Psychonomic Bulletin and Review (2018) 25:2175–2199
https://doi.org/10.3758/s13423-018-1461-6

Retest effects in working memory capacity tests: A meta-analysis

Jana Scharfen[1] · Katrin Jansen[1] · Heinz Holling[1]

Published online: 15 June 2018

Abstract

The repeated administration of working memory capacity tests is common in clinical and research settings. For cognitive ability tests and different neuropsychological tests, meta-analyses have shown that they are prone to retest effects, which have to be accounted for when interpreting retest scores. Using a multilevel approach, this meta-analysis aims at showing the reproducibility of retest effects in working memory capacity tests for up to seven test administrations, and examines the impact of the length of the test-retest interval, test modality, equivalence of test forms and participant age on the size of retest effects. Furthermore, it is assessed whether the size of retest effects depends on the test paradigm. An extensive literature search revealed 234 effect sizes from 95 samples and 68 studies, in which healthy participants between 12 and 70 years repeatedly performed a working memory capacity test. Results yield a weighted average of $g = 0.28$ for retest effects from the first to the second test administration, and a significant increase in effect sizes was observed up to the fourth test administration. The length of the test-retest interval and publication year were found to moderate the size of retest effects. Retest effects differed between the paradigms of working memory capacity tests. These findings call for the development and use of appropriate experimental or statistical methods to address retest effects in working memory capacity tests.

Keywords Meta-analysis · Retest effect · Practice effect · Working memory

Retest effects in working memory capacity tests

The repeated administration of a working memory capacity test is a common scenario in clinical settings as well as in behavioral and neuropsychological research. It is relevant, for example, when examining the rate of disease progression in Alzheimer's disease (e.g., Goldberg, Harvey, Wesnes, Snyder, & Schneider, 2015) or age-related cognitive decline in healthy adults (e.g., González, Tarraf, Bowen, Johnson-Jennings, & Fisher, 2013), or when evaluating the effectiveness of cognitive training programs (e.g., Jaeggi, Buschkühl, Jonides, & Perrig, 2008). When identical tests or alternate forms of a test are repeatedly administered, an improvement in test scores can be observed, a finding which is known as the practice effect, or retest effect (e.g., Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Heilbronner et al., 2010; Lievens, Buyse, & Sackett, 2005). On an individual level, failing to account for retest effects can easily lead to wrong treatment decisions, for example in patients with Alzheimer's disease or concussion. On an experimental level, not considering retest effects when evaluating a cognitive training would result in an overestimation of training effects. Thus, retest effects have to be taken into account when interpreting scores of repeated test administrations in both practical and in experimental settings (Green, Strobach, & Schubert, 2014; Heilbronner et al., 2010).

However, although retest effects are an established phenomenon, knowledge about the conditions under which

Jana Scharfen and Katrin Jansen declare a shared first authorship

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-018-1461-6) contains supplementary material, which is available to authorized users.

✉ Jana Scharfen
jana.scharfen@uni-muenster.de

Katrin Jansen
katrinjansen@uni-muenster.de

Heinz Holling
holling@uni-muenster.de

[1] Institute of Psychology, Westfälische Wilhelms-Universität Münster, Fliednerstr. 21, 48149 Muenster, Germany

they appear is relatively scarce. A meta-analysis by Hausknecht et al. (2007) yielded an overall retest effect of d = 0.24 in cognitive ability tests by integrating 75 samples. Moderator analyses indicated that identical test forms produced larger retest effects than alternate test forms, and that retest effects were larger in the presence of coaching before the second test administration.

More recently, Scharfen, Peters, and Holling (2018) meta-analyzed retest effect sizes in a wide range of cognitive abilities over several test repetitions and found effects of similar size that reached a plateau after the third test administration. Effect sizes were moderated by cognitive abilities and content measured by the test, test forms, test-retest intervals, and participant age.

