OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Age and Semantic Inhibition Measured by the Hayling Task: A Meta-Analysis

Teresa Cervera-Crespo[1,*], Julio González-Alvarez[2]

[1]*University of Valencia, Spain*
[2]*University Jaume I, Castellón, Spain*

*Corresponding author at: Facultad de Psicología, Departamento de Psicología Básica, University of Valencia, Blasco Ibanez, 21, 46010 Valencia, Spain. Tel: +34 963864823; fax: +34 983864822.
E-mail address: Teresa.Cervera@uv.es (T. Cervera-Crespo)

Accepted 3 October 2016

Abstract

Objective: Cognitive aging is commonly associated with a decrease in executive functioning (EF). A specific component of EF, semantic inhibition, is addressed in the present study, which presents a meta-analytic review of the literature that has evaluated the performance on the Hayling Sentence Completion test in young and older groups of individuals in order to assess the magnitude of the age effect.
Method: A systematic search involving Web of Science, PsyINFO, PsychARTICLE, and MedLine databases and Google Scholar was performed. A total of 11 studies were included in this meta-analysis, encompassing a total of 887 participants; 440 young and 447 older adults. The effect sizes for group differences on four measures of the Hayling test, latency responses and error scores on the Automatic and Inhibition sections of the test were calculated using the Comprehensive Meta-Analysis software package
Results: The results revealed large age effects for response latencies in both the Automatic (Hedges' $g = 0.81$) and Inhibitory conditions (Hedges' $g = 0.98$), though the latter two effect sizes did not differ from each other. In contrast, analysis of errors revealed a significant difference between the small effect seen in the Automatic condition (Hedges' $g = 0.13$) relative to the moderate effect seen in the Inhibition condition (Hedges' $g = 0.55$).
Conclusions: These results may be important for a better understanding of the inhibitory functioning in elderly individuals, although they should be interpreted with caution because of the limited number of studies in the literature to date.

Keywords: Hayling task; Meta-analysis; Aging

Inhibition can be defined as a process that suppresses irrelevant information that interferes with carrying out the task in progress (the task the individual is currently doing). This ability is necessary in order to overcome prepotent, well-learned, automated behaviors in novel situations that require a different response (Shallice, 1988). Inhibition is a central component of executive control, and its impairment impedes accomplishing complex cognitive tasks. The executive functioning (EF) model by Norman and Shallice (1986) proposes that inhibitory processes act under the control of the Supervisory Attentional System (SAS), which generates, selects, and initiates appropriate cognitive schemas in response to novel and challenging situations. The main characteristic of this model is the distinction it makes between automatic and controlled processes. The automatic activation of certain behaviors would not be sufficient in a situation that requires a novel sequence of actions, errors correction, or overcoming strong habitual responses.

Cognitive aging is characterized by changes in inhibitory capacities (Hasher, Stoltzfus, Zacks & Rypma, 1991). Compared to young adults, older adults frequently present less efficient inhibitory mechanisms, and it is often suggested that this decrease accounts for larger part of the cognitive changes associated with age. This decrease in inhibitory capacities with aging has been linked to brain changes, more precisely to changes in the frontal lobes. Clinical studies (Burgess & Shallice, 1996; Shallice, 1988; Stuss & Alexander, 2000) and neuroimaging studies (Hornberger, Geng & Hodges, 2011; West, 1996) confirm the involvement of the frontal lobes in inhibitory control. However, there is also evidence indicating that other brain

doi:10.1093/arclin/acw088

structures, in addition to the frontal lobes, contribute to the performance on tests that measure inhibition (Andrés & Van der Linden, 2001).

Some research suggests that inhibition deficits are also an important feature of the cognitive deterioration in the early stages of dementia (Amineva, Philips, Della Sala & Henry, 2004; Perry, Watson & Hodges, 2000) and could have a significant impact on everyday activities in older age (Lawton & Brody, 1969). Thus, the measurement of inhibitory cognitive capacities plays a central role in the early diagnosis of mild cognitive impairment and Alzheimer's disease (Nathan, Wilkinson, Stammers & Low, 2001), and it is a valuable indicator (along with other cognitive tests) in the prognosis of frontotemporal dementia (Hornberger, Piguet, Kipps & Hodges, 2008).

Specifically, failures in semantic inhibition are frequently measured with the Hayling Sentence Completion test (Burgess & Shallice, 1997). The tasks used in this test are based on the SAS theoretical framework which proposes that two cognitive processes control our actions and thoughts. One of them is responsible for the routine everyday tasks, and the other operates in non-routine more challenging tasks. In the Hayling test, the individual is presented with 30 sentences where the last word is missing. In the first 15 sentences (Automatic section), the word is strongly cued by the preceding context, and the individual has to complete the sentence by producing the missing word (e.g., in the sentence "This man has travelled everywhere around the…", the correct response would be "world"). In the next 15 sentences (Inhibition section), a word that makes no sense in the sentence context or is unrelated to the target word must be given by the individual. For instance, for the sentence "Most sharks attack very close to the…", the participant could give the word "table". Thus, in order to complete the sentence, the individual has to inhibit the automatic response *coast* and generate a new unrelated word.

In both the Initiation and Inhibition sections, the number of errors and the response time (RT) are registered. Thus, the test provides four raw measures: Mean RT in the Automatic and Inhibition sections and number of errors in the Automatic and Inhibition sections. The errors in both sections are weighted according to the following procedure (Burgess & Shallice, 1997): In the Automatic section, when the individual respond with the correct completion word he/she receives an error score of 0 (no error), while responses somewhat connected to the target word receive an error score of 1, and responses unrelated to the sentence receive an error score of 3. There is no answer that can obtain a 2-point score. In the Inhibition section, the system for weighing the errors is the opposite. Thus, a weighted error score can be calculated by adding the error points. Additionally, in line with Burgess and Shallice (1997), an "overall scaled score" can be constructed from the addition of scaled scores derived from RTs in the Automatic and Inhibition sections separately and errors made in the Inhibition section.

This test was initially designed by Burgess and Shallice (1997) to assess executive dysfunction in patients with frontal lobe lesions. These patients showed longer response latencies and more errors. More recently, the test's sensitivity to normal aging has also been evaluated in different studies (reviewed in the present meta-analysis). The Hayling test has frequently been used in the aging population to measure inhibition control because it has many advantages: it is easy to give, it takes a short time to complete, and visual impairment or motor difficulties are not an impediment for the individual. This test, which was designed for English-speaking individuals (Burgess & Shallice, 1997), has been adapted to other languages such as French (Andrés & Van der Linden, 2000), Spanish (Abusamra, Miranda & Ferreres, 2007), Chinese (Chan, Shum, Toulopoulou & Chen, 2008), and Brazilian-Portuguese (Fonseca, Oliveira, Gindri, Zimmermann & Reppold, 2010).

Bielak, Mansueti, Strauss and Dixon (2006) presented normative data for the Hayling test for healthy middle-aged and older English-speaking participants distributed in seven different age groups from 53 to 90 years old. The results of their study showed that advancing age was associated with poorer performance on the four measures of the test: RTs and number of errors in the Initiation and Inhibition sections. The results from the study by Abusamra and colleagues (2007) with people from 30 to over 65 years old indicated age effects, but only for RTs in the Automatic section. Other studies (Frias, Dixon & Srauss, 2006, 2009) that compared the performance of young–old (YO) to old–old (OO) participants on the test using the "overall scaled score" found lower scores in the OO than in the YO group. Moreover, the studies by Yeung, Fischer and Dixon (2009), and Wang and Su (2013) reported significant differences in the number of errors in the Inhibition section of the test. However, other studies (Lin, Chan, Zheng, Yang & Wang, 2007) found age differences only in the number of errors in the Automatic section.

In addition, some studies have compared older adults to young adults. In most of them, the older adults frequently produced more errors and longer response latencies than the young group, but an absolute consensus does not exist. For instance, most of the studies found an impaired performance in older adults on the error scores in the Inhibition section, but not in the Automatic section (Andrés & Van der Linden, 2000; Bastin & Van der Linden, 2003; Borella, Ludwig, Fagot & de Ribaupierre, 2011a; Collette, Germain, Hogge & Van der Linden, 2009; Collette, Grandjean, Lorant & Bastin, 2014; Wang & Su, 2013). However, the study by Belànger and Belleville (2009) reported significant differences between young and older groups in the number of errors in both the Automatic and Inhibition sections, while Tournier, Posta and Mathey (2014) found that older adults made fewer errors than younger adults.

Regarding the response latencies, Andrés and Van der Linden (2000), Belànger and Belleville (2009), and Belleville, Rouleau and Van der Linden (2006) found significant age differences in the RTs in the Inhibition section, but not in the Automatic section. On the other hand, Tournier and colleagues (2014) found significant differences between young and older groups in both Sections, whereas, Borella and colleagues (2011a) observed age differences, but only in those participants (both young and older groups) whose working memory (WM) scores were high (but not in the groups with lower WM scores).

Thus, it seems that there is no clear agreement about the exact effects of age on each of the different measures of the Hayling task. Methodological differences in the study design or different procedures used to administer the test could explain this apparent lack of agreement. For instance, the study by Borella and colleagues (2011a), whose objective was to explore the influence of WM capacity on the performance on the Hayling task, used an extreme group design, that is, participants whose scores on a WM task were either very high or very low (while the majority of studies used participants whose WM scores were in the middle range).

