects for hits, and large effects for false alarms and signal detection theory (SDT) measures. Younger participants outperformed older participants on most face recognition measures. Younger participants made more hits ($g_u = 0.31$) and fewer false alarms ($g_u = 0.95$), thus had better SDT recognition performance ($g_u = 1.01$) than older participants. These effects were largest for young faces, smaller for mixed-age faces, and smallest for older faces. Furthermore, older participants used a more liberal response criterion, that is, they were more likely to choose a face than younger participants ($g_u = 0.54$). Meta-regression analyses revealed that young faces (vs. mixed-age faces) and longer retention intervals were associated with greater differences between the age groups for hits but not for false alarms. Funnel plot and trim-and-fill analyses indicated the presence of a publication bias. Theoretical implications for future research and for older people as eyewitnesses are outlined.

Keywords: face recognition; memory; older adults; meta-analysis; signal detection theory.
 MEMORY FOR FACES IN OLD AGE: A META-ANALYSIS

Older people often complain about memory problems in everyday life. Forgetting other people's names or faces can be particularly embarrassing. Most research on memory of older people has been conducted on verbal memory. Here, we focus on visual memory, in particular, recognition memory of unfamiliar faces. In criminal cases, misidentifications have been one of the main contributing factors in wrongful convictions (www.innocenceproject.org).

Current State of Knowledge

To date, narrative literature reviews found evidence for age effects in face recognition and eyewitness identification studies (see Bartlett, 2014a; Bartlett & Memon, 2007). However, no formal meta-analysis has been conducted to statistically test the size of these effects on recognition of faces by young vs. older adults.

Face Recognition Studies

Typically, facial recognition studies have two phases, a presentation phase and a recognition phase. During the presentation phase, a large number of faces is presented, participants are told to remember them (intentional learning) or to judge them with respect to some characteristics (e.g., attractiveness, perceived age; i.e., an incidental learning paradigm). At recognition, previously seen stimuli are mixed with an (equal) number of distractors and participants judge them as "old" = previously seen or "new" = distractors.

Face researchers have repeatedly emphasized the use of real faces varying in pose and expression (e.g., Bruce, 1982; Burton, 2012; Sporer, 1992) that are not identical at study and test, and perhaps not even taken with the same camera (Burton, 2012). In our meta-analysis, we have coded studies to the extent they meet these desirable features.

Here, the focus is on unfamiliar faces, that is, faces of people not known to participants prior to the experiment. Using large numbers of faces allows researchers to test general principles of this type of stimuli as well as variations between stimuli based on perceived face qualities like attractiveness, distinctiveness, and memorability (e.g., Bainbridge, Isola, & Oliva, 2013) but also gender, race, and most pertinent for our review, age.
The goal of the present meta-analysis was to test age effects in memory for faces in face recognition studies, without the context of a complex event, by comparing recognition of younger (< 30 years) versus older participants (> 60 years) with faces of different age groups. Although we acknowledge from the outset that age-related declines may be continuous across the lifespan, we chose these age cut-off points in line with the great majority of studies. Whenever possible, however, we consider also more fine-grained distinctions between different age groups among the older adult group (cf. Bartlett, 2014b). Further, although other meta-analyses compared the performance of children and adults, the present meta-analysis focused on young and older adults only, mainly because differences in cognitive and memory processes between children and adults on the one hand and adults and older people on the other hand are not necessarily comparable.

**Age-related Changes of Memory: Neurological and Cognitive Aspects**

Ageing is often discussed as a deteriorative process. Advanced adult ageing is associated with systematic reductions in the efficiency of the sensory system (Schieber, 2006). Myriad age-related changes occur in the eye, the retina, and the ascending visual pathways in the nervous system. Further, reductions in the efficiency of the sensory system negatively affect visual acuity, contrast sensitivity, spatial vision, and speed perception, resulting in reduced encoding of information. These deficiencies, in turn, are thought to affect face recognition (Lott, Haegerstrom-Portnoy, Schneck, & Brabyn, 2005). Age-related changes of memory have also been attributed to age-related changes of brain structure (Coleman & O'Hanlon, 2008).

The main theoretical frameworks of memory in old age describe the memory processes and age-related changes from different perspectives. According to the *limited processing resources approach* (e.g., Craik, 1986), the available resources for encoding and retrieval of information are limited in older people and used in a different way. Particularly, older people's processing of resource-demanding information seems to be limited, especially when self-initiated processing is required (Maylor, 2005). An implication of this theoretical approach applied to face recognition would be that older people are less able to engage in "deeper" or more "elaborate" processing of faces which has been demonstrated to aid younger
people in the recognition of unfamiliar faces (Bower & Karlin, 1974; Sporer, 1991). Relevant data were provided by Smith and Winograd (1978) who found that both young and older participants showed better recognition performance when judging the friendliness of a face than when judging the size of the nose or after a standard intentional learning instruction. Although the college student group had a higher $d'$ score (2.46) than the two groups of older participants (high school educated: 1.58; college educated: 1.51) the interaction between encoding instructions and age was not significant. Thus, we need to be careful to apply theories developed with verbal material and tested with recall paradigms to face recognition. Interestingly, in Smith and Winograd's study older people showed a more lenient response criterion beta (resulting in more false alarms) than younger people. We will return to this point below.

Another widely cited explanation, the reduced processing speed hypothesis (Salthouse, 1996), proposed that cognitive performance deteriorates with the decrease of mental speed in old age because (a) relevant tasks cannot be accomplished in limited time, and (b) relevant information presented simultaneously may no longer be available when processing is completed. Consequently, older people show poor performance in memory even when given unlimited time (Maylor, 2005). Accordingly, when retrieval of information fails because of reduced processing speed, people would be more inclined to produce inaccurate memories. Support for this hypothesis was found in a cross-sectional study that examined event-related correlates of face recognition memory of 19 to 80 year old participants (Wolff, Wiese, & Schweinberger, 2012). This study demonstrated age-related slowing of face-processing speed beginning in early adulthood (aged 30 to 44) compared to a younger group (aged 19 to 29), with a steeper decline after the age of 60. Consequently, we would expect older adults to show weaker face recognition performance, regardless of the length of time allowed for encoding.

In more applied settings, the associated deficit hypothesis (Naveh-Benjamin, 2000) has been invoked to explain deficits of older compared to young people. According to this hypothesis, older people have difficulties linking a seen face with a name, an action or a specific context (e.g., Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Old & Naveh-Benjamin,
Memory for faces in old age: A meta-analysis

Here we review only studies using recognition of faces irrespective of context cues, so this hypothesis appears less relevant.

Despite differences in emphasis, these theoretical frameworks converge in the sense that they both assume a reduction in information processing with increasing age which is associated with weaker cognitive performance in older age. Moreover, Craik and Rose (2012) argued that the reduced processing speed and efficiency are the consequences of limited attentional resources. A meta-analysis, which in principle is correlational, cannot test which of these theories better fits the available data.

Social-Cognitive Aspects and Own-age Bias

When examining recognition memory for faces, it is important to consider both cognitive (above) and social-psychological approaches. As proposed in Sporer’s (2001) in-group/out-group model, people are better in recognizing faces belonging to their in-group than to an out-group. Group membership may be determined by one's ethnicity, gender, or age, and respective in-group performance advantages have been demonstrated (own-race, own-sex, own-age bias: Herlitz & Lovén, 2013; Meissner & Brigham, 2001; Rhodes & Anastasi, 2012; Shapiro & Penrod, 1986; Sporer, 2001; Wright & Sladden, 2003; Wright & Stroud, 2002). According to Sporer's model, when observers encounter an in-group person, they process his or her face in a rather automatic fashion. This entails configural, holistic processing which has been demonstrated to lead to better recognition performance. In contrast, with out-group faces, observers quickly label the face as an out-group member, which is considered of little personal interest for future interactions (cf. Malpass, 1990, social utility explanation). Thus, no further processing is engaged in, which lowers later recognition performance.

To our knowledge, this model is the only theoretical approach that predicts not only a deterioration in performance but also a more lenient response criterion for out-group faces. Accordingly, out-group faces are perceived as more similar to each other, that is, more "homogeneous" (out-group homogeneity effect: Quattrone & Jones, 1980). Encoding does not entail a search for individuating, distinctive features but rather relies on stereotypical features (e.g., "a skinhead; long hair; wrinkled face"). As these features are rather unspecific, they are
not very helpful at the recognition stage. At recognition, the labeling process enlarges the region of acceptance used to designate a face as previously seen. Thus, more faces will be considered as previously seen ("Yes" responses), irrespective of whether they appeared in the study list, that is, an increase in response bias (for a detailed description of the model and supporting evidence, see Sporer, 2001).

From a related theoretical perspective, older people may rely upon a generalized sense of familiarity (Hancock, Bruce, & Burton, 2000; Vokey & Read, 1992), which is considered sufficient to respond with "Yes, previously seen". This is tantamount to relying on a larger radius of acceptance in multiple dimension face space models (e.g., Valentine, 1991; see Sporer, 2001, for review). Consequently, a more lenient response bias would be expected with older people (main effect of participant age = age effect). To the extent that out-group faces are perceived as more homogeneous, this response bias of older people may be even more pronounced when recognizing young faces (Participant Age x Face Age interaction).

**Empirical evidence for these models.**

According to these categorization-individuation models, people tend to categorize and process faces of in- versus out-groups differently (Hugenberg, Miller, & Claypool, 2007; Sporer, 2001; Wiese, Komes, & Schweinberger, 2012). Accordingly, people may be less motivated to process out-group faces individually, perhaps even disregard them cognitively (Rodin, 1987; Sporer, 2001) and process them at a categorical level (Wiese et al., 2012).

In a series of five studies Rodin (1987) gained support for her *cognitive-disregard theory*, using both recall and recognition measures. Although youthfulness per se enhanced memorability, age-discrepancy acted as a disregard cue, which led to inferior face recognition of targets discrepant in age both in a laboratory experiment (Experiment 2)\(^1\) and in a field study after a brief encounter (Experiment 3). Attractiveness as well as gender of targets moderated the effect (Experiment 4). Experiment 5 indicated that the effect was not restricted to recognition but affected also recall of personal attributes.

\(^1\) Unfortunately, not enough statistical data were presented to include these studies in our meta-analysis.
More direct evidence for visual attentional processes at encoding comes from a study that examined recognition performance and visual scan patterns of younger and older people while they were observing faces of different age groups (He, Ebner, & Johnson, 2011). Independent of their age, people tended to look longer at faces of their own age group than at other-age faces. Participants also reported to have more contact with people of their own-age than the other-age group. Consequently, people were better in recognizing own-age than other-age faces. Interestingly, this study demonstrated that self-reported amount of contact with young and older people predicted face recognition (He et al., 2011).