Calamia, Markon, and Tranel (2012) investigated retest effects in a variety of tests used frequently by clinical neuropsychologists, and found effect sizes ranging between mean differences of $d = 0.16$ and $d = 0.34$ across different domains. Common moderators were the use of alternate forms, the age of participants, clinical diagnoses of participants, and the length of the test-retest interval. For the subtests Digit Span, Arithmetic and Letter Number Sequencing of the Wechsler Adult Intelligence Scale (over several versions, e.g., English- language, German-language and editions, e.g., WAIS-R, WAIS-III), which were among the chosen measures for the domain of auditory attention and working memory, effect size estimates ranged between $d = 0.17$ and $d = 0.27$. At the same time, their results suggested that there was considerable variation across different measures within most of the domains examined. These findings raise the question whether these results can be generalized to measures of working memory such as N-back or complex span tasks, which are popular in neuropsychological and behavioural research (Kane & Engle, 2002; Conway et al., 2005). Moreover, the number of outcomes included for tests of working memory capacity was below 20 for some of the measures of working memory capacity, which is especially problematic when investigating effects of potential moderators. Finally, the statistical model applied by Calamia et al. (2012) did not account for dependencies among effect sizes, although separate effect sizes were calculated for multiple outcomes from the same study.

To the best of our knowledge, there has not been any meta-analysis on retest effects focusing on working memory tests in a representative sample of healthy subjects. Also, in most of the related meta-analyses mentioned above, retest effects resulting from multiple test administrations were neglected. Thus, the objective of the present meta-analysis is to provide estimates of the overall sizes of retest effects for multiple test administrations in working memory tests and to investigate potential moderators, using a multilevel model to account for dependencies among effect sizes. It will be shown that the common phenomenon of retest effects can be replicated in working memory capacity tests in various contexts.

Working memory

Working memory has become one of the most popular and important constructs in cognitive and differential psychology over the last years. Indeed, working memory processes play an important role in many complex cognitive operations, such as text and reading comprehension (Daneman & Carpenter, 1980), mathematic skills (Peng, Namkung, Barnes, & Sun, 2015) or reasoning (Kyllonen & Christal, 1990), and working memory was shown to be correlated with intelligence (Ackerman, Beier, & Boyle, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005). In recent years, research on working memory mainly focused on the effects of working memory training, resulting in a debate about the possibility of its transfer to complex cognitions (Au et al., 2015; Au, Buschkühl, Duncan, & Jaeggi, 2016; Karbach & Verhaeghen, 2014; Melby-Lervåg & Hulme, 2013, 2016; Shipstead, Redick, & Engle, 2012; Soveri, Antfolk, Karlsson, Salo, & Laine, 2017), while neglecting the effect that results from simple retesting.

Working memory was first introduced by Baddeley and Hitch (1974) who proposed the three-component model of working memory, a memory system consisting of a central executive and two slave systems, the phonological loop and the visuospatial sketchpad. Later, the model was expanded by another subsystem: the episodic buffer (Baddeley, 2000). Since Baddeley's conceptualization, different frameworks and models of working memory have been proposed, for example by Cowan (1999), Kane and Engle (2002), Unsworth and Engle (2007), and Oberauer (2009). In all models, working memory is conceptualized as a system responsible for the temporary storage and processing of information, where information is likely to be held in chunks, and it is restricted by the capacity limit of storage systems and the effectiveness of executive processes. Moreover, working memory is linked to long-term memory. Models differ in the specification of subcomponents and processes engaged when working memory is used.

Tests of working memory capacity

Depending on specifications of components and processes proposed by different models of working memory, several methods have been developed to assess working memory capacity. Most instruments focus on individual differences in working memory storage capacity. The choice of a method to assess working memory capacity depends on both the research domain and specific research interests (Conway et al., 2005). Whereas complex span tasks dominate in the

field of cognitive psychology, the N-back task has become increasingly popular within neuropsychological research (Kane & Engle, 2002). In clinical settings, different versions of tasks from the Wechsler Adult Intelligence Scale IV (Wechsler, 2008), such as digit span (backward) and letter number sequencing, are frequently used to assess working memory capacity (Calamia et al., 2012).