Regarding the procedure adopted to administer the test and register the responses, there are also some differences. While some studies adopted the classic paradigm by Burgess and Shallice (1997) (Bastin & Van der Linden, 2003; Belleville et al., 2006; Collette et al., 2009, 2014; Morrone, Declercq, Novella & Beshe., 2010; Stites, Federmeier & Stine-Morrow, 2013; Wang & Su, 2013), other studies used adapted versions of the original test. For example, in the original procedure, the sentences in the Automatic section were presented first, followed by the sentences in the Inhibition section (blocked format). All the sentences were presented orally to the individual, and the responses were registered manually. The RTs were also measured manually with a stop-watch. Other studies (Belànger & Belleville, 2009; Borella et al., 2011a; Tournier et al., 2014) used a computerized procedure in where the sentences were visually and/or orally presented on a computer, which also registered the RT and the errors. Moreover, these three studies used a switching format (or "unblocked"), in other words, all the sentences in the Automatic and Inhibition sections were presented in a randomized order (so that the participant has to continually switch between the instruction to respond with the semantically-related word or the unrelated word), instead of the classic "blocked" format, where the participants were first presented with the sentences in the Automatic section, and then those in the Inhibition section. In addition, the study by Tournier and colleagues (2014) used a different procedure to obtain the responses; instead of asking the participants to generate the required word, the participants were presented with two words and had to choose the required one.

Given the theoretical and clinical importance of a correct knowledge about the age-related decrease in inhibitory functioning, the present study aimed to review and synthesize findings from published studies that evaluated the performance on the Hayling task in young and older participants. The objective was to shed more light on the exact effects of normal aging on the performance on each of the dependent measures of this test, and to interpret the results of each study in the context of all the other studies. The use of large samples in this research area is not common; therefore, a meta-analytic study is a good way to approach this issue. By combining the results of all these studies through meta-analysis, the magnitude of the age effect can be better clarified. To date, no meta-analytic studies have been conducted with this purpose in mind.

In the present study, the overall (summarized) estimated effect sizes were calculated for each of the four measures of the Hayling test (RTs and error scores for the Initiation and Inhibition sections) from published studies that compared young (control group) and older participants. Their magnitudes provided a quantitative measure of the differences between young and older adults on the different dependent measures of the test.

In addition, in this study, no attempt was made to summarize the results of studies comparing different age decades because the few existing published studies that report data about this test in different decades or age groups (Abusamra et al., 2007; Bielak et al., 2006; Oliveira, Pedron, Gonçalves-Gurgel, Tozzi-Reppold & Fonseca et al., 2012) did not use equivalent age ranges in each group. In the same way, no attempt was made to summarize the results comparing YO to OO because the few existing published studies reported different outcomes on the Hayling test. For instance, Lin and colleagues (2007) reported the error scores in the Automatic and Inhibition sections, while Wang and Su (2013) and Yeung and colleagues (2009) reported the error scores only in the Inhibition section. On the other hand, the studies by Frias and colleagues (2006, 2009) reported the "overall scaled score".

Materials and Methods

Search for and Selection for Primary Studies

A systematic search involving Web of Science, PsyINFO, PsychARTICLE, and MedLine databases and the Internet (e.g., Google Scholar) was performed using the combination of the following terms as search parameters: "Hayling sentence

completion test" OR, "Hayling test", OR "Hayling task", AND "older adults", OR "aging". No limits were applied regarding publication dates. Additionally, a manual search was conducted for articles cited on the reference lists of the initial pool of selected articles and in the indexes of the journals that publish most of the papers in the field.

The following inclusion criteria were used to select studies:

(a) The study had to be published in a peer-reviewed journal.
(b) The study had to include a control group of young participants and one or more groups of older participants over 60 years of age with no cognitive, neurological, and psychiatric disorders, drug or alcohol addiction, and/or sensory impairment. The study could include additional groups of older participants, such as a group with some type of dementia, but the data corresponding to these groups were not included in the present analysis.
(c) The study had to report at least one of the four measures from the Hayling test.
(d) The study had to report means and standard deviations (*SD*s) for at least one of the Hayling measures, or other statistics convertible to effect size Cohen's *d* (and Hedges' *g*), such as *t*-tests or univariate *F*-tests. In addition, the way the scores were calculated had to be clearly explained in the article.

The study selection was carried out by the authors, who independently screened search results for initial eligibility based on the title, abstract, and full-text reading of all the potentially eligible studies. In order to establish the reliability of the article inclusion, inter-rater agreement was calculated by using Cohen's Kappa, with 89% agreement, resulting in 92 studies preselected from the electronic search supplemented by one study obtained by scanning reference lists of previous reviews, resulting in a total of 93 articles. After excluding duplicate articles, 73 articles were included in the initial pool. Based on the inclusion criteria, this initial pool of published articles was then reduced to 11 articles. The degree of inter-reviewer agreement was 91%. Fig. 1 shows the flowchart for the published studies included and excluded in the current meta-analysis.

It is worth explaining in more detail that five of the preselected studies had to be excluded because they did not meet the (d) criteria: These five articles reported some measure of the Hayling test (from combining some of the four raw measures), from which we could not derive the raw measures. For instance, Amiri and colleagues (2014) reported a "final score"; Bailey and Henry (2008) reported the "overall scaled score" described by Burgess and Shallice (1997); Borella, Carretti and Beni (2008) calculated "correct scores in the Automatic section–Inhibition section", but did not report them separately; Borella, Delaloye, Lecerf, Renaud and Ribaupierre (2009) calculated an "interference index", and the study by Borella, Ghisletta and de Ribaupierre (2011b) did not offer information about the way the performance on the Hayling test was scored.

Among the selected studies in the present meta-analysis, two of them included (along with the young and healthy older groups of participants) one or more groups of participants with mild cognitive impairment and/or Alzheimer's disease (Belànger & Belleville, 2009; Belleville et al., 2006). In these cases, the data corresponding to these two groups of participants with dementia were not included in the meta-analysis. In the same way, another study (Wang & Su, 2013) reported data from two groups of older participants: YO and OO (as well as the data from the control group of young participants), but the data from the OO group was not included in the present meta-analysis because the mean age of this group was higher than the one corresponding to the rest of the studies in the selected pool.

In addition, four of the studies reported two or more sets of data: (i) the study by Borella and colleagues (2011a) reported two sets of data, one corresponding to participants with high WM (both young and older groups), and the other for participants with low WM; (ii) the study by Tournier and colleagues (2014) reported two sets of data corresponding to two formats of the Hayling test: blocked versus unblocked designs; and finally, (iii) the study by Collette and colleagues (2014) presented two sets of data corresponding to participants (both young and older) with "standard" versus "strong" episodic memory (EM). Although these studies present differences in their participant characteristics and methodology, they were included in the present meta-analysis, and later an analysis of moderator variables was performed in order to study their possible impact on the effect size. In all, the present meta-analysis included 11 studies (14 sets of data).

Coding the Measures of the Hayling Test

The pool of selected studies was examined in order to extract relevant data to perform the meta-analysis. The authors coded the different dependent variables of the Hayling test independently. Agreement between coders was 97%. Any disagreement between raters was resolved by discussion. Four dependent measures from the Hayling test were coded: (a) RT in the Automatic section, (b) RT in the Inhibition section, (c) error scores in the Automatic section, and (d) error scores in the Inhibition section. Shorter RTs and fewer errors in both sections indicated better performance, that is, better semantic

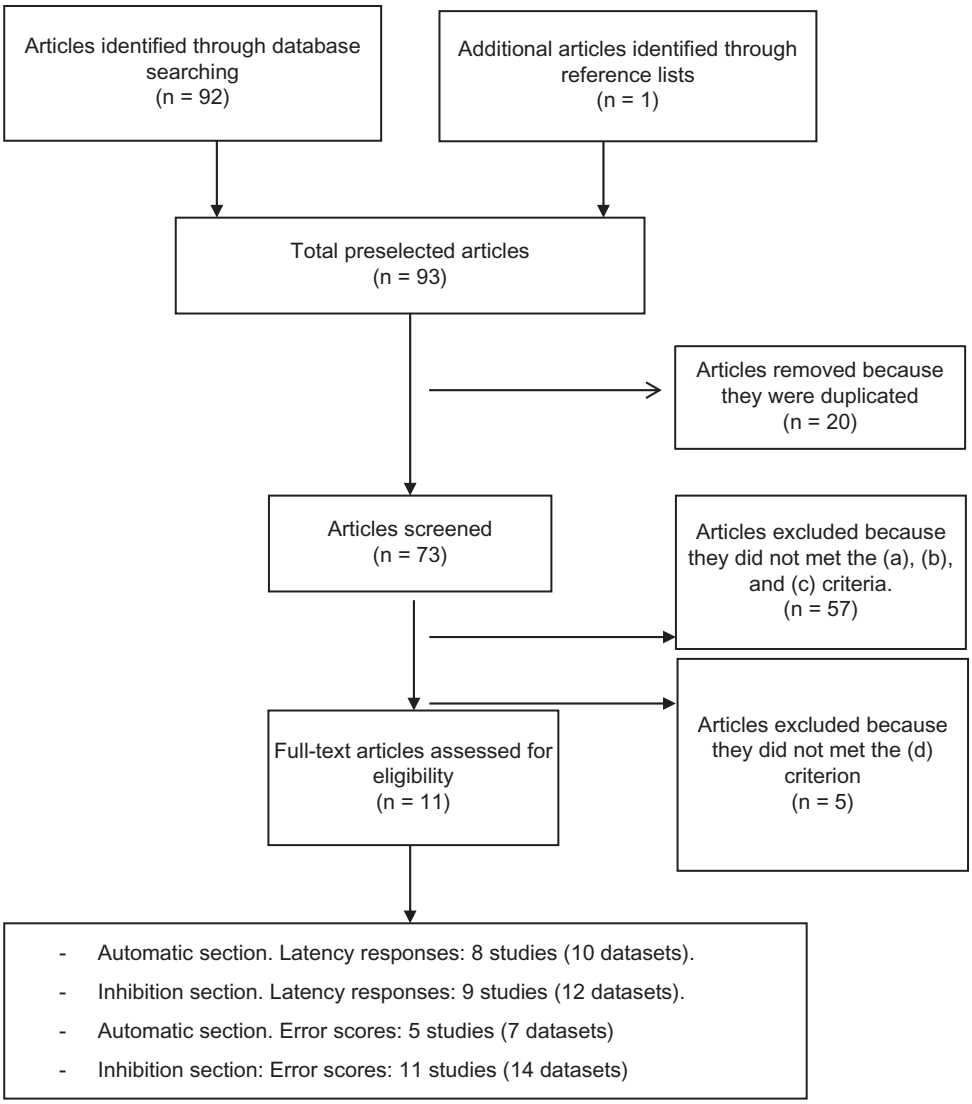

Fig. 1. Flowchart for the inclusion of the studies in the present meta-analysis.

inhibitory capacity. Thus, higher scores (worse performance) were typically observed in older participants, compared to younger participants, in the literature.