The latter findings are compatible with the contact hypothesis, namely that recognition performance is related to the amount and quality of social contacts with the particular group (Meissner & Brigham, 2001; Sporer, 2001). Consequently, younger and older people would be expected to recognize faces of their own group better as a function of contact and experience with their own-age than with other-age groups.

Additional support for this assumption comes from a quasi-experimental study with geriatric nurses who had frequent contact with older people (Wiese, Wolff, Steffens, & Schweinberger, 2013). In contrast to a control group, these nurses were equally likely to recognize old-age and own-age faces. A comparable pattern was found with trainee teachers for child faces (Harrison & Hole, 2009; cf. also the field study on the own-race bias of White teachers tested before and after a teaching visit to a Black African community, cited in Valentine, Chiroro, & Dixon, 1995).

On the other hand, one could argue that older people have had a life-long experience with younger people and in later years with older people while younger people may predominantly have interacted with younger people. Consequently, younger people may show deficits with older relative to younger faces while older people may not show such differences as a function of face age (e.g., Short, Semplonius, Projetti, & Mondloch, 2014). In other words, there should be a disordinal, not a cross-over interaction (see Sporer, 2001). Contrary to this assumption, in Rhodes and Anastasi’s (2012) meta-analysis the own-age bias was present for hits and discriminability independently of participant age across the lifespan, beginning in early childhood until old age. An own-age bias in false alarms was observed.
with young adults only. This suggests that not the life-long, but the more recent experience may be crucial. Another possibility is that factors other than the own-age bias might account for people’s proneness to a particular response bias, for instance stable "cognitive traits" (Kantner & Lindsay, 2012).

**Face Recognition of Unfamiliar vs. Familiar Faces**

Neuropsychological studies have demonstrated that certain factors may affect processing and recognition of familiar and unfamiliar faces differently (see Johnston & Edwards, 2009, for a review). For instance, viewpoint, facial expression and context influence recognition of unfamiliar but not of familiar faces. Unfamiliar faces, unlike familiar faces, seem to be encoded in a viewpoint dependent manner, and thus their recognition is sensitive to image changes. To study face recognition, not just picture recognition, stimuli should not be identical at study and test (Bartlett, 2014b; Bruce, 1982; Sporer, 1992). Studies which exclusively relied on picture recognition or used unnaturally cropped or computer-generated faces were excluded (e.g., Germine, Duchaine, & Nakayama, 2011; Hildebrandt, Wilhelm, Schmiedek, Herzmann, & Sommer, 2011).

To test the importance of face changes, we coded studies with respect to changes (e.g., expression and/or view/pose) between encoding and test and conducted moderator analyses with face change as a predictor (see Table 1 for an overview of study characteristics coded).
<table>
<thead>
<tr>
<th>Study</th>
<th>Mean age of older age group</th>
<th>Face Age</th>
<th>Encoding task</th>
<th>Encoding time (s)</th>
<th>Retention interval (min)</th>
<th>No. of faces at encoding</th>
<th>No. of faces (distractors) at recognition</th>
<th>Gender of faces</th>
<th>Change of faces at test</th>
<th>Recognition task</th>
<th>Visual acuity</th>
<th>Health screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Anastasi &amp; Rhodes (2006, Exp. 1)</td>
<td>65.7</td>
<td>Young vs. Middle-aged vs. Old</td>
<td>Rating of attractiveness</td>
<td>7</td>
<td>18</td>
<td>24</td>
<td>48 (24)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N</td>
<td>N/A</td>
<td>SR</td>
</tr>
<tr>
<td>1.2 Anastasi &amp; Rhodes (2006, Exp. 2)</td>
<td>69.5</td>
<td>Young vs. Middle-aged vs. Old</td>
<td>Categorizing age groups</td>
<td>7</td>
<td>2 days</td>
<td>24</td>
<td>48 (24)</td>
<td>Mixed</td>
<td>Expression</td>
<td>Y/N</td>
<td>N/A</td>
<td>SR</td>
</tr>
<tr>
<td>2.2 Bartlett &amp; Fulton (1991, Exp. 2)</td>
<td>69.0</td>
<td>Mixed-age</td>
<td>Rating of pleasantness</td>
<td>10</td>
<td>0</td>
<td>48</td>
<td>72 (24)</td>
<td>Mixed</td>
<td>Expression</td>
<td>Y/N; Familiarity</td>
<td>Snellen chart 20/40</td>
<td>WAIS</td>
</tr>
<tr>
<td>3.1 Bartlett et al. (1989, Exp 1)</td>
<td>66.7</td>
<td>Mixed-age</td>
<td>Rating of pleasantness</td>
<td>10</td>
<td>0</td>
<td>48</td>
<td>72 (24)</td>
<td>Female</td>
<td>Expression, pose</td>
<td>IRJ</td>
<td>Snellen chart 20/40</td>
<td>WAIS</td>
</tr>
<tr>
<td>3.2 Bartlett et al. (1989, Exp 2)</td>
<td>69.6</td>
<td>Mixed-age</td>
<td>Rating of pleasantness</td>
<td>10</td>
<td>0</td>
<td>48</td>
<td>72 (24)</td>
<td>Female</td>
<td>Expression, pose</td>
<td>IRJ</td>
<td>Snellen chart 20/40</td>
<td>WAIS</td>
</tr>
<tr>
<td>5.1 Bastin &amp; van der Linden (2003)</td>
<td>64.4</td>
<td>Mixed-age</td>
<td>Intentional</td>
<td>Unlimited</td>
<td>30</td>
<td>18</td>
<td>36 (18)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N (or 2-AFC)</td>
<td>N/A</td>
<td>HT, SR, ST</td>
</tr>
<tr>
<td>7.1 Firestone et al. (2007)</td>
<td>70.5</td>
<td>Young vs. Old</td>
<td>Rating of photo quality, age of face</td>
<td>7</td>
<td>5</td>
<td>24</td>
<td>48 (24)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N</td>
<td>SR</td>
<td>PS</td>
</tr>
</tbody>
</table>
## Table 1 continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Mean age of older age group</th>
<th>Face Age</th>
<th>Encoding task</th>
<th>Encoding time (s)</th>
<th>Retention interval (min)</th>
<th>No. of faces at encoding</th>
<th>No. of faces (distractors) at recognition</th>
<th>Gender of faces</th>
<th>Change of faces at test</th>
<th>Recognition task</th>
<th>Visual acuity</th>
<th>Health screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fulton &amp; Bartlett (1991)</td>
<td>71.4</td>
<td>Young vs. Middle-aged vs. Old</td>
<td>Rating of pleasantness</td>
<td>10</td>
<td>0</td>
<td>24</td>
<td>72 (48)</td>
<td>Mixed</td>
<td>Expression, pose</td>
<td>IRJ</td>
<td>Snellen chart 20/40</td>
<td>SR, WAIS</td>
</tr>
<tr>
<td>Lamont et al. (2005)</td>
<td>74.0</td>
<td>Young vs. Old</td>
<td>Intentional</td>
<td>5</td>
<td>0</td>
<td>30</td>
<td>60 (30)</td>
<td>Male</td>
<td>Identical</td>
<td>Y/N</td>
<td>RTT</td>
<td>N/A</td>
</tr>
<tr>
<td>Mason (1986)</td>
<td>75.0</td>
<td>Young vs. Old</td>
<td>Face-name pairs</td>
<td>10</td>
<td>0</td>
<td>40</td>
<td>80 (40)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N, name recognition</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Naveh-Benjamin et al. (2009, Exp. 1)</td>
<td>72.7</td>
<td>Mixed-age</td>
<td>Face-name pairs; intentional</td>
<td>3</td>
<td>0</td>
<td>24</td>
<td>16 (8)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N, name recognition</td>
<td>N/A</td>
<td>SR</td>
</tr>
<tr>
<td>Naveh-Benjamin et al. (2009, Exp. 2)</td>
<td>73.2</td>
<td>Mixed-age</td>
<td>Fit of face-name pairs; incidental</td>
<td>3</td>
<td>0</td>
<td>48</td>
<td>32 (16)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N, name recognition</td>
<td>N/A</td>
<td>SR</td>
</tr>
<tr>
<td>Olofsson &amp; Bäckman (1996)</td>
<td>73.8</td>
<td>Mixed-age</td>
<td>Face-name pairs</td>
<td>10</td>
<td>15</td>
<td>16</td>
<td>32 (16)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N, name recognition</td>
<td>N/A</td>
<td>SR, VT</td>
</tr>
<tr>
<td>Savaskan et al. (2007)</td>
<td>67.3</td>
<td>Mixed-age</td>
<td>Intentional</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>60 (30)</td>
<td>Mixed</td>
<td>Expression</td>
<td>Y/N, expression recognition</td>
<td>N/A</td>
<td>GDS, MMSE, SR</td>
</tr>
<tr>
<td>Smith &amp; Winograd (1978)</td>
<td>65.0</td>
<td>Mixed-age</td>
<td>Rating of nose/friendliness/Intentional</td>
<td>8</td>
<td>0</td>
<td>50</td>
<td>60 (30)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Bryce &amp; Dodson (2013)</td>
<td>72.0</td>
<td>Young vs. Old</td>
<td>Intentional</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>24 (12)</td>
<td>Mixed</td>
<td>Expression</td>
<td>Y/N, RKG</td>
<td>N/A</td>
<td>BDI, SR, VT</td>
</tr>
<tr>
<td>He et al. (2011)</td>
<td>73.9</td>
<td>Young vs. Old</td>
<td>Incidental (inspect pictures)</td>
<td>4</td>
<td>10</td>
<td>48</td>
<td>96 (48)</td>
<td>Mixed</td>
<td>Identical</td>
<td>Y/N</td>
<td>MARS, RPVS</td>
<td>SR, DSST</td>
</tr>
</tbody>
</table>
Memory for faces in old age: A meta-analysis