Consistent with the common ground of the theoretical frameworks described above, a majority of tasks supposed to assess working memory capacity focuses on the requirement of simultaneous storage and processing (Conway et al., 2005). A psychometric study by Schmiedek, Lövdén, and Lindenberger (2014) indicates that complex span tasks, memory updating tasks, sorting span (also known as transformation span) tasks, and N-back tasks all measure the same construct: working memory capacity. Conway et al. (2005) and Wilhelm, Hildebrandt, and Oberauer (2013) further argue that running memory tasks reflect the same "keeping-track", or updating processes as memory updating or N-back tasks, and thus can be used as working memory measures as well. Ultimately, tasks requiring mainly coordination, i.e. the integration of elements into a structure, provide another method to assess working memory capacity, even though they do not necessarily have a storage component (Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003). This is consistent with the conceptualization that working memory operates by integrating elements into a common cognitive coordinate system (Oberauer, 2009). Although, according to the framework by Kane and Engle (2002), measures of inhibition such as Stroop, antisaccade, and dichotic listening tasks (Conway & Kane, 2001; Engle, 2001) reliably predict individual differences in working memory capacity, recent factor-analytical research does not support this claim (Wilhelm et al., 2013). Similarly, measures of task switching were not found to load substantially on the same factor as commonly acknowledged measures of working memory capacity (Oberauer et al., 2000, 2003). Furthermore, it has been argued that both change detection and serial recall tasks reflect the same components as complex span tasks (Unsworth & Engle, 2007; Unsworth, Fukuda, Awh, & Vogel, 2014), but presumably to a different extent. Serial recall tasks draw mainly on primary memory, while complex span tasks always require secondary memory resources (Unsworth & Engle, 2007). Consistent with the assumption that serial recall is not the same as working memory capacity is the finding that the dorsolateral prefrontal cortex shows higher activation for a sorting span task than for a simple span task (D'Eposito, Postle, & Rypma, 2000). A similar assumption holds for change detection tasks, which reflect mainly the scope of attention and attentional control (Unsworth et al., 2014). For these reasons, measures of inhibition, task switching, change detection and serial recall were not approved as measures of working memory capacity in this analysis.

A brief description of the paradigms of working memory capacity tasks can be found in Table 1. In accordance with the working memory model described by Baddeley (2000), working memory capacity tests can be further distinguished according to the test modality, that is, whether they contain verbal-numerical or visuospatial stimuli (Oberauer et al., 2000, 2003).

Retest effects

There are various possible explanations for the emergence of retest effects. Lievens, Reeve, and Heggestad (2007) subsume the causes of retest effects which have been proposed under three categories: Firstly, retest effects might reflect a true gain in the construct measured by the test. For memory tests, this phenomenon has been termed the *testing effect* (Roediger & Karpicke, 2006), referring to the finding that the repeated administration of a memory tests enhances long-term retention.

However, Lievens et al. (2007) argue that it seems unlikely that a latent complex cognitive ability, i.e. a highly *g*-loaded construct, can be enhanced by simple retesting and that an improved test performance can rather be explained by the following two causes. It is thus unlikely that this explanation applies to retest effects in working memory capacity tests, which share substantial amounts of variance with measures of fluid intelligence (Kyllonen & Christal, 1990).

Secondly, retesting might lead to the reduction of debilitating construct-irrelevant factors, such as test anxiety or the lack of familiarity with the material and response mode of the test, and thereby increase the validity of retest scores compared to baseline scores (Anastasi, 1981; Freund & Holling, 2011; Kliegl & Baltes, 1987).

This perspective is supported by research on situational factors accounting for an increase in test scores in cognitive ability tests (Matton, Vautier, & Raufaste, 2009). As tests of working memory capacity often confront participants with unfamiliar task demands (consider, e.g., a 3-back task), one could expect that participants become more and more used to these demands during the first test administration and perform better in a retest situation, when they are already familiar with the task in the beginning of the test administration. In accordance, an fMRI study on changes in cerebral activation after working memory practice by Jolles, Grol, van Buchem, Rombouts, and Crone (2010) revealed changes in the activation of the bilateral dorsolateral prefrontal cortex and the superior parietal cortex both in a practice group and a control group which only participated in pre- and post-tests, possibly reflecting an increase in familiarity with the test.