All the studies included in the present meta-analysis reported at least one of these four measures, which were entered as means and *SD*s. It should be noted that one of the studies (Tournier et al., 2014) reported the standard errors, which were transformed into *SD*s. In some cases, the authors were contacted to request these statistics. Additionally, some variables that were considered relevant for the study were coded, such as the mean age (and *SD*), years of formal education, and vocabulary in the young and older groups. The participants' language and the procedure followed to administer the test were also coded (Table 1).

Study Quality Assessment

The methodological quality of the studies included was assessed by means of an 11-item quality checklist derived from STROBE (Strengthening the Reporting of Observational Studies in Epidemiology, www.strobe-statement.org) adapted to meet the objectives of the present meta-analysis. These criteria were that the study had to present (i) clear objectives and hypotheses; (ii) sufficient characteristics of study participants from both samples regarding number of participants; (iii) regarding gender distribution; (iv) and demographic characteristics; (v) eligibility criteria for the elderly participants as well as

Table 1. Characteristics of studies and data sets (numbers in parenthesis are standard deviations)

Study	Data set	No. of participants		Mean age		Years of education		Language	Vocabulary Mill-Hill		Hayling format	Study quality
		Young	Older	Young	Older	Young	Older		Young	Older		
Andrés and Van der Linden (2000)		46	48	22.8 (2.8)	65 (3.9)	14.5 (1.3)	15.1 (2.2)	French	35.3 (3.6)	38.1 (2.7)	Classic	11
Bastin and Van der Linden (2003)		64	64	21.73 (2.36)	64.36 (3.03)	14.47 (1.96)	14.17 (1.81)	French	NR	NR	Classic	9
Belleville et al. (2006)		12	12	22 NR	72.7 NR	13.8 (1.7)	11 (1.9)	French	34.4 (5.2)	35.3 (6.7)	Classic	9.5
Belànger and Belleville (2009)		20	16	NR	NR	NR	NR	French	10.8[a] (2.5)	11.3[a] (1.9)	Adapted[b]	9
Collette et al. (2009)		40	40	24.15	67	NR	NR	French	36.6 (3.6)	37.52 (5.7)	Adapted[c]	10
Borella et al. (2011a)	Low working memory	48	52	23.04 (3.5)	70.62 (5.02)	16 (0.0)	12.87 (3.01)	French	36.64 (2.74)	38.12 (2.57)	Adapted[d]	10
	High working memory	52	52	21.81 (2.41)	68.73 (5.8)	16 (0.0)	14.38 (3.71)	French	37.03 (2.73)	40.31 (2.57)	Adapted[d]	
Morrone et al. (2010)		30	30	24.5 (2.81)	70 (3.32)	16.7 (1.3)	13.73 (1.5)	French	27.77 (2.39)	29.10 (2.44)	Classic	9
Stites et al. (2013)		18	18	24.4 (18–26)	65.5 (62–83)	NR	NR	English	NR	NR	Classic	7
Wang and Su (2013)		32	42	26.5 (1.65)	69.21 (2.5)	14.47 (1.92)	14.05 (1.46)	Chinese	NR	NR	Classic	11
Tournier et al. (2014)	Blocked	30	31	20.68 (1.89)	69.61 (7.35)	13.60 (1.0)	12.84 (2.73)	English	35.04 (3.0)	40.2 (2.04)	Adapted[e]	11
	Unblocked[f]										Adapted[g]	
Collette et al. (2014)	Standard episodic memory	24	24	22.1 (2.15)	68.9 (2.8)	14.4 (1.8)	14.0 (1.0)	French	22.75 (4.30)	27.29 (3.87)	Classic	10
	Strong episodic memory	24	24	22.1 (2.0)	68.2 (3.1)	14.5 (1.7)	13.9 (2.0)	French	24.67 (3.64)	28.83 (2.91)		

Note: NR, not reported.
[a]Measured by subtest in Weshler Adult Intelligence Scale.
[b]Switching format, computerized, visually, and orally presented.
[c]Switching format, computerized, visually presented.
[d]The same sentences are used in the Automatic and Inhibition sections.
[e]Two-response choice, computerized.
[f]The same participants.
[g]Two-response choice, computerized, switching (or unblocked) format.

information about cognitive assessment; (vi) a clear definition of all the variables; (vii) a clear explanation of assessment methods for each measure of the Hayling test; (viii) a description of any efforts to address potential confounders; (ix) a report on summary measures of the dependent variable; (x) a description of statistical methods; and (xi) an interpretation of the results. It is worth noting that, in the present meta-analysis, each study had different objectives. In some of them, the measurement of the performance on the Hayling test was not their primary objective. Thus, they may not provide exhaustive information about the Hayling task compared to other studies in the pool.

The presence criterion received 1 point, and absence received 0 points. The total maximum score was 11 points. The two authors evaluated each study independently, and the Pearson correlation coefficient between their scores gave a value of $r = 0.98$. Table 1 shows the quality values for each study (mean of the two judges). Most of the studies received between nine and eleven points (between 77% and 100% of the total score). The quality scores were not used to weigh the studies or to discard them from the meta-analysis.

Meta-Analysis Procedure

Overall, the 11 studies included in this meta-analysis encompassed a total of 887 participants; 440 young and 447 older adults. The mean age of the older participants ranged from 64 to 72 years old, and from 20 to 26 years old for the group of young participants (Table 1). All of the older individuals were reported to have normal cognitive status and good health.

The present meta-analysis was carried out using the Comprehensive Meta-analysis software (version 2.2.064) (Borenstein, Hedges, Higgins & Rothstein, 2005). Four separate analyses were conducted, one for each of the four measures on the Hayling test. As not all of the studies in the pool reported outcomes for the four measures on the Hayling test, each meta-analysis included a different number of studies. Effect sizes for each study in each pool were calculated using Cohen's d, with corrections for small sample sizes using the so-called Hedges' g (Hedges & Olkin, 1985). Cohen's d (and Hedges' g) is a standardized metric obtained by calculating the difference between two means divided by the pooled SD for the two groups. In the present meta-analysis, the older participants made up the target group (1st sample), and the younger participants were consider the control group (2nd sample). Cohen's d was calculated using the formula:

$$d = \frac{M_1 - M_2}{SD_{pooled}},$$

where M_1 and M_2 are the means for the 1st and 2nd samples, and SD_{pooled} is the pooled SD for the samples. SD_{pooled} is properly calculated by

$$SD^*_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}.$$

Thus, this metric allows the comparison of measures that employ different scales. Then, Cohen's d values were converted to Hedges' g using the formula by Hedges and Olkin (1985):

$$g \cong d\left(1 - \frac{3}{4(n_1 + n_2) - 9}\right).$$

The effect size value of Hedges' g reflects the magnitude of the age effect on the performance on the Hayling test for the dependent variable measured. Values of g are interpreted according to Cohen's guidelines (Cohen, 1992), where effect sizes of 0.2 are considered small, 0.5 are considered medium, and 0.8 are considered large. In the present study, a positive value of the effect size indicates an increase with age in the studied variable (either RTs or error scores). In other words, the older individuals have longer RTs and more errors than the younger ones, which is interpreted as poorer performance. By contrast, a negative value of the effect size indicates a decrease with age in the RTs and errors, which is interpreted as better performance by the older group, compared to the young group. The closer an effect size is to 0, the smaller the difference is between the age groups. Furthermore, a 95% confidence interval (CI) was calculated for each study effect size, to establish whether it was statistically significant.

If the number of available studies to compute the effect size was larger than three, an overall (summarized) effect size was calculated using a random-effects model. This model was selected because it was considered more representative of the data, in the present meta-analysis, than the fixed-effects model. The selection of the model was based on our expectations about

whether or not the studies shared a common effect size. As the studies in the pool had differences in the way the Hayling task was administered, it was unlikely that all of them would be functionally equivalent. Thus, we cannot a priori assume homogeneity in the magnitudes of the effect sizes of the studies (Borenstein, Hedges, Higgins & Rothstein, 2009; Hedges, 1994). The overall effect size represents a weighted average of the effect sizes of the individual studies in the pool. The weight assigned to each study to compute the overall effect size was the inverse variance (Hedges & Olkin, 1985). In the random effects the weight includes two variance components: within-study variance and between-study variance.

Forest-plots were also obtained to examine the distribution of the effect sizes of individual studies in the pool, and to estimate the impact of possible outliers. In order to examine the variations in effect sizes across studies, an overall Q-statistic was used to test homogeneity, that is, to test whether the effect sizes of the studies could be assumed to have come from a single population (Hedges & Olkin, 1985). The Q-statistic has a chi-square distribution with $k-1$ degrees of freedom, where k is the number of studies. A significant Q (if Q has a value of $p < 0.05$) indicates heterogeneity of the studies' effect sizes. In addition, the I^2 was used (Higgins & Thompson, 2002) as an indicator of the extent of the heterogeneity, and as a complement to the Q-test. I^2 values of 25%, 50%, and 75% suggest small, moderate, and large heterogeneity, respectively.

If there was heterogeneity in the effect sizes, an analysis of potential moderator variables was performed. The categorical variables were (i) format used to administer the Hayling test (classic vs. adapted), and (ii) language of the participants (English, French, Chinese). The quantitative variables were (i) mean age, (ii) years of education, and (iii) vocabulary of the elderly sample (Table 1). Note that the gender distribution of the participants was not used as a moderator because there was an insufficient number of studies (<10) reporting this information. Table 1 shows that one study employed participants with high versus low WM (Borella et al., 2011a), and another study employed participants with standard versus strong EM (Collette et al., 2014). These two variables (WM and EM) were not used as moderator variables because there was only one study in the pool with each of these characteristics.