*Note.* Face Age: Young; Mixed = Young, (Middle-aged) and Old; Young vs. Middle-aged vs. Old = age of faces as a repeated measurement factor; Young vs. Old = both as a repeated measurement factor; Lamont et al. (2006) varied the number of faces as a between-participants factor (encoding: 20 vs. 40; test: 40 vs. 80); 2-AFC: two-alternative forced choice test; BDI: Beck Depression Inventory (Beck, Rush, Shaw, & Emery, 1979); DSST: Digit-Symbol-Substitution Test: measures visual-motor processing speed (Wechsler, 1981); GDS: Geriatric Depression Scale; GMT: Guild Memory Test (Gilbert, Levee, & Catalano, 1974); HT: Hayling Test (Burgess & Shallice, 1996); IRJ: Identity/recognition judgment with identical/old, old-but changed [expression, pose], new; MARS: MARS Letter Contrast Sensitivity Test (Arditi, 2005); MMSE: Mini-Mental State Examination; N/A: not available; PS: Prescreening through the Rothman Research Institute volunteer pool; RKG: Recognition with remember/know/guess judgment; RPVS: Rosenbaum Pocket Vision Screener (Rosenbaum, Granham-Field Surgical Co Inc, New York, NY); RTT: Reading Test types (Faculty of Ophthalmologists, London, 1987); SR: Self-report; ST: Verbal fluency: modification of the Stroop test (Stroop, 1935); VT: Vocabulary test; WAIS: Vocabulary test of the second half of the Wechsler Adult Intelligence Scale (Wechsler, 1981); Y/N: Yes-No recognition test.
Signal Detection Theory

To measure memory performance and response bias, many studies rely on signal detection theory (SDT; e.g., Macmillan & Creelman, 2005). SDT is based on the assumption that the degree of perceived familiarity (i.e., degree of memory strength) of a new face and a previously seen face are discriminable from each other (O’Toole, Bartlett, & Abdi, 2000). Therefore, the experienced level of familiarity should be higher for a previously seen face compared to a new face. Observers can either correctly recognize a previously seen face as "Old" (hit) or incorrectly reject it as "New" (false rejection). Similarly, they can either correctly reject (correct rejection) or incorrectly recognize a "New" face as previously seen (false alarm). Hits and false alarms suffice to describe recognition performance, given that false and correct rejections are merely complementary measures to hits and false alarms (Macmillan & Creelman, 2005).

Hits and false alarms can be integrated into combined measures of recognition performance or "sensitivity", usually $d'$. A $d'$ of zero denotes no discrimination, and continuously increasing values of, for example, 1.0, 2.0 or 3.0 (to infinity) indicate stronger recognition performance. Besides sensitivity, people also differ in response tendencies. Most authors who did analyze response biases reported the criterion measure $c$, which is positive when hits exceed false alarms, negative when false alarms exceed hits, and zero if there is no difference (see Macmillan & Creelman, 2005). From an applied perspective, both a person's recognition performance as well as his or her response bias are of importance. For instance, a study that examined recognition memory of middle-aged adults (35-49 years) and older adults (75-89 years) using verbal learning test materials demonstrated that older adults showed a more liberal response criterion than their younger counterparts (Huh, Kramer, Gazzaley, & Delis, 2006). Further, inhibition--an executive function associated with the frontal brain lobe--was a significant predictor of response bias, suggesting a more liberal response criterion with reduced control in inhibition among older adults.

Mirror effect.

One way to test whether age affects only performance or also response bias is to test for a "mirror effect" (Glanzer & Adams, 1985) of hits and false alarms. A mirror effect describes
the pattern of results in a Yes-No recognition experiment where a stimulus (here: a face) is accurately recognized as previously seen when it had been presented before (a hit), and accurately rejected as new when it had not been seen at presentation. Although some studies did not find such mirror effects in face recognition (e.g., Hancock et al., 2000; Vokey & Read, 1992) or face matching studies (Megreya & Burton, 2007), in a large scale meta-analysis of the own-race bias Meissner and Brigham (2001) found such a symmetrical pattern—although the effect sizes for false alarms were almost twice as large as those for hits.

Some authors have also postulated (and found) changes in response criterion along with a mirror effect (Hirshman, 1995; McClelland & Chappell, 1998). In line with Meissner and Brigham's assumption of a mirror effect and a concomitant change in response criterion, we also expected to find both effects in our meta-analysis.

One way to test for mirror effects is to examine the patterns of hits and false alarms separately for young and older participants. Unfortunately, researchers did not report the respective intercorrelations between dependent variables, so no meta-analytic synthesis of correlations was possible. Hence, we had to rely on visual inspection of unweighted means for hits and false alarms.

**Summary of Hypotheses of the Present Meta-analysis**

The present meta-analysis tested an overall age effect in face recognition across studies that compared young adults with an older age group, focusing on natural, unfamiliar faces. In line with a general age-related decline in neurological and memory processes considered to be involved in face recognition (Bartlett, 2014a; Coleman & O’Hanlon, 2008; Maylor, 2005) we predicted that across all studies, older people show decreased face recognition performance compared to younger people (i.e., an age effect). This should be reflected in both decreased hits and increased false alarms as well as in smaller $d'$ values in older age. This age effect may be smaller in the recognition of older compared to young faces (own-age bias). As predicted from Sporer's (2001) in-group/out-group model, we expected a more liberal response criterion of older compared to younger people. Finally, we hypothesized that variables that make the task more difficult for participants (i.e., number of faces to be recognized, short encoding times, face changes between study and test, longer retention
interval, additional tasks at test) would be particularly detrimental for old age participants and should hence be positively associated with effect sizes of the old-age effect.

**Method**

**Search Strategies**

Different search strategies were used to obtain studies for the current meta-analysis. Computerized search was conducted of PsycINFO, EDOC, MEDLINE, and Web of Science (Science Citation Index Expanded, Social Sciences Citation Index and Arts and Humanities Citation Index), using free text search and search engines, that is, ZPID, DGPs, Google Scholar, and Dissertation Abstracts. To identify all possible studies, different combinations of the keywords face recognition, identification, accuracy, age, elderly, old using Boolean operators "AND" or "OR", and the wildcard character "*" were used.

Further, manual searches of relevant journals and reference sections of published articles were conducted. Finally, after completing the literature research, authors of the existing studies were contacted to obtain unpublished studies. The literature search revealed a large number of studies, including both lineup and face recognition studies. The current meta-analysis is restricted to studies using a recognition paradigm showing a large number of target faces at encoding that were randomly mixed with distractors at testing, displaying one face at a time. Studies reporting only neuropsychological measures were excluded. Person identification studies using single targets that were exposed either live or in a film later to be recognized in a lineup were recently reviewed in a meta-analysis by Kocab, Martschuk and Sporer (2017).

To be included, studies had to compare at least two groups of participants, viz. young people above 16 years and older persons above 60 years. In the primary studies participants with any intellectual or psychiatric impairments had been excluded. Stimuli must have been photographs of strangers of young or older target faces, not of familiar people like celebrities. Experiments with transformed, inverted or computer-generated faces were excluded. Dependent variables were hits, false alarms, and/or signal detection theory measures of performance ($d'$, $A'$) or bias ($c$, beta). The reported statistics had to provide data to calculate effect sizes (e.g., proportions, frequencies, means and standard deviations, or other statistical
information such as $F$, $t$, or $p$ values). If sufficient data were not reported, authors were contacted for providing missing information. Although we searched for published and unpublished studies in English, German, French and Russian, only English studies were found. A total of 19 studies met the inclusion criteria, with $n_Y = 656$ young ($M_Y = 24.1$ years) and $n_O = 755$ older people ($M_O = 70.1$ years).

The following studies were excluded from the analyses: Bartlett and Fulton (1991, Exp. 1) and Flicker, Ferris, Crook, and Bartus (1990) used a continuous recognition paradigm: Participants were instructed to identify repeated faces in an ongoing series of faces without a clear distinction between the encoding and recognition phase; Bastin (2008), Bastin and van der Linden (2006), and Old and Naveh-Benjamin (2008) used a two-alternative-forced-choice recognition paradigm, that is, at test participants were presented with pairs of faces consisting of a target and a distractor, and were instructed to choose the previously studied face. In Bartlett, Strater and Fulton (1991, Exp. 2), the sample was the same as in their first experiment. In order not to violate the assumption of independence of effect sizes, Exp. 2 was excluded. Rodin (1987, Exp. 2) did not report enough information to calculate an effect size. Ebner and Johnson (2009) reported only results for corrected recognition but not for the dependent measures of interest for this meta-analysis.

**Study Characteristics and Coded Variables**

Table 1 shows the main characteristics of the included studies. To ensure reliability of coding, each study was coded independently by two coders. Inter-coder reliabilities for categorical variables were between $kappa = .66$ and $kappa = 1.00$. Intra-class correlation coefficients ($ICCs$) for continuous variables were between $ICC = .79$ and $ICC = 1.00$. Discrepancies were discussed until full agreement was reached. The most important study characteristics coded are discussed below.

**Encoding task.**

The encoding task was coded continuously (e.g., number of faces, exposure time in seconds) and categorically (e.g., instruction: intentional vs. incidental). In intentional encoding participants were informed of the impending recognition test beforehand. Incidental encoding tasks involved categorizing faces by age, rating their pleasantness, etc. Encoding
times varied between 1.5 (Bastin & van der Linden, 2003) and 12 seconds (Bartlett et al., 1991), with \( Mdn = 7 \) seconds. The number of faces at encoding was between 12 (D'Argembeau & van der Linden, 2004) and 50 (Smith & Winograd, 1978), \( Mdn = 24 \).

**Retention interval.**

The retention interval, that is, the time between encoding and recognition of the faces, varied between none (i.e., immediate; e.g., Bartlett & Fulton, 1991) and 48 hours (Anastasi & Rhodes, 2006, Exp. 2), \( Mdn = 0 \).

**Recognition task.**

In addition to the Yes-No recognition paradigm that requires participants to indicate whether each face had been seen before ("Yes") or not ("No") (e.g., Lamont, Steward-Williams, & Podd, 2005) some studies also tested the associations between names and faces (Naveh-Benjamin et al., 2009). Other studies tested the influence of change of expression or change of pose on memory for faces, instructing participants to classify each face as "identical", "changed", or "new".

Change of faces at test was coded with levels no change, change of appearance, pose, expression, or both expression and pose. Continuous variables were number of targets and number of distractors. The number of faces (targets and distractors) varied between 16 (Naveh-Benjamin et al., 2009) and 96 (He et al., 2011), \( Mdn = 48 \). Recognition of names, ratings of familiarity or of friendliness were not subject of the present meta-analysis and were therefore not analyzed.

**Participant screening.**

In the majority of studies (89.5%) participants were screened for visual acuity and/or with performance tests, dementia tests or tests for psychopathological problems (see Table 1).

**Data Analyses**

Separate effect sizes (Hedges \( g_u \) adjusted for sample bias; Borenstein, 2009) were calculated for hits, false alarms, and signal detection measures of performance (\( d' \)) and response bias (\( c, beta \)). For each dependent variable, one single effect size per study was calculated, if necessary averaging across experimental conditions. Some studies reported several dependent variables for the same participants, but not the correlations between them.
To avoid dependencies among effect sizes, only one of the measures was used (e.g., the effect size for changed faces but not the one for identical faces).

Homogeneity of effect sizes was tested using the $Q$ statistic (Hedges & Olkin, 1995), supplemented by the descriptive index of heterogeneity $I^2$ (Shadish & Haddock, 2009). This value indicates the proportion of effect size variation due to heterogeneity between studies rather than to sampling error. Recommendations to interpret values of $I^2$ are as follows: $I^2 = .25$ is considered as small, $I^2 = .50$ as medium, and $I^2 = .75$ as large heterogeneity (Higgins & Thompson, 2002). However, $I^2$ should not be interpreted in isolation but only in combination with the amount of true heterogeneity (see Borenstein, Higgins, Hedges, & Rothstein, 2017).

For the global estimate of effect sizes, both fixed effects models (FEM) and random effects models (REM) were calculated. The FEM assumes that a single effect size estimates the same fixed population parameter, with heterogeneity due to within-study sampling error. The REM assumes a random distribution of individual effect sizes that differ from the population mean by subject-level sampling error plus a random variance parameter (Lipsey & Wilson, 2001). Although we report both models, interpretation will focus on the REM results only.

Separate analyses were conducted for young faces, mixed-age faces (results for young and older faces not reported separately), and older faces. Some studies varied face age as a within-participants variable. These were, however, not combined in a single meta-analysis in order not to violate the independence assumption of effect size estimates.

To test for moderator effects, studies with young and mixed-age faces were analyzed together. Hierarchical weighted mixed-model (methods of moments) meta-regression analyses using macros developed by Lipsey and Wilson (2001) were conducted to test the effects of the mean age of the older age group and of face age (young vs. mixed-age) at step 1. At step 2, the number of faces at encoding, retention interval, change of pose and/or expression at test, and the presence of additional tasks beside face recognition were added to the model.
The nonparametric trim-and-fill method was applied to assess possible problems of publication bias (Duval & Tweedie, 2000a, 2000b; Sutton, 2009). Trim-and-fill analyses were carried out for young and mixed-age faces combined to increase the number of studies, and after removing outliers to avoid biased results. Effect sizes were plotted against standard errors to assess the distribution of the funnel plot (Sterne & Egger, 2001; Sutton, 2009). Based on the assumption that an asymmetric funnel plot arises from missing unfavorable and nonsignificant studies, the trim-and-fill method estimates the number of missing studies, imputes them in the funnel plot and estimates an “adjusted” effect size (Duval & Tweedie, 2000a, 2000b). Although trim-and-fill analyses could be carried out with an arbitrary amount of studies (Duval & Tweedie, 2000b), tests for funnel plot asymmetry would be unlikely to be useful in meta-analyses with small number of studies (Sterne, Becker, & Egger, 2005).

In addition to the graphical method, the statistical linear regression test was applied (Egger, Smith, Schneider, & Minder, 1997; Sutton, 2009). Thereby, the standard normal deviate is regressed against the inverse variance, which is similar to adjusting a line to Galbraith’s radial plot. We also calculated the commonly used failsafe N (Orwin, 1983; Rosenthal, 1979) due to its pioneering role. However, we advise of recent criticisms of failsafe N in the meta-analytic literature (see Becker, 2005).

Meta-analyses and meta-regression analyses were conducted using SPSS Macros provided by Lipsey and Wilson (2001) and cross-validated with self-programmed Excel macros. For trim-and-fill analyses and forest plots, the R package ‘metafor’ was used (Viechtbauer, 2010).

Results

Weighted Mean Effect Size Analyses

Figures 1 to 3 display the forest plots of individual effect sizes and their 95% confidence intervals (CIs), as well as the estimated mean effects of fixed and random effects models with and without outliers. Effect sizes for young, middle-aged and older faces are separated to avoid dependencies among effect sizes and to appreciate possible own-age effects.
Figure 1. Forest plot for hits and the mean weighted effect sizes with and without outlier.
**Figure 2.** Forest plot for false alarms and the mean weighted effect sizes with and without outlier.
Memory for faces in old age: A meta-analysis

<table>
<thead>
<tr>
<th>Author(s), Year (Exp.)</th>
<th>N</th>
<th>Type</th>
<th>M_Y</th>
<th>M_O</th>
<th>Z</th>
<th>VC</th>
<th>Q</th>
<th>$i^2$</th>
<th>$N_{ls}$</th>
<th>Hedges $g_u$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recognition Performance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young Faces</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamont et al., 2005</td>
<td>96</td>
<td>$d'$</td>
<td>2.03</td>
<td>1.35</td>
<td>0.90</td>
<td>0.95</td>
<td>0.45</td>
<td>1.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulton &amp; Barlett, 1991</td>
<td>69</td>
<td>$d'$</td>
<td>1.76</td>
<td>1.08</td>
<td>0.93</td>
<td>0.43</td>
<td>1.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He et al., 2011</td>
<td>49</td>
<td>$d'$</td>
<td>1.99</td>
<td>1.07</td>
<td>1.22</td>
<td>1.61</td>
<td>1.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firestone et al., 2007</td>
<td>48</td>
<td>$d'$</td>
<td>2.36</td>
<td>1.59</td>
<td>1.44</td>
<td>0.80</td>
<td>2.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mason, 1986</td>
<td>86</td>
<td>$d'$</td>
<td>3.45</td>
<td>1.02</td>
<td>1.99</td>
<td>1.47</td>
<td>2.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mixed-Faces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D'Argembeu &amp; van der Linden, 2004</td>
<td>64</td>
<td>$d'$</td>
<td>1.99</td>
<td>0.96</td>
<td>0.71</td>
<td>0.21</td>
<td>1.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bastin &amp; van der Linden, 2003</td>
<td>128</td>
<td>$d'$</td>
<td>1.70</td>
<td>1.10</td>
<td>0.86</td>
<td>0.50</td>
<td>1.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett et al., 1989 [1]</td>
<td>30</td>
<td>$d'$</td>
<td>2.18</td>
<td>1.67</td>
<td>0.93</td>
<td>0.17</td>
<td>1.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Winograd, 1978</td>
<td>160</td>
<td>$d'$</td>
<td>2.05</td>
<td>1.30</td>
<td>1.01</td>
<td>0.67</td>
<td>1.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Savaskan et al., 2007</td>
<td>45</td>
<td>$d'$</td>
<td>1.60</td>
<td>0.97</td>
<td>1.11</td>
<td>0.45</td>
<td>1.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett et al., 1989 [2]</td>
<td>32</td>
<td>$d'$</td>
<td>1.05</td>
<td>0.50</td>
<td>1.17</td>
<td>0.42</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olofsson &amp; Bäckman, 1996</td>
<td>96</td>
<td>$d'$</td>
<td>3.26</td>
<td>2.18</td>
<td>1.28</td>
<td>0.78</td>
<td>1.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old Faces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamont et al., 2005</td>
<td>96</td>
<td>$d'$</td>
<td>2.24</td>
<td>2.25</td>
<td>-0.01</td>
<td>-0.44</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He et al., 2011</td>
<td>49</td>
<td>$d'$</td>
<td>1.42</td>
<td>1.24</td>
<td>0.21</td>
<td>-0.02</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulton &amp; Barlett, 1991</td>
<td>69</td>
<td>$d'$</td>
<td>1.30</td>
<td>1.05</td>
<td>0.38</td>
<td>-0.11</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firestone et al., 2007</td>
<td>48</td>
<td>$d'$</td>
<td>2.59</td>
<td>2.16</td>
<td>0.65</td>
<td>0.07</td>
<td>1.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mason, 1986</td>
<td>86</td>
<td>$d'$</td>
<td>2.28</td>
<td>1.20</td>
<td>0.78</td>
<td>0.34</td>
<td>1.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Response Criterion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young and Mixed-Faces</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D'Argembeu &amp; van der Linden, 2004</td>
<td>64</td>
<td>$c$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.00</td>
<td>-0.49</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olofsson &amp; Bäckman, 1996</td>
<td>96</td>
<td>$c$</td>
<td>-0.01</td>
<td>-0.11</td>
<td>0.25</td>
<td>-0.21</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bastin &amp; van der Linden, 2003</td>
<td>128</td>
<td>$c$</td>
<td>0.20</td>
<td>0.11</td>
<td>0.28</td>
<td>-0.08</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firestone et al., 2007</td>
<td>48</td>
<td>$c$</td>
<td>0.23</td>
<td>0.01</td>
<td>0.27</td>
<td>-0.25</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Winograd, 1978</td>
<td>218</td>
<td>$t$</td>
<td>1.36</td>
<td>0.80</td>
<td>0.39</td>
<td>0.12</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulton &amp; Barlett, 1991</td>
<td>69</td>
<td>$c$</td>
<td>-0.12</td>
<td>-0.34</td>
<td>0.63</td>
<td>0.15</td>
<td>1.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett et al., 1991</td>
<td>39</td>
<td>$c$</td>
<td>N/A</td>
<td>N/A</td>
<td>0.72</td>
<td>0.07</td>
<td>1.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mason, 1986</td>
<td>86</td>
<td>$t$</td>
<td>2.61</td>
<td>0.72</td>
<td>1.17</td>
<td>0.40</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett et al., 1989 [1]</td>
<td>30</td>
<td>$c$</td>
<td>-0.24</td>
<td>-0.65</td>
<td>1.19</td>
<td>0.43</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartlett et al., 1989 [2]</td>
<td>30</td>
<td>$c$</td>
<td>-0.44</td>
<td>-0.81</td>
<td>0.71</td>
<td>0.61</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FE Model</strong> (k = 5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.46***</td>
<td>0.59</td>
<td>19.97</td>
<td>0.549</td>
<td>111</td>
<td>0.47</td>
<td>0.33</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RE Model</strong> (k = 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.52***</td>
<td>0.69</td>
<td>19.97</td>
<td>0.549</td>
<td>48</td>
<td>0.52</td>
<td>0.29</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3. Forest plot for recognition performance $d'$ and response criterion $c$ (or beta), and the mean weighted effect sizes with and without outliers.
Before synthesizing effect sizes, two different methods were applied to check for outliers for each dependent variable: (a) graphic representations of individual effect sizes (and 95% CIs), along with the weighted mean effect sizes (see the forest plots in Figures 1-3); (b) Hedges and Olkin's (1985, Chapter 12) method of calculating standardized residuals and homogeneity tests, excluding potential outliers consecutively using a self-programmed Excel spreadsheet.