Q5 For the categorical moderators the random-effects ANOVA-analog of the Q-statistic was used to determine whether the moderator contributed to the effect size variability. Both between ($Q_{between}$) and within (Q_{within}) group variances were calculated to determine whether they were statistically significant. A significant $Q_{between}$ indicates that there is significant variability between the levels of the categorical variable moderator, whereas a significant Q_{within} indicates that there is still significant variability within each effect size that is not being explained by the categorical moderator. For the quantitative moderators, simple weighted meta-regression was performed for each moderator to examine how the moderator is related to the variation in effect sizes across studies. (Hedges, 1994). A Z-test of the unstandardized regression coefficient (b_j) was used to determine the statistical significance. The comparison between the effect sizes for the Automatic and Inhibition conditions for the RTs and for the Error scores of the Hayling task was also performed using the ANOVA-analog of the Q-statistic (random effects model).

Publication bias was tested by means of constructing the funnel plots of the effect sizes. In this graphic, the effect sizes are plotted on an x-axis, whereas the standard error is on the y-axis. When there is publication bias, the plot tends to be asymmetrical. Although funnel plots are helpful for exploratory purposes, they are limited because of the subjectivity involved in evaluating the shapes of the distributions. Thus, Egger's test for publication bias (Egger, Smith, Schneider & Minder, 1977) was calculated, as well as Duval and Tweedie's (2000) trim and fill procedure for imputation of lost studies from the funnel. In addition Rosenthal's fail-safe N (Rosenthal, 1995) was estimated, which provides an estimate of the number of unpublished studies with non-significant findings that would be needed to reduce a significant mean effect size across studies to non-significance (Lipsey & Wilson, 2001). A larger N indicates that greater confidence can be placed in the significance of the current findings. All of these computations were performed using was Comprehensive Meta-Analysis software. The MOOSE (Meta-analysis of Observational studies in Epidemiology, www.strobe-statement.org) guidelines were followed throughout the entire meta-analytic process.

Results

Mean RT in the Automatic Section

Mean effect sizes and heterogeneity. Eight (10 data sets) out of 12 Hayling task studies reported the means and *SD*s of the RTs in the Automatic section for older and young (control) groups. Table 2 shows the Hedges' g effect sizes, the 95% CIs, variance, and relative weight for each study, as well as the summarized effect size (and 95% CI). Because higher RTs on the Hayling test indicate worse performance, a positive effect size (Hedges' g) indicates a disadvantage for the older group, whereas negative effect sizes show a disadvantage for the younger sample. As Table 2 shows, the overall effect size of age on semantic inhibition, based on 10 effect sizes in the pool, was 0.81 (95% CI, 0.35–1.27, $Z = 3.44$, $p < 0.01$), indicating that

older participants had significantly longer latencies than the younger ones on the Automatic section of the Hayling test, and that this effect is high according to Cohen's criterion (Cohen, 1992). The Q-statistic was calculated and indicated significant heterogeneity in the studies' effect sizes ($Q(9) = 78.12$, $p < 0.01$), and the I^2 index ($I^2 = 88.48\%$) showed that 88.48% of the variability in effect sizes can be attributed to something other than sampling error.

An inspection of the forest-plot (Fig. 2) showed that the study by Borella and colleagues (2011a) (set 2: individuals with high WM) presented a small negative effect size of −0.03, and the study by Tournier and colleagues (2014) (the two data sets) showed the largest effect sizes in the pool (see also forest-plot). After excluding these two studies from the calculation of the overall effect size, a Hedges' g value of 0.52 (95% CI, 0.33–0.71, $Z = 5.30$, $p < 00.1$) was obtained. The $Q(6)$ value of 6.81 ($p = 0.34$) and the I^2 value of 0% indicated no heterogeneity.

As there was significant heterogeneity, an examination of moderator variables was performed. The results showed that the test format was significant ($Q_{between}(1) = 4.02$, $p < 0.05$), although within-group variance was also significant ($Q_{within}(8) = 74.10$, $p < 0.01$), indicating that the model may not be well specified, and that other moderator variables can exist. The largest effect size was for the study by Tournier and colleagues (2014) (the two data sets). It is worth noting that this study was the only one that used an adapted format with two-response choices instead of an open response. According to the authors, this format was employed with the objective to the involvement of other non-inhibitory processes, such as anticipating the answer in the Inhibition section prior to hearing the sentence (especially when the unblocked format was administered), by providing two alternative words from which the participant had to choose.

On the other hand, the moderator variable language of participants did not show a significant relationship with Hedges' g ($Q_{between}(1) = 0.81$, $p = 0.36$). For the quantitative moderators, a weighted simple meta-regression analysis was conducted for

Table 2. Results of the meta-analysis for the response latencies in the Automatic section of the Hayling task

Study name	Hedges's g	Variance	95% confidence intervals	Relative weight
Andrés and Van der Linden (2000)	0.31	0.04	−0.09–0.71	10.72
Bastin and Van der Linden (2003)	0.47	0.03	0.12–0.82	10.92
Belleville et al. (2006)	1.16	0.18	0.32–1.99	8.42
Belànger and Belleville (2009)	0.75	0.12	0.09–1.42	9.38
Collette et al. (2009)	0.48	0.05	0.04–0.92	10.54
Borellla et al. (2011a) (low working memory)	0.76	0.04	0.35–1.16	10.71
Borella et al. (2011a) (high working memory)	−0.19	0.04	−0.57–0.19	10.80
Stites et al. (2013)	0.10	0.11	−0.54–0.73	9.53
Tournier et al. (2014) (blocked)	1.93	0.09	1.33–2.53	9.73
Tournier et al. (2014) (unblocked)	2.69	0.12	2.00–3.38	9.26
Overall	0.81	0.06	0.35–1.27	

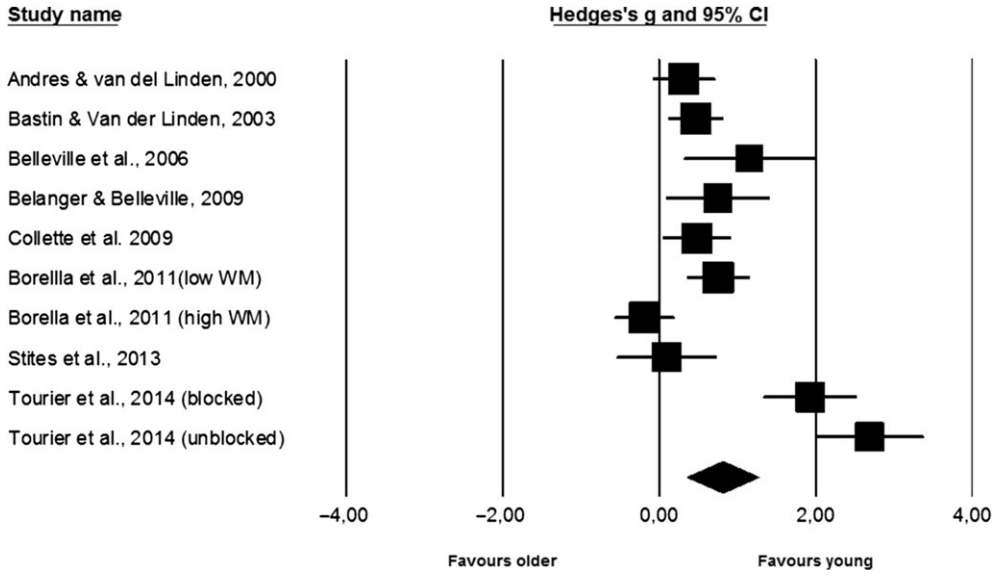

Fig. 2. Forest-plot of latency responses in the Automatic section of the Hayling task. WM, working memory.

each moderator, with Hedges' g as a dependent variable. As expected, there was a statistically significant positive relationship between Hedges' g and the "mean age" of the elderly samples ($b_j = 0.10$, $Z = 3.43$, $p < 0.01$), and a negative statistically significant relationship with education ($b_j = -0.46$, $Z = -4.76$, $p < 0.01$). This result indicated that, as the mean age of the elderly sample increases, the effect size increases. On the other hand, as years of education increased, the effect sizes decreased. The variable vocabulary was not significant ($b_j = 0.01$, $Z = 1.05$, $p = 0.28$).

Finally, it is worth noting that the study by Borella and colleagues (2011a) (Data set 1) employed participants who were selected based on their high scores on WM (measured by an adaptation of the "reading span task", Delaloye, Ludwig, Borella, Chicherio & de Ribaupierre, 2008). It seems that, under these circumstances, older participants performed better than younger ones. Although it is possible that higher than normal WM could moderate the differences between young and older adults in their performance on the Hayling test, this meta-analysis cannot explore this possibility because the study by Borella and colleagues (2011a) is the only one in the pool with these subject characteristics.

Finally, the presence of a publication bias in our results was examined by using the funnel plot (Fig. 3), which was moderately symmetrical. Egger's test for publication bias showed a value for the intercept (B_0) of 8.22, with $t(4) = 2.48$, $p < 0.05$, indicating significance. On the other hand, the trim and fill method imputed three studies, and the fail-safe N was 197, indicating that we would need to locate and include 197 null studies for the effect to be nullified.