Next, FEMs and REMs were calculated for young, mixed-age and older faces. Below we focus on the REMs, which were calculated both for all reported studies and after removal of outliers if necessary. Results without outliers provide a somewhat more conservative estimate. Figure 4 provides an overview of all mean effect sizes (with 95% CIs) using the REM without outliers.

**Figure 4.** Summary of mean weighted effect sizes (without outliers) using the random effects model for hits, false alarms, recognition performance and response bias for young, mixed-age and old-age faces.
Memory for faces in old age: A meta-analysis

Hits.

Overall, 26 hypothesis tests were identified reporting age effects on hits (8 for young, 10 for mixed-age, and 8 for older faces; Figure 1). Outlier analyses indicated one extreme case for young faces (Anastasi & Rhodes, 2006, Exp. 2, one for mixed-age faces (Naveh-Benjamin et al., 2009, Exp. 1), and two for older faces (Anastasi & Rhodes, 2006, Exp. 1 and 2; see Figure 1). The experiments by Anastasi and Rhodes (2006) differed from the other studies in several ways: (1) the encoding task was incidental, asking participants to rate the face age as young, middle-aged or old (Exp. 1); (2) an age rating was used as the basis for categorizing faces into same-age/other-age group at the retrieval phase, instead of using objective age ranges; (3) pose of target faces was changed between encoding and recognition phase; (4) retention interval was 48 hours in Experiment 2, and therefore much longer than in any other study included. The difference in encoding task and the much longer retention interval may have contributed to the extreme outcomes. Naveh-Benjamin et al. (2009, Exp. 1) instructed their participants to study 24 face-name pairs and tested them (1) with a Yes-No face recognition task, (2) a name recognition task, and (3) recognition of face-name pairs. It is not clear to us why this study produced a negative age effect despite the complexity of these multiple tasks.

After outliers were removed, homogeneity tests were no longer significant (ps > .05). The mean weighted effect size indicated higher hits for young participants than their older counterparts when viewing young faces, $g_u = 0.49 \ [0.22, 0.76], \ p < .001, \ VC = .07$, and to a lesser extent when viewing mixed-age faces, $g_u = 0.34 \ [0.15, 0.53], \ p < .001, \ VC = .02$. A marginally significant negative effect was found between the age groups for older faces, $g_u = -0.24 \ [-0.49, 0.00], \ p = .053, \ VC = .03$. While younger participants had higher hits than their older counterparts when viewing young and mixed-age faces, the effect reversed such that older participants tended to have higher hits when viewing older faces.

Three studies reported hits for middle-aged faces. Therefore, no separate meta-analysis was conducted. Anastasi and Rhodes (2006) observed higher hits for young compared to older adults (Exp. 1: $g_u = 1.05 \ [0.56, 1.55]$; Exp. 2: $g_u = 1.26 \ [0.81, 1.72]$), whereas Fulton
and Bartlett (1991) found no differences between age groups ($g_u = 0.10 [-0.37, 0.58]$ for identical, and $g_u = 0.06 [-0.42, 0.53]$ for changed faces).

**False alarms.**

Twenty-six hypothesis tests were identified to test the effects on false alarms (8 for young, 10 studies for mixed-age, and 8 for older faces; see Figure 2). Outlier analyses revealed one extreme case for young and for older faces, respectively (Mason, 1986; see Figure 2), and one extreme case for mixed-age faces (Naveh-Benjamin et al., 2009, Exp. 1; see Figure 2). In Mason's study (1986), the task to study not only faces but also their associated names within a very limited time frame, and the large number of faces used, may explain why older adults had many more false alarms than young adults. The study by Naveh-Benjamin et al. (2009, Exp. 1) was described above.

After excluding outliers homogeneity tests were no longer significant, $p > .05$. Young participants made fewer false alarms than older participants when recognizing young, $g_u = 0.89 [0.67, 1.12]$, $p < .001$, $VC = .02$, mixed-age, $g_u = 1.07 [0.86, 1.27]$, $p < .001$, and older faces, $g_u = 0.72 [0.53, 0.91]$, $p < .001$, $VC = .00$.

Finally, for the three studies using middle-aged faces older participants had significantly higher false alarms than their younger counterparts (Anastasi & Rhodes, 2006, Exp. 1: $g_u = 0.90 [0.41, 1.38]$; Anastasi & Rhodes, 2006, Exp. 2: $g_u = 1.12 [0.67, 1.57]$; Fulton & Bartlett, 1991: $g_u = 0.96 [0.46, 1.46]$).

In summary, regardless of target age, older adults were more likely to incorrectly choose a new face as previously seen than young adults (Figures 2 and 4).

**Recognition performance.**

For signal detection measures of recognition performance ($d'$) five studies were included for young, seven for mixed-age, and five for older faces (Figure 3). For young faces, Mason (1986) was identified as an outlier (see above) and removed from the analyses (see Figure 3). The analyses showed large effects in favor of young participants for young faces: $g_u = 1.06 [0.80, 1.32]$, $p < .001$, $VC = .00$, and mixed-age faces: $g_u = 0.98 [0.80, 1.17]$, $p < .001$, $VC = .05$, and a medium effect for old faces: $g_u = 0.40 [0.11, 0.70]$, $p < .001$, $VC = .05$. Parallel to the results for hits and false alarms, young adults outperformed their
older counterparts in face recognition performance, regardless of target age. However, the differences between the age groups were smaller for older faces than for young and mixed-age faces.

**Response bias.**

Ten studies reported enough information to calculate the mean effect size for response bias for combined young and mixed-age faces (see Figure 3). Outlier analyses revealed no outliers. The mean weighted effect size was positive, $g_u = 0.52$ [0.30, 0.75], $p < .001$, $VC = .07$, indicating that older participants used a more lax decision criterion than their younger counterparts. In other words, compared to young adults, older adults were more likely to choose a face as previously seen, regardless of whether the face was a target or a distractor.

**Mirror Effect in Face Recognition**

To further grasp the sensitivity and specificity of the age groups, hits (sensitivity) and false alarms (1-specificity) were plotted against each other (see Figures 5a and 5b). Effects for all face age groups were included (young, middle-aged, older and mixed-age faces). Savaskan et al. (2007) was excluded as an outlier as the hits were unusually low for both age groups ($M_y = .35$; $M_o = .32$; see Figure 1). Young adults showed a negative relationship between hits and false alarms, indicating that with an increase in hits false alarms decreased (Figure 5a), thus demonstrating the expected mirror-effect pattern. In contrast, the relationship was reversed for older adults. Higher hits were not necessarily associated with lower false alarms (Figure 5b). Instead there was either no or a positive relationship between hits and false alarms for older adults, indicating that they did not discriminate well between targets and distractors. Only for young-aged faces Figure 5b implies a mirror effect, whereas for middle-aged, mixed or older faces there appears to be a positive association which would indicate a lenient response bias.
Figure 5a and 5b. Scatterplots of the correlations between hits and false alarms of young (Panel A) and older adults (Panel B).
Moderator Analyses with Meta-regression

The meta-analyses reported above were carried out separately for hits, false alarms and signal detection measures of performance and bias. We also separated analyses for young, mixed-age, and old faces. The latter separation was necessary because all studies that used old faces also used young faces (to test for an own-age bias). Because studies with both younger and older faces used identical methodologies their effect sizes are stochastically dependent (see Gleser & Olkin, 2009), and hence should not be pooled with studies using different methodologies that yield independent effect sizes.2

To avoid statistical dependencies, we pooled studies with independent data for young faces and studies with mixed-age faces--none of the studies used both--to conduct moderator analyses.3

Most meta-analyses in the (eyewitness and face recognition) literature have used separate blocking analyses (as an analogue to ANOVAs; see Lipsey & Wilson, 2001) with different moderator variables as predictors. However, this approach has been criticized because it does not adequately take the intercorrelations of predictor variables into account. To demonstrate, Table 2 displays the intercorrelations of potential moderators (encoding time (in seconds), retention interval (in minutes), number of faces at encoding, number of faces at 

2 Because the correlations between the dependent variables were not reported we could not estimate effect sizes taking these dependencies into account. We also decided against robust variance estimation methods (Hedges, Tipton, & Johnson, 2010) as an alternative due to the relatively low number of studies and the ensuing lack of power (a minimum of 40 studies are needed for the robust variance estimation when dependent and nondependent effects are included in the model; Hedges et al., 2010).

3 For signal detection measures, the number of studies (k = 12 for d’ and k = 10 for c or beta) was too small (and heterogeneity too small after removal of outliers; see Figure 3) to warrant meta-regression analyses. For SDT performance, the mean effect size was $g_u = 1.01 [0.86, 1.16], Z = 13.16, p < .001, VC = .000, Q(10) = 5.92, p = .822$ (after removing Mason, 1986, as an outlier). For response bias, the mean effect size is reported in Figure 3.
test, change of view (no change, 1 change, 2 changes), as well as the weighted semi-partial and zero-order correlations between the effect sizes (hits, false alarms) and the predictor variables used in the meta-regressions. As is apparent from Table 1, many studies shared some features, for example, using a longer encoding time when a large number of faces was used. Thus, such similarities lead to confounded predictors. While some of the correlations were small, they nonetheless indicated that ignoring them would introduce confounds in blocking analyses (Pigott, 2012). However, the intercorrelations were not too high to create problems of multicollinearity.

As Table 2 shows, the number of faces at encoding were highly correlated with the number of faces at test, thus making the latter redundant as a predictor. Less obvious, and hence a potential confound in moderator analyses, are the medium to high correlations of encoding time with number of faces at test and with the number of changes at test. Presumably, researchers used longer encoding times to compensate when the task became more difficult.