Mean RT in the Inhibition section. Nine studies (12 data sets) compared the RTs of older and young individuals in the Inhibition section. It should be noted that: (i) the study by Borella and colleagues (2011a) is the only one in the pool where the sentences presented in the Inhibition section were the same ones presented in the Automatic section (instead of using a different set of sentences in each section of the test, as in the classic procedure by Burgess and Shallice, 1997), but with the instruction to provide an incongruent word to complete the sentence; and (ii) the article by Collette and colleagues (2009) reported the RTs for the Inhibition – Automatic/Inhibition + Automatic formula, from which we derived the RT for the Inhibition section. Table 3 shows the effect sizes, the 95% CIs for each study, and the summarized effect size for the total pool of studies. As in the case of RTs in the Automatic section, higher RTs in the Inhibition section of the Hayling taskt indicate worse performance, and a positive effect size (Hedges' g) indicates a disadvantage for the older participants, while negative effect sizes show a disadvantage for the young participants. The significant overall effect size Hedges' g for eight studies was 0.98 (95% CI, 0.52–1.44, $Z = 4.15$, $p < 0.01$), indicating that older participants showed worse performance on the latencies in the Inhibition section of the test, and that this effect was high (Cohen, 1992). The Q value obtained indicated significant heterogeneity in the effect sizes of the studies ($Q(11) = 106.06$, $p < 0.001$), and the I^2 index ($I^2 = 89.62\%$) indicated the amount of this heterogeneity; that is, 89.62% of the variability in effect sizes can be attributed to something other than sampling error. As in the case of the previous section (RTs in the Automatic section), an inspection of the forest-plot (Fig. 4) indicated that, the studies by Borella and colleagues (2011a) (data set corresponding to participants with high WM) and Tournier and colleagues (2014) showed the lowest and largest effect sizes, respectively, in the context of the other studies. After excluding them from the analysis, the overall effect size was moderate (Hedges' $g = 0.63$; 95% CI = 0.38–0.83, $Z = 4.99$,

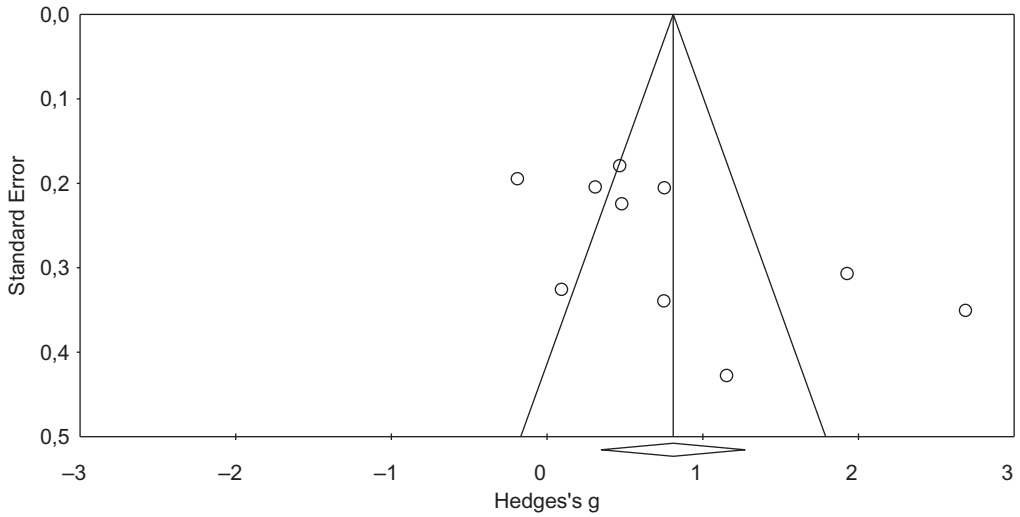

Fig. 3. Funnel plot for latency responses in the Automatic section of the Hayling task. The solid vertical line represents the weighted average effect size.

Table 3. Results of the meta-analysis for the latency responses in the Inhibition section of the Hayling task

Study name	Hedges's *g*	Variance	95% confidence intervals	Relative weight
Andrés and Van der Linden (2000)	0.75	0.05	0.32–1.17	8.87
Bastin and Van der Linden (2003)	0.22	0.03	−0.13–0.56	9.10
Belleville et al. (2006)	1.56	0.21	0.67–2.45	7.08
Belànger and Belleville (2009)	1.46	0.14	0.73–2.19	7.76
Collette et al. (2009)	0.46	0.05	0.02–0.90	8.83
Borella et al. (2011a) (low working memory)	0.42	0.04	0.03–0.81	8.97
Borella et al. (2011a) (high working memory)	−0.03	0.04	−0.41–0.35	9.00
Stites et al. (2013)	0.60	0.11	−0.05–1.25	8.05
Collette et al. (2014) (standard episodic memory)	0.43	0.08	−0.14–0.99	8.40
Collette et al. (2014) (strong episodic memory)	0.70	0.09	0.12–1.27	8.36
Tournier et al. (2014) (blocked)	2.72	0.12	2.03–3.41	7.90
Tournier et al. (2014) (unblocked)	3.09	0.14	2.35–3.82	7.70
Overall	0.98	0.06	0.52–1.44	

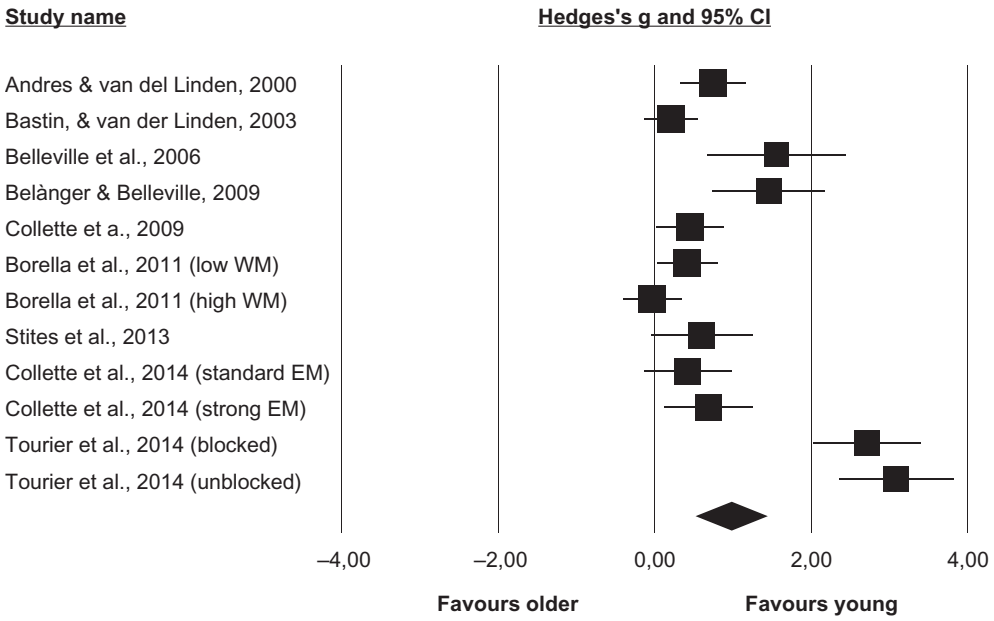

Fig. 4. Forest-plot of latency responses in the Inhibition section of the Hayling task. WM, working memory; EM, episodic memory.

$p = 0.04$), and I^2 with a value of 51.28% indicated that the amount of heterogeneity decreased, but was still significant according to the Q-statistic, with a value of 23.96 ($df = 9$) ($p = 0.02$) ($I^2 = 66.61\%$).

As the homogeneity analysis was significant, an analysis of moderators was performed, revealing that the language of participants was significant ($Q_{between}(1) = 52.20$, $p < 0.01$), respectively, although the within-group variances were also significant ($Q_{within}(10) = 53.86$, $p < 0.01$). The test format was not significant ($Q_{between}(1) = 2.69$, $p > 0.05$). For the quantitative moderators, the meta-regression analysis showed a statistically significant positive relationship between Hedges' g and the mean age of the elderly samples ($b_j = 0.10$, $Z = 3.30$, $p < 0.01$), and a negative statistically significant relationship with education ($b_j = -0.40$, $Z = -3.47$, $p < 0.01$). The variable vocabulary did not show a significant relationship with the effect sizes ($b_j = 0.30$, $Z = 1.48$, $p = 0.14$).

As previously described, the publication bias was analyzed by inspecting the funnel plot (Fig. 5) and calculating Eggers's test, which was significant ($B_0 = 8.11$, with $t(10) = 3.79$, $p < 0.01$), suggesting publication bias. On the other hand, the trim and fill algorithm imputed zero studies and the fail-safe N was 356.

Error scores in the Automatic section. Table 4 shows the overall effect size (and CI) calculated across five studies (seven data sets) that measured the error scores on the Hayling test in young and older groups, as well as the effect sizes of each

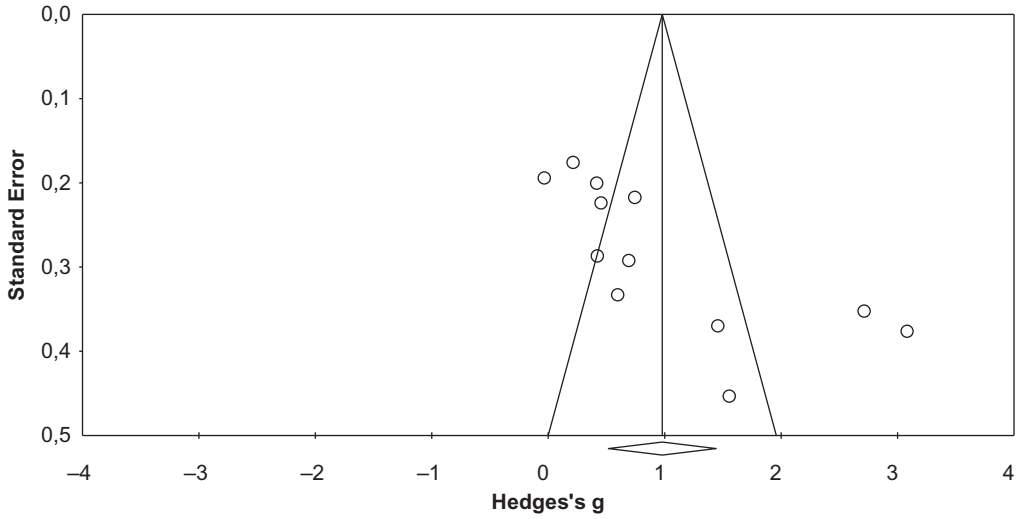

Fig. 5. Funnel plot for latency responses in the Inhibition section of the Hayling task. The solid vertical line represents the weighted average effect size.