Further support for the interdependencies of the predictor variables became apparent from the semi-partial correlations, which were calculated using ordinary least squares weighted regression analyses to partial out shared variance with other predictor variables. Compared to zero-order correlations, some of the semi-partial correlations differed considerably in strength and direction. For instance, although the correlation between hits and number of faces at encoding was large ($r = .49$), the semi-partial correlation that controlled for other predictors was zero ($sr_k = .02$). The difference was reversed for change of faces: While the small correlation ($r = -.08$) increased considerably for hits ($sr_k = -.24$), the medium to large correlation ($r = .40$) decreased to a small effect for false alarms after controlling for other variables ($sr_k = .14$).
# Table 2

Intercorrelations of Predictor Variables for all Studies (excluding Smith & Winograd, 1978) and Weighted Semi-partial (and Zero-order) Correlations for Hits and False Alarms

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Age of Older Age Group</th>
<th>Face Age</th>
<th>Encoding Time</th>
<th>Retention Interval</th>
<th># of Faces at Encoding</th>
<th># of Faces at Test</th>
<th># of Foils at Test</th>
<th>Change of Faces</th>
<th>Additional Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age of Old Age Group</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Face Age (Young vs. Mixed)</td>
<td><strong>-.43</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoding Time (s)</td>
<td>.04</td>
<td>-.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention Interval (min)</td>
<td>-.08</td>
<td>-.26</td>
<td>.16</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Faces at Encoding</td>
<td>.20</td>
<td>.21</td>
<td>.17</td>
<td>-.28</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Faces at Test</td>
<td>.13</td>
<td>-.22</td>
<td><strong>.45</strong></td>
<td>-.08</td>
<td><strong>.67</strong></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Foils at Test</td>
<td>.24</td>
<td>-.40</td>
<td>.34</td>
<td>.01</td>
<td>.36</td>
<td><strong>.87</strong></td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changes of Pose and/or Expression at Test</td>
<td>-.25</td>
<td>.35</td>
<td><strong>.62</strong></td>
<td>.09</td>
<td>.15</td>
<td>.26</td>
<td>.12</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Additional Task(s) besides Face Recognition</td>
<td><strong>.43</strong></td>
<td>.00</td>
<td>-.08</td>
<td>-.20</td>
<td>-.03</td>
<td>-.36</td>
<td>-.31</td>
<td>-.34</td>
<td>1.00</td>
</tr>
</tbody>
</table>

| Effect Size Hits                              | -.29                        | -.30     | .07           | **.49**            | -.02                   | --                 | --                | -.24             | -.24            |
|                                               | (.22)                       | (.24)    | (.09)         | (.58)             | (.49)                  | --                 | --                | (.08)            | (.39)           |
| Effect Size False Alarms                      | -.30                        | -.01     | .15           | .11               | .30                    | --                 | --                | .14              | .11             |
|                                               | (.25)                       | (.19)    | (.34)         | (.13)             | (.34)                  | --                 | --                | (.40)            | (.12)           |

*Note.* Significant (with df = 16) correlations > .40 are highlighted in bold face. Semi-partial correlations are reported only for predictor variables included in meta-regressions. *These two variables were removed from meta-regression to avoid collinearity and to reduce the number of predictors in the model.*
To avoid these confounds, we only used weighted meta-regression analyses which statistically control for mutual pairwise as well as multivariate dependencies in predictors. This was only possible for studies reporting recognition of young and mixed-age faces which we pooled to obtain a sufficient number of studies to conduct meta-regression analyses.\footnote{It is not quite clear what the minimum number of studies should be to conduct meta-regression analyses but there is serious concern about the lack of power of meta-regression analyses when \( k \) is small (Hedges & Pigott, 2004).} Predictor variables were entered as blocks in hierarchical meta-regressions, after the outliers noted above were removed.\footnote{Although including the outliers would result in significant results for some of the predictors we decided to exclude them so results would not be distorted by one or two outliers.}

Because we pooled studies using both young faces and mixed-age faces, in step one we included both the mean age of participants in the old age group and age of faces (young vs. mixed-age) as predictors (see Tables 3 and 4). If there is an own-age bias, effect sizes should be larger for recognition of young faces compared to those using mixed-age faces. In step 2 we entered methodological characteristics described in Table 1 into Tables 3 and 4 as predictors.

**Hits.**

A total of 16 effect sizes were available for the meta-regression analyses for hits (without the two outliers by Anastasi & Rhodes, 2006, Exp. 1 and 2). The mean weighted effect size was \( g_u = 0.31 \) [0.17, 0.45], \( Z = 4.31, p < .001, VC = .015, Q(15) = 18.34, p = .245 \).

Results of the mixed-model meta-regression for hits are presented in Table 3. At step 1, the effect sizes for hits were surprisingly larger when the mean age of the older participants was younger and for studies using young (vs. mixed-age) faces. That is, the higher the mean age of the older participants the smaller was the difference between them and younger adults. At the same time, young adults did not perform as well when they viewed mixed-age faces compared to young faces only.
### Table 3

**Hierarchical Weighted Mixed-Effect Meta-regression for Hits for Young and Mixed-Age Faces (k = 16)**

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>B</th>
<th>beta</th>
<th>p</th>
<th>R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Q(2) = 5.62,  p = .060</td>
<td></td>
<td></td>
<td>.306</td>
</tr>
<tr>
<td>Residual</td>
<td>Q(13) = 12.72,  p = .470</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Age of Old Age Participant Group</td>
<td>-05</td>
<td>-.66</td>
<td>.033</td>
<td></td>
</tr>
<tr>
<td>Face Age (Young vs. Mixed)</td>
<td>-.37</td>
<td>-.68</td>
<td>.029</td>
<td></td>
</tr>
<tr>
<td><strong>Step 2:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Q(7) = 14.90,  p = .037</td>
<td></td>
<td></td>
<td>.812</td>
</tr>
<tr>
<td>Residual</td>
<td>Q(8) = 3.44,  p = .904</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Age of Old Age Participant Group</td>
<td>-.05</td>
<td>-.60</td>
<td>.208</td>
<td></td>
</tr>
<tr>
<td>Face Age (Young vs. Mixed)</td>
<td>-.30</td>
<td>-.54</td>
<td>.203</td>
<td></td>
</tr>
<tr>
<td>Encoding Time (s)</td>
<td>.01</td>
<td>.13</td>
<td>.756</td>
<td></td>
</tr>
<tr>
<td># of Faces at Encoding</td>
<td>-.00</td>
<td>-.04</td>
<td>.917</td>
<td></td>
</tr>
<tr>
<td>Retention Interval</td>
<td>.03</td>
<td>.63</td>
<td>.038</td>
<td></td>
</tr>
<tr>
<td>Changes of Pose and/or Expression at Test</td>
<td>-.13</td>
<td>-.37</td>
<td>.314</td>
<td></td>
</tr>
<tr>
<td>Additional Task(s) besides Face Recognition</td>
<td>-.19</td>
<td>-.33</td>
<td>.301</td>
<td></td>
</tr>
<tr>
<td>Regression constant</td>
<td>4.16</td>
<td>.00</td>
<td>.137</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Q statistics shown are from mixed-effects meta-regression models and test whether the independent variables (1) mean age of the old-age participant group and face age, and (2) study characteristics account for an additional portion of the variance in effect sizes. Two outliers (Anastasi & Rhodes, 2006, Exp. 1 and 2) were removed from the model. Before removing the outliers: Step 2 Model Q(7) = 44.79,  p < .001; Residual Q(10) = 4.55,  p = .919; R^2 = .908. Significant results are highlighted in boldface.*
These effects were no longer significant at step 2 when other variables were entered as predictors and thus controlled for, indicating that other factors mitigated the effect of mean age of the older group and of face age. The only significant predictor was retention interval: With a slightly longer retention interval, effect sizes for hits were higher than when tests followed more or less immediately the encoding phase. Although $R^2$ was very high ($R^2 = 81.2\%$ explained variance, as opposed to $R^2 = 30.6\%$ at step 1), none of the other predictors was significant. This may be due to mutually controlling these factors (or to low power). Particularly in the case of low statistical power (many predictors with few studies) moderator effects are difficult to detect (Hedges & Pigott, 2004).

**False Alarms.**

After removing one outlier (Mason, 1986), the mean weighted effect size was $g_u = 0.95$ [0.78, 1.12], $Z = 11.17$, $p < .001$, $VC = .047$, $Q(16) = 26.37$, $p = .049$. None of the predictors (the same as used for hits above) at step 1 and at step 2 yielded significant contributions (see Table 4).
Table 4

Hierarchical Weighted Mixed-Effect Meta-regression for False Alarms for Young and Mixed-Age Faces (k = 17)

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>B</th>
<th>beta</th>
<th>p</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Q(2) = 1.80,  (p = .406)</td>
<td></td>
<td></td>
<td>.112</td>
</tr>
<tr>
<td>Residual</td>
<td>Q(14) = 14.36,  (p = .423)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Age of Old Age Participant Group</td>
<td>-.03</td>
<td>-.26</td>
<td>.336</td>
<td></td>
</tr>
<tr>
<td>Face Age (Young vs. Mixed)</td>
<td>.09</td>
<td>.13</td>
<td>.624</td>
<td></td>
</tr>
<tr>
<td>Step 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Q(7) = 6.02,  (p = .537)</td>
<td></td>
<td></td>
<td>.408</td>
</tr>
<tr>
<td>Residual</td>
<td>Q(9) = 8.74,  (p = .462)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Age of Old Age Participant Group</td>
<td>-.05</td>
<td>-.42</td>
<td>.247</td>
<td></td>
</tr>
<tr>
<td>Face Age (Young vs. Mixed)</td>
<td>.00</td>
<td>.01</td>
<td>.987</td>
<td></td>
</tr>
<tr>
<td>Encoding Time (s)</td>
<td>.03</td>
<td>.24</td>
<td>.517</td>
<td></td>
</tr>
<tr>
<td># of Faces at Encoding</td>
<td>.01</td>
<td>.33</td>
<td>.274</td>
<td></td>
</tr>
<tr>
<td>Retention Interval</td>
<td>.00</td>
<td>.11</td>
<td>.713</td>
<td></td>
</tr>
<tr>
<td>Changes of Pose and/or Expression at Test</td>
<td>.10</td>
<td>.21</td>
<td>.588</td>
<td></td>
</tr>
<tr>
<td>Additional Task(s) besides Face Recognition</td>
<td>.08</td>
<td>.11</td>
<td>.747</td>
<td></td>
</tr>
<tr>
<td>Regression constant</td>
<td>3.70</td>
<td>.00</td>
<td>.215</td>
<td></td>
</tr>
</tbody>
</table>

Note. \(Q\) statistics shown are from mixed-effects meta-regression models and test whether the independent variables (1) mean age of the old-age participant group and face age, and (2) study characteristics account for an additional portion of the variance in effect sizes. One outlier (Mason, 1986) was removed from the model. Before removing the outlier: Step 2 Model \(Q(7) = 5.45, p = .606\); Residual \(Q(10) = 10.38, p = .407; R^2 = .344\).
Publication Bias

Sensitivity analyses were conducted to test and adjust for publication bias with the nonparametric trim-and-fill method by plotting effect sizes against standard errors (Duval & Tweedie, 2000a, 2000b). Figures 6a to 6c present the funnel plot for (a) hits, (b) false alarms, and (c) recognition performance for young and mixed-age faces with the imputed studies (circled dots), if applicable. The dashed vertical line indicates the mean weighted effect without the imputed studies.