Table 4. Results of the meta-analysis for the error scores in the Automatic section of the Hayling task

Study name	Hedges's *g*	Variance	95% confidence intervals	Relative weight
Belleville et al. (2006)	0.25	0.11	−0.40–0.89	12.13
Belànger and Belleville (2009)	0.50	0.11	−0.15–1.15	11.99
Borella et al. (2011a) (low working memory)	0.00	0.04	−0.39–0.39	17.22
Borella et al. (2011a) (high working memory)	0.00	0.04	−0.38–0.38	17.38
Stites et al. (2013)	−0.83	0.12	−1.49–0.16	11.76
Tournier et al. (2014) (unblocked)	0.83	0.07	0.31–1.35	14.55
Tournier et al. (2014) (blocked)	0.08	0.06	−0.42–0.58	14.98
Overall	0.13	0.03	−0.20–0.46	

Q6 study. It should be noted that: (i) in the study by Belànger and Belleville (2009), the scoring scale was reversed (compared to what was used in the original task by Burgess and Shallice (1997), and therefore, the scores represent correct responses; and (ii) the study by Stites and colleagues (2013) reported the number of correct responses. In both cases, these scores were converted to errors to compute the effect size. The summarized effect size for the seven sets of data was 0.13 (95% CI, −0.20–0.46, $Z = 0.75$, $p = 0.46$), indicating that the older participants performed worse, that is, they produced more errors than the younger participants, although this effect is small and non-significant. On the other hand, the Q-statistic indicated significant heterogeneity in the effects sizes of the studies in the pool ($Q(6) = 17.23$, $p < 0.05$, and I^2 was 65.16%).

The forest-plot (Fig. 6) shows that the study by Stites and colleagues (2013) presented a higher (and negative) effect size than the rest of the studies, with a Hedges' g value of −0.83. It is should be noted that, in this study, the Hayling task was administered to participants only for cognitive screening purposes, and little information is provided in the article about the procedure used to administer the task. Thus, we cannot speculate about the reason for this result. After excluding this study from the pool, a larger (but still not significant) overall effect size of 0.24 (95% CI, −0.02–0.5, $Z = 1.72$, $p < 0.05$) was obtained and all the studies in the pool became homogeneous ($Q(5) = 8.83$, $p = 0.12$, $I^2 = 43.39$). No attempt was made to study the possible effect of moderators because of the insufficient number of studies in this pool, following the recommendation by Borenstein and colleagues (2009), or the publication bias (Higgins & Altman, 2008).

Error scores in the Inhibition section. A total of 11 studies (14 data sets) were computed in the analysis, where the measured variable was the error score in the Inhibition section. In the case of the studies by Belànger and Belleville (2009), Borella and colleagues (2011a), and Stites and colleagues (2013), correct scores were converted to error scores as in the previous section. Table 5 shows the effect sizes (and CI) for each study in the pool and the summarized effect size. As this table reveals, a moderate and significant overall effect size of 0.55 (95% CI, 0.36–0.74, $Z = 5.62$, $p < 0.01$) was obtained, indicating that the older

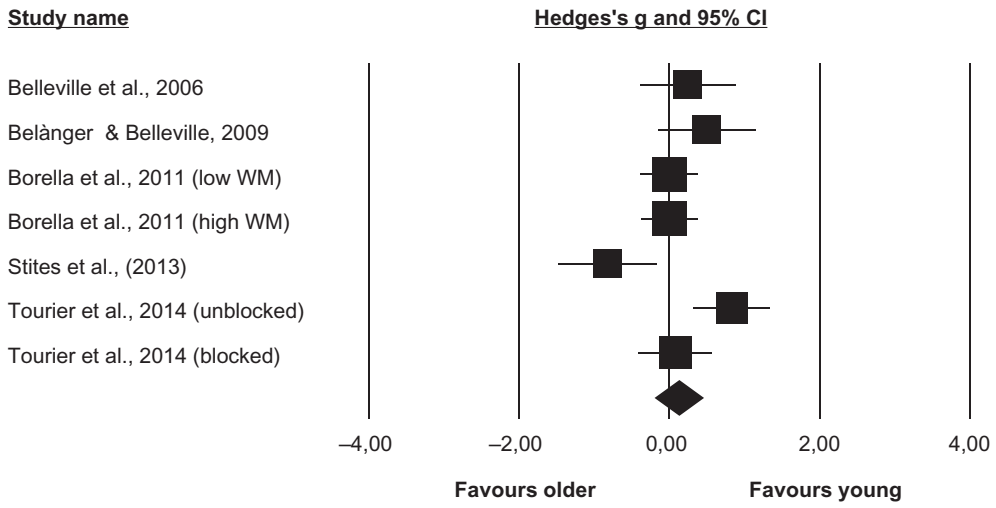

Fig. 6. Forest-plot of error scores in the Automatic section of the Hayling task. WM, working memory.

Table 5. Results of the meta-analysis for the error scores in the Inhibition section of the Hayling task

Study name	Hedges's *g*	Variance	95% confidence intervals	Relative weight
Andrés and Van der Linden (2000)	0.65	0.04	0.24–1.06	8.74
Bastin and Van der Linden (2003)	0.79	0.03	0.43–1.16	9.59
Belleville et al. (2006)	0.53	0.16	−0.25–1.32	4.22
Belànger and Belleville (2009)	0.81	0.12	0.14–1.48	5.25
Collette et al. (2009)	0.89	0.08	0.33–1.44	6.59
Morrone et al. (2010)	0.75	0.07	0.23–1.26	7.07
Borella et al. (2011a) (high working memory)	0.66	0.04	0.26–1.06	8.90
Borella et al. (2011a) (low working memory)	0.55	0.04	0.16–0.94	9.09
Stites et al. (2013)	0.32	0.11	−0.32–0.96	5.53
Wang and Su (2013)	0.63	0.06	0.17–1.10	7.82
Collette et al. (2014) (strong episodic memory)	0.96	0.09	0.37–1.55	6.14
Collette et al. (2014) (standard episodic memory)	0.64	0.08	0.07–1.21	6.36
Tournier et al. (2014) (blocked)	−0.37	0.07	−0.87–0.13	7.32
Tournier et al. (2014) (unblocked)	−0.09	0.06	−0.58–0.41	7.38
Overall	0.55	0.01	0.36–0.74	

participants produced more errors in the Inhibition section than the young controls. An analysis of the variability in the effect sizes of the 14 data sets revealed heterogeneity, with $Q(13) = 26.64$, $p < 0.01$, and $I^2 = 51.21\%$.

As Fig. 7 shows, in the context of the other studies, the study by Tournier and colleagues (2014) (two data sets) presented low negative effect sizes of −0.37 and −0.09, corresponding to the blocked and unblocked formats, respectively. These values indicate that the older participants made fewer errors than the younger ones, in contrast to what is frequently observed. One possibility for this result is the procedure used in this study. According to the authors, it was designed to reduce the age-related decline in the strategies that participants (especially the younger ones) commonly used to accomplish the task on the Inhibition section of the test, such as not paying attention to the sentence and anticipating the response. Additionally, the authors suggested a cohort effect to explain these results, that is, the possibility that the young and older participants had different educational backgrounds, favoring the older group.

After excluding this study, the summarized effect size was moderate, with a Hedges' g value of 0.69 (95% CI, 0.54–0.83, $Z = 9.54$, $p < 0.01$). The values of $Q(Q = 3.82$, $p = 0.97)$ and an I^2 value of 0% indicated homogeneity of the effect sizes from the pool of studies. Because there was heterogeneity, an analysis of moderators was performed, showing that the test format was significant with ($Q_{between}(1) = 4.59$, $p < 0.05$) (although the within variance was also significant ($Q_{within}(8) = 18.86$, $p < 0.01$), as well as language ($Q_{between}(2) = 21.48$, $p < 0.01$) (Q_{within} was not significant). On the other hand, the meta-regression analysis showed a statistically significant positive relationship between the studies effects sizes and vocabulary ($b_j = -0.04$, $Z = -2.53$, $p < 0.01$). The variables mean age of the elderly sample and education did not show significant relationships with the effect sizes.

In order to evaluate publication bias, the funnel plot was constructed (Fig. 8). In addition, Eggers's regression was not significant, with $B_0 = -0.54$, $t(12) = 0.28$, $p = 0.38$, indicating no evidence of publication bias. On the other hand, the trim and fill calculation yielded three imputed studies. The fail-safe N was 227.

Comparison of the Effect Sizes for the Automatic and Inhibition Conditions of the Hayling Test

Regarding the RTs, the results show (see the preceding sections) that the values of the effect sizes for the Automatic and Inhibition conditions were both high (according to Cohen's guidelines, Cohen, 1992) and very similar, with a difference of less than 0.20, and their CIs overlapped considerably. On the other hand, the values of the effect sizes for the error scores corresponding to the Automatic and Inhibition conditions were larger (low and moderate, respectively, according to Cohen's guidelines), and their CI overlapped much less. Thus, it was consider necessary to statistically test whether these differences in the effect size values between the Automatic and Inhibition conditions were significant. Both between ($Q_{between}$) and within

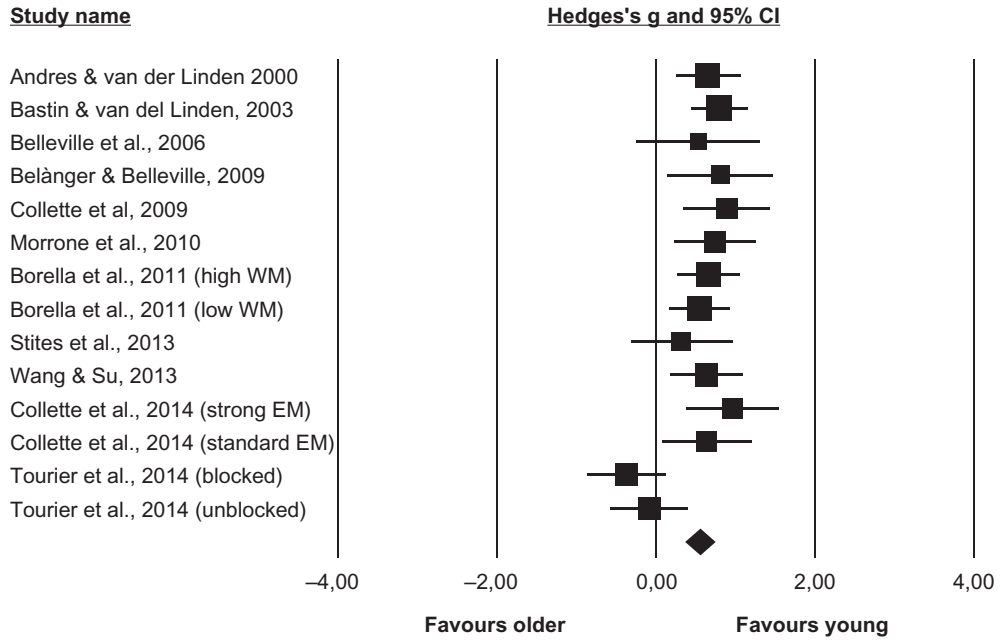

Fig. 7. Forest-plot of error scores in the Inhibition section of the Hayling task. WM, working memory; EM, episodic memory.

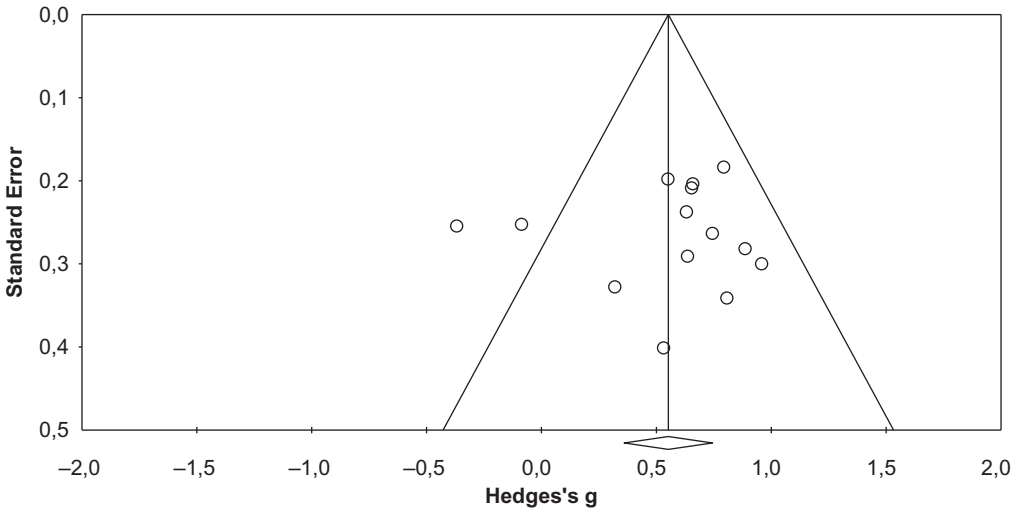

Fig. 8. Funnel plot for error scores in the Inhibition section of the Hayling task. The solid vertical line represents the weighted average effect size.

(Q_{within}) group variances were calculated yielding a value of ($Q_{between}(1) = 0.30$, $p = 0.58$) ($Q_{within}(20) = 184.18$, $p < 0.01$ was significant). Thus, there were no significant differences between Automatic and Inhibition conditions. Regarding the error scores, we obtained ($Q_{between}(1) = 14.36$, $p < 0.01$) (although the within variance was also significant ($Q_{within}(19) = 43.87$, $p < 0.01$), indicating that the differences between Automatic and Inhibition conditions were significant.

Discussion

A review of the literature that evaluated the performance on the Hayling task in younger and older adults showed that, while some studies found statistically significant age effects, others failed to reach statistical significance, giving the appearance of inconsistent results in this research field. Thus, a systematic literature search and a quantification of age-related differences in the performance on the Hayling test were conducted. Findings from this meta-analysis revealed that older adults performed worse than young adults on the four measures of the Hayling test: RTs and error scores in the Automatic and Inhibition sections. The magnitude of the age effect was larger for the latency responses (in both the Automatic and the Inhibition sections) than for the error scores in both Sections. The overall effect size obtained for the latency responses was high in both the Automatic (Hedges' $g = 0.81$) and Inhibition (Hedges' $g = 0.98$) parts, but this difference was not significant. On the other hand, the overall effect size for the error scores was low (Hedges' $g = 0.13$) in the Automatic section and moderate (Hedges' $g = 0.55$) in the Inhibition section, and this difference was significant. This finding agrees with the research literature in the field of cognitive aging, where slow of processing has been considered a central feature of the cognitive changes associated with age, and its relationship with inhibition has been pointed out in some studies (Salthouse, 1991).

With regard to the error scores, the overall effect size in the Automatic section was small. This result agrees with what was expected according to the findings reported in the literature. Moreover, some studies did not report this measure because of ceiling effects (Andres et al., 2000; Bastin & Van der Linden, 2003; Collette et al., 2009, 2014; Morrone et al., 2010; Wang & Su, 2013), as the task of completing the sentence with a congruent word when the sentence was highly predictable was very easy for healthy older individuals (and younger ones). By contrast, in the Inhibition section of the test, adding an incongruent word to a highly semantically constricted sentence is difficult for older individuals because it requires the inhibition of the overlearned automatic response. Congruently, the results of the present meta-analysis indicated that the magnitude of the age effect on the error scores in the Inhibition section was moderate.

Furthermore, the four analyses (one for each measure on the test) indicated that there was heterogeneity in the effect sizes of the studies in each of the four pools. In the case of the RTs (both in the Automatic and Inhibition sections), the heterogeneity in the effect sizes could be due to the influence of two studies that, in the context of the other studies, presented lower (Borella et al., 2011a) and higher (Tournier et al., 2014) effect sizes. Both studies differed from the other studies in the special characteristics of their experimental designs. On the other hand, in the Borella and colleagues (2011a) study, the effect size of the data set corresponding to individuals with high WM was negative (although small), indicating a slight advantage of the older participants, in contrast to the rest of the studies (and to the participants with lower WM in the same study). This study used an extreme groups design, where the participants (both young and older) were selected based on their scores on a WM task. Thus, one set of data from this study corresponded to participants with high WM (participants whose scores were in the upper tercile), and another set of data corresponded to participants with low WM (participants in the lower tercile) (participants with scores in the second tercile were excluded). These types of participants were not used in the rest of the studies, which typically used participants whose WM spans were in the middle range. On the other hand, the study by Tournier and colleagues (2014) (both data sets) showed the highest effect size values in the pool, in the expected direction, (favoring the younger participants). Based on the explanation given by the authors, one possible explanation for the high age effect sizes on the RTs in both sections of the test found in this study, would be the use of a forced-response choice format (instead of pronouncing the target word as in the rest of the studies). This type of response could be a disadvantage for older adults, compared to younger ones because, they needed more time to respond, although processing speed was a variable that was statistically controlled in this study.

There was also heterogeneity in the studies' effect sizes on the dependent measures of error scores in the Automatic and Inhibition sections. It is interesting to note that, for the error scores in the Inhibition section, the two sets of data in the study by Tournier and colleagues (2014) presented lower effect sizes than the other studies. In other words, the older participants performed better than the younger ones. It seems that the procedure used to administer the Hayling task produced longer response latencies (discussed above), but fewer errors. It should be noted that the objective of the study by Tournier and colleagues (2014) was to adjust the original task by Burgess and Shallice (1997) in order to keep younger participants from using some strategies that improve their performance on the task, such as pre-selecting the incongruent word to complete the sentence (according to the instructions given by the experimenter in the Inhibition section) before the sentence has been

completely heard and processed. According to the authors, this strategy benefits younger individuals more than older ones, who do not seem to use these types of strategies (Lemaire, 2010). By asking the participants to respond by choosing between two words, the use of this strategy was eliminated.

In addition to the format used to administer the test, some participant characteristics, such as language, years of education and vocabulary, might affect the age effect size on the different measures of the Hayling test, although their effects were different on each measure, making difficult to draw conclusions. Further research with more studies would be needed to address this issue.

In conclusion, the observed heterogeneity in the effect sizes across studies appears to partly reflect methodological differences or participant characteristics. Thus, the results of the present meta-analysis suggest that it is important for researchers and clinicians to select the procedure used to administer the test consider certain participant characteristics, such as WM capacity. However, the potential role of this variable could not be examined in this meta-analysis because there was only one study in the pool with this characteristic. Thus, the results of the present study represent an initial approach to the assessment of the effects of aging on semantic inhibitory capacity, measured by performance on the Hayling test.

Funding

This research was supported in part by research grant FF12014-54088-P (Ministry of Science and Innovation of Spain).

Conflict of Interest

None declared.

References

Abusamra, V., Miranda, M. A., & Ferreres, A. (2007). Evaluación de la iniciación verbal en español. Adaptación y normas del test Hayling. *Revista Argentina de Neuropsicología*, *9*, 19–32.

Amineva, H., Philips, L. H., Della Sala, S., & Henry, J. D. (2004). Inhibitory functioning in Alzheimer disease. *Brain*, *127*, 949–964. doi:10.1093/brain/awh045.

Amiri, M., Pouliot, P., Bonnéry, C., Leclerc, P. O., Desjardins, M., Lesage, F., et al (2014). An exploration of the effect of hemodynamic changes due to normal aging on the fNIRS response to semantic processing of words. *Frontiers in Neurology*, *5*, 249. doi:10.3389/fneur.2014.00249.