The funnel plot in Figure 6a showed a symmetric distribution of the effect sizes for hits. The trim-and-fill analyses confirmed the funnel plot analyses, namely absence of a publication bias for hits. The fail-safe $N$ revealed that $k = 74$ studies using the fixed-effects model or $k = 55$ studies using the random-effects model would be necessary to render the effects nonsignificant for hits.

By contrast, the funnel plot in Figure 6b showed an asymmetric tendency for false alarms and in Figure 6c for recognition performance, indicating smaller effect sizes missing at the left. The trim-and-fill method suggested the possibility of four missing studies on the left side of the funnel for false alarms and of three missing studies for recognition performance. The re-estimated weighted mean effect size for false alarms was $g_{u} = 0.85$ [0.73, 0.97], $Z = 14.06, p < .001, k = 21$, which only little deviates from the original value of $g_{u} = 0.94$ [0.81, 1.07]. The fail-safe $N$ indicated that $k = 909$ using the fixed-effects model or $k = 544$ using the random-effects model would be required to make the effect for false alarms nonsignificant. For recognition performance, the re-estimated weighted mean effect was $g_{u} = 0.93$ [0.79, 1.07], $Z = 13.35, p < .001, k = 14$, again only slightly smaller than the original value of $g_{u} = 1.01$ [0.86, 1.16]. The fail-safe $N$ for SDT performance was $k = 472$ using both fixed- and random-effects models.
Figure 6a, 6b, and 6c. Funnel plots and trim-and-fill analyses for (a) hits, (b) false alarms, and (c) recognition performance without outliers. Solid circles are included studies, open circles are imputed studies. The solid vertical line denotes the re-estimated mean effect size, while the dashed vertical line indicates the original mean weighted effect size.
Discussion

Different theoretical approaches discussing age-related changes in the sensory system (Lott et al., 2005; Schieber, 2006), in brain structure (Coleman & O’Hanlon, 2008), and in cognitive processes and processing speed (Craik, 1986; Salthouse, 1996) have commonly been invoked to account for deteriorating memory performance at old age. These factors can also be assumed to affect face recognition performance with increasing age. Although comprehensive narrative reviews of face recognition studies of young versus older people exist (Bartlett, 2014a; Bartlett & Memon, 2007), little was known about the size of these effects.

The present meta-analysis examined age effects on recognition memory for faces containing 79 comparisons between young and older adults on different dependent measures (hits, false alarms and signal detection measures d’ and c). Included studies focused solely on the recognition of unfamiliar faces without a complex event. As hypothesized, young people outperformed their older counterparts, showing small to medium effects for hits and large effects for false alarms and SDT performance measures (small: $g_U = 0.2$; medium: $g_U = 0.5$; large: $g_U \geq 0.8$; according to Cohen’s, 1988, recommendations). The effects observed here are even more noteworthy in light of the large-scaled meta-analysis of 128 studies by Shapiro and Penrod (1986) who found much smaller effects for most of the variables investigated in the literature. Our effects were even larger than those for the own-race bias in Meissner and Brigham’s (2001) meta-analysis who found a mean weighted $g_U = 0.24$ for hits and $g_U = 0.39$ for false alarms. Thus, the effects observed in our meta-analysis can be considered quite robust, in particular because the effects were largely homogeneous once outliers had been removed. Notably, in all studies reviewed older participants had been prescreened to exclude any persons which showed more general cognitive or memory deficits. Thus, our estimates are likely underestimates compared to more representative populations. Our results also show that age effects are not only present in recall and cued recall but also in recognition memory paradigms.

Although the effects on hits were not quite as high as those on false alarms, the general pattern followed a mirror effect as postulated by signal detection theory (Glanzer & Adams,
1985) that had also been observed in the literature on the cross-race effect. However, the mirror effect found here was not quite symmetric, and, interestingly, only present with young participants who showed the expected negative association between hits and false alarms across studies (Figure 5a). With older participants, the pattern was non-existent or tended to be even positive, depending on the age of the to-be-recognized faces (Figure 5b). However, these conclusions are based on visual inspection of the graphs only and could not be tested with inferential statistics because the original studies did not report the necessary intercorrelations.

Age effects differed as a function of face age: Differences between young and older adults were larger for young and mixed-age faces than for older faces, and reversed for hits for older faces, indicating the presence of an own-age bias. While younger adults outperformed older adults when recognizing young faces, older adults were better at correctly recognizing older faces. Analogous own-group advantages have been well documented regarding the own-race bias (Meissner & Brigham, 2001; Sporer, 2001). Also analogous to the own-race bias, older people displayed a more liberal response criterion than younger people, with a medium effect size. In summary, compared to older people young people were better in discriminating new faces from previously seen faces and less likely to accept new faces as previously seen.

We further explored potential effects of moderator variables. Due to small to medium size intercorrelations between predictor variables, we employed meta-regression analyses that control for confounding effects (rather than using a series of blocking analyses as is frequently found in the meta-analytic literature; see Pigott, 2012).

Surprisingly, the above-mentioned age effect became weaker when mean age of the older participants was taken into consideration: The higher the age of older people the less was the difference in hits between young and older people. This result contradicts previous research on biological markers of cognitive functioning, which found a decline in memory that is more rapid in older age (Anstey, 2011). One possible explanation are possible confounds inherent in the experimental designs across studies: Mean age of older people was negatively correlated with face age (young vs. mixed-age) \( (r = -.43) \). Thus, it seemed that the
higher the age of older people the more likely were the primary studies to report results for young and older faces separately. Yet when mean age of older people was controlled for, face age remained a significant predictor, indicating the presence of an own-age bias (Rhodes & Anastasi, 2012).

Perhaps surprisingly, retention interval was the only moderator to reach statistical significance in the meta-regression with hits when controlling for the remaining variables (face age, mean age of older participants, encoding time, number of faces at encoding, changing between encoding and test, and additional tasks at test). Longer retention intervals were associated with larger differences between the age groups for hits but not for false alarms. Considering the different mirror effect patterns of young and older people shown in Figures 5a and 5b, these differences between hits and false alarms indicate that the recognition process and forgetting over time may operate differently in young versus older age. However, practically all retention intervals were minimal in length (less than one hour), clearly indicating the need to study longer retention intervals of days, weeks, or months in future studies. To test whether memory decline over time is linear or curvilinear, future studies should measure recognition at least at three points in time (Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Rubin & Wenzel, 1996), perhaps also with within-participants designs.

None of the other moderators were associated with hits, and none of the moderators were associated with false alarms. These findings, however, do not necessarily rule out moderator effects, considering the relatively low number of studies available and the ensuing low power of our moderator analyses (Hedges & Pigott, 2004).

**Recognition Performance**

In addition to separate face recognition measures, the current meta-analysis considered sensitivity of recognition performance, namely the SDT discrimination index $d'$, which simultaneously takes both hits and false alarms into account (Macmillan & Creelman, 2005). From an applied perspective, the SDT performance measure is of greater importance than the individual recognition measures hits and false alarms. In other words, making conclusions about someone's face recognition performance based solely on hits would be misleading. If
false alarms increase along with hits, this would only indicate a change in response bias but not in recognition performance.

The age groups differed by one standard deviation in $d'$ demonstrating a clear advantage of young age. Discrimination between old and new faces, therefore, seemed to be significantly impaired with age. The reduced processing speed hypothesis (Salthouse, 1996) proposes that mental speed declines with age. Cognitive processes that require working memory may therefore decline with age, which results in poorer performance and inaccurate memory of older people (Maylor, 2005). Salthouse (1996) showed that reduced processing speed in age led to decreased ability to accomplish tasks or retrieve relevant information. Similarly, Wolff et al. (2012) demonstrated that face-processing speed declined with age, with the strongest decline after 60. As already indicated in the introduction, meta-analyses cannot provide answers regarding the types of face information extracted at encoding and used in decision processes and the role working memory plays. We do encourage future researchers to test participants regarding their (visual) working memory capacity and to assess response latencies to gain further insight into these processes.

**Response Bias**

The response criterion measures a participant's tendency to denote a face as "seen" or "not seen" before. As a consequence of setting a more liberal criterion, an individual might correctly recognize most of the faces presented because of his or her proneness to choose. In contrast, another individual might operate more cautiously or conservatively at the face recognition task, leading not only to fewer false alarms, but also fewer hits. In fact, the tendency towards a conservative or liberal response criterion and a tendency to false memory has been linked to personality traits (Kantner & Lindsay, 2013), and to age-related structural changes in the frontal brain lobe (Huh et al., 2006).

Overall, the tendency for a liberal response seemed to be a robust effect in older age, which we could only test for young and middle-aged faces. Older adults may rely more on a generalized sense of familiarity (Hancock et al., 2000; Vokey & Read, 1992) when deciding whether or not they have seen a face before. Future studies should routinely investigate response bias with faces of different ages. More frequent choosing rates were also observed
with older compared to young people in studies using a lineup paradigm in the meta-analysis by Kocab et al. (2017), and in a large scale field study by Martschuk et al. (2018).\

**Own-age Bias**

Social-psychological approaches to face recognition stress the importance of in-group and out-group categorization of the target face for later recognition (Hugenberg et al., 2007; Sporer, 2001). Possible reasons for the own-age bias are cognitive processes (He et al., 2011), personal motivation (Wiese et al., 2012) and (recent) contact frequency with out-group members (Rhodes & Anastasi, 2012; Wiese et al., 2013). Moreover, the own-age bias has been shown to be a robust effect across age groups from children to old age (Rhodes & Anastasi, 2012).

The present meta-analysis supported these assumptions, showing decreased differences between age groups with increasing face age, that is, a Participant Age by Face Age interaction (see Figure 4). Participant age differences decreased with increasing target age and became negative for older age faces for hits. Specifically, older participants showed an increase, whereas their younger counterparts showed a slight decrease in hits with older targets. Similarly, differences in false alarms and in recognition performance $d'$ between the age groups decreased when participants were shown older faces compared to young faces.

With mixed-age faces, young participants chose significantly fewer unknown faces as previously seen than older participants.

**Practical Implications**

The findings demonstrated that older individuals are not only less likely to correctly recognize a person they previously saw but also have a greater probability of incorrectly

---

6 But see also Colloff, Wade, Wixted, and Maylor (2017), who used an online study in which the target had a highly distinctive mark [scar or tattoo on his cheek] and in order to create fair lineups in three out of four conditionns the lineups were digitally manipulated to prevent the distinctive target from standing out, and found older adults did not just set an overall more lenient response criterion than young adults. With unfair lineups, older adults were highly overconfident and confidence was not a good predictor of accuracy.
indicating a face as previously seen compared to younger people. This effect was robust across the different measures and was stronger for false alarms than for hits, with little heterogeneity. Although the studies included in this meta-analysis all used old-age groups with people older than 60, declines in recognition performance may occur even earlier (e.g., Germine et al., 2011).