Andrés, P., & Van der Linden, M. (2000). Age-related differences in supervisory attentional system functions. *Journal of Gerontology, Series B. Psychological Sciences and Social Sciences*, *55*, 373–380. doi:10.1093/geronb/55.6.P373.

Andrés, P., & Van der Linden, M. (2001). Supervisory attentional system in patients with frocal frontal lesions. *Journal of Clinical and Experimental Neuropsychology*, *23*, 225–239. doi:10.1076/jcen.23.2.225.1212.

Bailey, P. E., & Henry, J. D. (2008). Growing less empathic with age: disinhibition of the self-perspective. *Journal of Gerontology: Psychological Sciences*, *63B*, P219–P226. doi:10.1093/geronb/63.4.P219.

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology*, *17*, 14–24. doi:10.1037/0894-4105.17.1.14.

Belànger, S., & Belleville, S. (2009). Semantic inhibition impairment in mild cognitive impairment: A distinctive feature of upcoming cognitive decline? *Neuropsychology*, *23*, 592–606. doi:10.1037/a0016152.

Belleville, S., Rouleau, N., & Van der Linden, M. (2006). Use of the Hayling task to measure inhibition of prepotent responses in normal aging and Alzheimer's disease. *Brain and Cognition*, *62*, 113–119. doi:10.1016/j.bandc.2006.04.00.

Bielak, A. A. M., Mansueti, L., Strauss, E., & Dixon, R. A. (2006). Performance on the Hayling and Brixton tests in older adults: Norms and correlates. *Archives of Clinical Neuropsychology*, *21*, 141–149. doi:10.1016/j.acn.2005.08.006.

Borella, E., Carretti, B., & Beni, R. D. (2008). Working memory and inhibition across the adult life-span. *Acta Psicologica*, *128*, 33–44. doi:10.1016/j.actpsy.2007.09.008.

Borella, E., Delaloye, C., Lecerf, T., Renaud, O., & Ribaupierre, A. (2009). Do age differences between young and older adults in inhibitory tasks depend on the degree of activation of information? *European Journal of Cognitive Psychology*, *21*, 445–472. doi:10.1080/09541440802613997.

Borella, E., Ghisletta, P., & de Ribaupierre, A. (2011b). Age differences in text processing: The role of working memory, inhibition, and processing speed. *Journal of Gerontology: Psychological Sciences*, *66*, 311–320. doi:10.1093/geronb/gbr002.

Borella, E., Ludwig, C., Fagot, D., & de Ribaupierre, A. (2011a). The effect of age and individual differences in attentional control: A sample case using the Hayling test. *Archives of Gerontology and Geriatrics*, *53*, e75–e80. doi:10.1016/j.archger.2010.11.005.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2005). *Comprehensive meta-analysis version 2*. Engelwood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. London: Wiley. doi:10.1002/9780470743386.ch29.

Burgess, P. W., & Shallice, T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia*, *34*, 263–273. DOI:10.1016/0028-3932(95)00104-2.

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton tests*. Bury St. Edmunds, England: Thames Valley Test Company. doi:10.1097/NMD.0000000000000366.

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*, 201–216. doi:10.1016/j.acn.2007.08.010.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155.

Collette, F., Germain, S., Hogge, M., & Van der Linden, M. (2009). Inhibitory control of memory in normal aging: Dissociation between impaired intentional and preserved unintentional processes. *Memory (Hove, England)*, *17*, 104–122. doi:10.1080/09658210802574146.

Collette, F., Grandjean, J., Lorant, C., & Bastin, C. (2014). The role of memory traces quality in directed forgetting: A comparison of young and older participants. *Psychologica Belgica*, *54*, 310–327. doi:10.5334/pb.au.

Delaloye, C., Ludwig, C., Borella, E., Chicherio, C., & de Ribaupierre, A. (2008). L'Empan de lecture comme épreuve mesurant la capacité de mé moire de travail: Normes basées sur une population francophone de 775 adultes jeunes et âgés. *European Review of Applied Psychology*, *58*, 89–103. doi:10.1016/j.erap.2006.12.004.

Duval, S., & Tweedie, R. L. (2000). Trim and fill: A simple funnel plot based method for testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. doi:10.1111/j.006-341×2000.00455x.

Egger, M., Smith, G., Schneider, M., & Minder, C. (1977). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. doi:10.1136/bmj.315.7109.629.

Fonseca, R. P., Oliveira, C. R., Gindri, G., Zimmermann, N., & Reppold, C. T. (2010). Teste Hayling: Um instrumento de avaliação de componentes das funções executivas. In Hutz, C. (Ed.) *Avaliação psicológica e neuropsicológica de crianças e adolescentes* (pp. 337–364). São Paulo: Casa do Psicólogo. doi:0.1590/S1413-82712011000100015.

Frias, C. M., Dixon, R. A., & Srauss, E. (2006). Structure of four executive functioning tests in healthy older adults. *Neuropsychology*, *20*, 206–214. doi:10.1037/0894-4105.20.2.206.

Frias, C. M., Dixon, R. A., & Srauss, E. (2009). Characterizing executive functioning in older special populations: From cognitively elite to cognitively impaired. *Neuropsychology*, *23*, 778–791. doi:10.1037/a0016743.

Hasher, L., Stoltzfus, E. R., Zacks, R. T., & Rypma, B. (1991). Age and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 163–169.

Hedges, L. V. (1994). Fixed effects models. In Cooper, H., & Hedges, L. V. (Eds.) *The handbook of research synthesis* (pp.285–299). New York, NY: Russell Sage Foundation. doi:10.1002/(SICI)1097-0258(19970330)16.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Higgins, J. P. T., & Altman, D. G. (2008). Assessing risk of bias in included studies. In Higgins, J. P. T., & Green, S. (Eds.) *Cochrane handbook for systematic reviews of interventions* (pp.187–241). Chichester, UK: Wiley, Chapter 8.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statatistical Medicine*, *21*, 1539–1558. doi:10.1002/sim.1186.

Hornberger, M., Geng, & Hodges, J. R. (2011). Convergent grey and white matter evidence of orbitofrontal cortex changes related to disinhibition in behavioral variant of frontotemporal dementia. *Brain*, *134*, 2502–2512. doi:10.1093/brain/awr173.

Hornberger, M., Piguet, C., Kipps, C., & Hodges, J. R. (2008). Executive function in progressive and nonprogressive behavioral variant of frontotemporal dementia. *Neurology*, *71*, 1481–1488. DOI:10.1212/01.wnl.0000334299.72023.c8.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologis*, *9*, 179–186. doi:10.1093/geront/9.3.

Lemaire, P. (2010). Cognitive strategy variations during aging. *Current Directions in Psychological Sciences*, *19*, 363–369. doi:10.1177/0963721410390354.

Lin, H., Chan, R. C. K., Zheng, L., Yang, T., & Wang, Y. (2007). Executive functioning in healthy elderly Chinese people. *Archives of Clinical Neuropsychology*, *22*, 501–511. doi:10.1016/j.acn.2007.01.028.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis*, applied social research methods s*eries. Thousand Oaks: Sage Publications.

Q10 MOOSE. Meta-analysis of observational studies in epidemiology. Retrieved from www.strobe-statemetn.org.

Morrone, I., Declercq, C., Novella, J. L., & Besche, C. (2010). Aging and inhibition process: The case of metaphor treatment. *Psychology and Aging*, *25*, 697–701. doi:10.1037/a0019578.

Nathan, J., Wilkinson, D., Stammers, S., & Low, L. (2001). The role of tests of frontal executive function in the detection of mild dementia. *International Journal of Geriatric Psychiatry*, *16*, 18–26. doi:10.1002/1099-1166(200101)16.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. (Center for Human Information ProcessingTechnical Report No. 99, rev. ed.). In Davidson, R. J., Schartz, G. E., & Shapiro, D. (Eds.) *Consciousness and self-regulation: Advances in research* (pp. 1–18). New York: Plenum Press. doi10.1007/978-1-4613-2900-8.

Oliveira, C. R., Pedron, A. C., Gonçalves-Gurgel, L., Tozzi-Reppold, C., & Fonseca, R. P. (2012). Executive functions and sustained attention. Comparison between age groups of 19–39 and 40–59 years old. *Dement Neuropsychology*, *6*, 29–34. doi:10.1016/B978-0-12-380882-0.00005-X.

Perry, R. J., Watson, P., & Hodges, J. R. (2000). The nature and staging of attention dysfunction in early (minimal and mild) Alzheimer's disease: Relationship to episodic and semantic memory impairment. *Neuropsychologia*, *38*, 252–271.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192. doi:10.1037/0033-2909.118.2.183.

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*, *2*, 179–183. doi:10.1111/j.1467-9280.1991.tb00127.x.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, England: Cambridge University Press. doi:0.1017/ CBO9780511526817.

Stites, M. C., Federmeier, K. D., & Stine-Morrow, E. A. L. (2013). Cross-age comparisons reveal multiple strategies for lexical ambiguity resolution during natural reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1823–1841. doi:10.1037/a0032860.

STROBE (Strengthening the reporting of observational studies in epidemiology). Retrieved from www.strobe-statement.org.

Stuss, D. T., & Alexander, M. (2000). Executive functions and the frontal lobes: A conceptual view. *Psychological Research*, *63*, 289–298. doi:10.1007/s004269900007.

Tournier, I., Posta, V., & Mathey, S. (2014). Investigation of age-related differences in an adapted Hayling task. *Archives of Gerontology and Geriatrics*, *59*, 599–606. doi:10.1016/j.archger.2014.07.016.

Wang, Z., & Su, Y. (2013). Age-related differences in the performance of theory of mind in older adults: A dissociation of cognitive and affective components. *Psychology and Aging*, *28*, 284–291. doi:10.1037/a0030876.

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

Yeung, S. E., Fischer, A. L., & Dixon, R. A. (2009). Exploring effects of type 2 diabetes on cognitive functioning in older adults. *Neuropsychology*, *23*, 1–9. doi:10.1037/a0013849.