These findings have general and forensic implications. In daily life, social interactions with new acquaintances (as opposed to people we know) might be complicated with increased age by the fact that memory for faces is impaired, and even more when the interaction is with younger people. Of course, this meta-analysis did not test whether these effects hold in live interactions (including conversations) but the present findings indicate that in older age people tend to "recognize" faces of strangers whom they think they have encountered before. As some of the individual studies reviewed show, these effects may be even stronger when the persons are encountered in a different context (or have changed in appearance--although our moderator analyses did not confirm this). Training programs for older adults "to improve your memory" could be designed to counteract age-related deficits, emphasizing the need for "individuation" of faces.

In criminal investigations, the face recognition paradigm reviewed here resembles more a "showup" or "street identification" than a traditional lineup with a suspect and several distractors. Considering the importance of person identifications in criminal investigations, our results should be considered by triers of fact (including police, attorneys, jurors and judges) who evaluate the accuracy of witness identification decisions. When evaluating identification decisions, these evaluators have to consider all possible factors that may affect the accuracy of an identification, especially if witnesses are older. The sensory system, eyesight and cognitive functioning are important, and all these factors should be evaluated individually. For instance, not every young person would be a more reliable observer than an older person when his or her visual ability is limited. Likewise, not every older person would be an unreliable witness just because they have reached a particular age. If possible, medical and psychiatric history should be taken into consideration, as these factors impair memory and recognition.
An important factor is a participant’s age in relation to the age of the person-to-be-recognized. The own-age bias has been repeatedly demonstrated, both in the present meta-analysis on studies with the face recognition paradigm and in those using eyewitness identification paradigms (although evidence in the latter studies is not so clear; see Kocab et al., 2017; Rhodes & Anastasi, 2012). Accordingly, older individuals should be more reliable in recognizing older faces than younger faces, and young individuals should be less reliable in recognizing faces as the target face increases in age. As most crimes are committed by younger people, and the proportion of older people in the population, as well as their likelihood of victimization, increases, a possible own-age bias may be an important factor to consider.

Further, older individuals displayed a more liberal response criterion than younger people. Although attitudinal variables may not be related to identification decisions (at least not in the context of own-race bias research: Meissner & Brigham, 2001) older participants have been shown to have more positive attitudes towards the police (Gallagher, Maguire, Mastrofski, & Reisig, 2001) and to be more prosecution prone in a large scale representative study using the Juror Bias Scale (Goodman-Delahunty, 2018). Hence we hypothesize that older people may also be more likely to choose someone in an identification/face recognition task even if their memory trace is weak. There is at least some prior research that indicated that personality and age were directly related to setting of response criteria (Huh et al., 2006; Kantner & Lindsay, 2013); for a general review of individual differences and facial recognition, see Hosch, 1994).

However, we have been unable to locate research that directly compared young vs. older people on attitudinal variables like Belief in a Just World (Jones & Brimbal, 2017; Lipkus, 1991), Right-wing Authoritarianism (Altemeyer, 1998; Zakrisson, 2005), or Legal Authoritarianism (Narby, Cutler, & Moran, 1993) in relation to the setting of a more liberal response criterion. While most researchers in the past have favored cognitive explanations for age effects in recognition performance we encourage future researchers to also investigate social and personality variables to test whether they may explain the setting of response criteria in face recognition and lineup studies.
In identification or showup procedures, possible countermeasures against a response bias (e.g., fair lineup instructions: Memon, Hope, Bartlett, & Bull, 2002) should be further empirically tested and optimized for different age groups (children, younger and older adults). Also, explicitly allowing a "don't know" option could be explored.

**Limitations of the Present Meta-analysis**

One of the most frequently voiced objections to meta-analysis is the "file-drawer problem" (Rosenthal, 1979), that is, the absence of unpublished data. If unpublished studies are missing, the overall effect size is likely to be overestimated, based on the assumption that unpublished studies are likely to have been negative or not significant.

To counteract the overestimation of effect sizes different methods were applied in the current meta-analysis. First, known researchers in the field and authors of identified primary studies were contacted to obtain unpublished data. Second, the literature search included not only English but also German, French and Russian studies to overcome a language bias. Unfortunately, the search for unpublished and non-English studies did not add any additional studies to the current meta-analysis. Even though some authors responded, these studies did not meet the inclusion criteria. In addition, missing data in some of the published studies led to exclusion of these data even though we had contacted the researchers of the primary studies for clarification. Third, the nonparametric trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) was conducted to assess the possible impact of publication bias and to calculate an "adjusted" effect size. Due to the small number of studies that reported the response criterion the impact of publication bias using the trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) could not be assessed for this dependent measure.

The trim-and-fill analysis (Duval & Tweedie, 2000a, 2000b) was conducted when combining young and mixed-age faces for hits, false alarms, and recognition performance, respectively. Results indicated the potential absence of four smaller or negative effect sizes for false alarms, and the potential absence of three smaller effect sizes for recognition performance; thus a possible overestimate of the weighted mean effect size may have been present for these measures. The difference between the results obtained in the meta-analysis
and the "adjusted" effect sizes was negligible. There was no indication of a publication bias for hits using this method.

A further limitation of the meta-analysis associated with missing data was nonreporting of inter-effect-size covariances between dependent measures in a study. This made it impossible to include all possible effects in the meta-analysis without violating the assumption of independence of effect sizes (Cheung & Chan, 2014; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). The random variance estimation method was not conducted, because (1) the method assumes equal correlations between measures—but our data indicated that this was not the case; and (2) at least 40 studies are required to yield stable results (Hedges et al., 2010). Multi-level meta-analyses that take within-cluster dependencies into account could not be applied, because some of the studies reported only data for one effect size, and therefore did not meet the criteria for Level 2 clustering (for a critical review, see Cheung & Chan, 2014). The sample-wise approach would lead to biased results due to the assumed heterogeneity of the effect sizes when all measures would have been entered into the same model (Cheung & Chan, 2004; Hunter & Schmidt, 2004).

Finally, our meta-regression findings need to be interpreted with caution for different reasons. While the meta-regression results at step 1 seem to be robust, the absence of significant effects in the full model are likely a function of low power (Hedges & Pigott, 2004) and the fact that some of the intercorrelated predictor variables share mutual variance. To gain further insight into these interdependencies we calculated the semi-partial correlations between predictor variables and hits and false alarms.7 To increase the validity of our meta-regression analyses (a) we excluded outliers from analyses to prevent distortions of effects; (b) we specified predictors based on theory a priori and not based on "ex post facto observed correlations" (Schmidt, 2017, p. 470); (c) we reported unstandardized and standardized regression weights, and (d) intercorrelations and semi-partial correlations of predictor variables.

7 These analyses were recommended by an anonymous reviewer.
Methodological Implications and Future Directions

As discussed above, one of the limitations of the meta-analysis was the presence of missing values. This was particularly present for the signal detection measures for performance ($d'$) and response bias ($c$). Less than two-thirds of the studies reported a discrimination index and only a handful of studies reported data on response criterion. Hence, we encourage future researchers not only to report the individual face recognition measures but also the combined statistics showing participants' SDT performance and response criterion measures, as these measures can only be calculated with original data.

Although many studies tested face recognition for faces changed between study and test (e.g., pose, expression), such natural variations should be routinely employed (even in neurological studies). To control for possible confounds of distinctiveness, attractiveness, familiarity, perceived age and other face attributes that covary with face age, large pools of faces should be used. Faces could be pre-rated on these attributes and assigned either randomly or counter-balanced in complex experimental designs. One of the advantages of the face recognition paradigm over eyewitness identification paradigms is the maximization of stimulus sampling (Wells & Windschitl, 1999).

Furthermore, future studies should manipulate and extend the delay between encoding and retrieval; more than half of the studies tested recognition performance immediately, most studies within 25 minutes, and only one study after 48 hours (Anastasi & Rhodes, 2006, Exp. 1). Hence, the implications of the present findings are limited to an immediate or short delay and memory and retrieval processes after longer delays remain unknown. Over extended delays, the role of verbal and visual (rehearsal) processes and of context reinstatement could also be tested with young and older adults.

Finally, future research should include confidence as an additional measure to the "Old"/"New" recognition paradigm or as response alternatives in rating experiments (i.e., high to low certainty that the stimulus was "old"/"new"; Macmillan & Creelman, 2005). Such designs will enable to plot empirical Receiver Operating Curves, which are more fine-grained measures of sensitivity. Confidence has been routinely measured in eyewitness research, and findings revealed that high confidence of choosers is a reliable indicator of identification
Memory for faces in old age: A meta-analysis

accuracy (Sporer, Penrod, Read, & Cutler, 1995; Wixted & Wells, 2017). Confidence was also more strongly related to accuracy with younger than with older adults in an identification field study (Martschuk, Sporer, & Sauerland, 2018; but see Colloff et al., 2017, who found older adults could adjust their confidence ratings appropriately to achieve a similar likelihood of being correct at each level of confidence as younger adults in fair but not in unfair lineups). Future studies need to address the confidence-accuracy issue with different targets (e.g., race), procedures, retention intervals and other possible moderators related to old age.

Conclusions

The present meta-analysis demonstrated that face recognition performance was weaker for older than for young adults. Further, older adults tended to choose more often, demonstrating a more liberal response bias which led to more false alarms, whereas young adults tended to use a more conservative response criterion. This is in line with the meta-analytical findings for eyewitness identification studies (Kocab et al., 2017). This conclusion was also supported when hits and false alarms were plotted against each other for each age group separately: While the mirror effect was present for young participants, showing a negative association between hits and false alarms, there was no visible association between hits and false alarms for older people, perhaps even a positive relationship.

Further, our findings showed the importance of face age. In line with previous research (e.g., Meissner & Brigham, 2001; Rhodes & Anastasi, 2012) the present study demonstrated an own-group advantage which was strongest for hits (i.e., older adults made more hits than their younger counterparts when recognizing old faces), and to a lesser extent for recognition performance d’ (i.e., the age differences were smaller for mixed-age and older faces compared to young faces). Finally, longer retention intervals (as opposed to immediate retrieval) tended to be associated with greater differences between the age groups but future research should test this with longer delay intervals.
References

References marked with an asterisk indicate studies included in the meta-analysis.