Table 1 Description of the different task paradigms

Paradigm	Description	Examples
N-back	Participants monitor a sequence of stimuli (e.g., letters, digits) and are required to indicate whether the current stimulus is the same as the one n steps back in the sequence.	N-back (Kirchner, 1958), dual N-back (Jaeggi et al., 2008)
Complex span	Participants perform a processing task (e.g., evaluating sentences or math equations) while memorizing stimuli for later sequential recall.	reading span (Daneman & Carpenter, 1980), computation span (Turner & Engle, 1989)
Memory updating	Participants are presented with a number of stimuli and a sequence of updating operations which they have to apply to the stimuli. Finally, participants recall the result of the updating operations.	memory updating (numerical or spatial) (Salthouse et al., 1991)
Running span	Participants are presented with a sequence of stimuli of an unknown length. At the end of the sequence, they are asked to recall either a predefined number of the final stimuli or as many stimuli as they remember.	running memory span (Pollack et al., 1959)
Transformation span	Participants are presented with a sequence of stimuli and either have to recall them in a predefined order (e.g., backwards), or indicate whether a number presented together with another stimulus (e.g., letter) matches the position of the stimulus in a predefined order (e.g., alphabetical).	digit span backward, letter number sequencing (Wechsler, 2008), alpha span (Craik, 1986)
Coordination	Participants view a display of changing stimuli (e.g., numbers in a matrix), and are asked to give a response whenever the stimuli fulfill a predefined relational criterion (e.g., numbers in a row or column can all be divided by three).	coordination verbal, coordination spatial (Oberauer, 1993), flight control (Oberauer et al., 2003)

Thirdly, retest effects might reflect an increase in method-specific knowledge (e.g., knowledge of test-specific strategies), specific item content knowledge, or test-specific narrower cognitive abilities (e.g., memorization of items). In summary, this explanation includes all the factors which might "contaminate" the scores of the second (or any later) test administration. Lievens et al. (2007) found that retest scores in cognitive ability tests were less *g*-loaded and instead more strongly reflected variance due to memory. The notion that score gains might reflect narrow ability components has been shared by a few other researchers (Jensen, 1998; Lubinski, 2000; Arendasy & Sommer, 2013). In a study on the reliability of verbal-numerical working memory tasks (Beckmann, Holling, & Kuhn, 2007), exploratory factor analyses revealed an additional factor at the second test administration, possibly reflecting a task-specific practice effect. In addition, there is evidence that when taking a working memory test, participants are likely to develop strategies such as rehearsal processes in form of covert retrieval in complex span tasks (Turley-Ames & Whitfield, 2003; Friedman & Miyake, 2004; McCabe, 2008), or selective strategies in memory updating tasks (Shing, Schmiedek, Lövdén, & Lindenberger, 2012). As an example, Dunning and Holmes (2014) found an increase in the number of participants who used a grouping strategy in a visuospatial working memory task and a visualization strategy in a verbal working memory task from a baseline administration to a retest administration.

When seeking to increase working memory capacity, training studies often use strategy-induction or instruction on how to work on a task. Studies from the field of testing the limits of working memory capacity (Oberauer & Kliegl, 2001, 2006; Oberauer, Farrell, Jarrold, & Lewandowsky, 2016) make use of this approach and find evidence that strategy induction increases test performance (Baltes & Kliegl, 1992; Kliegl, Smith, & Baltes, 1989; Kliegl & Baltes, 1987).

Hence, causes of retest effects in cognitive ability tests should, in an analogous manner, apply to tests of working memory capacity. Based on the research evidence presented so far and because causes of retest effects can be adapted to working memory tests, it is hypothesized that there are retest effects in working memory capacity tests (Hypothesis 1).

When participants are assessed with the same test more than twice, an increase in test scores of cognitive ability tests was observed also between later test administrations (Hausknecht et al., 2007; Scharfen et al., 2018). Studies on retest effects in various neuropsychological measures, amongst others tests of working memory capacity, demonstrated that retest effects can be observed up to the third or fourth test administration, and then tend to reach a plateau, meaning that scores stop increasing (Collie, Maruff,

Darby, & McStephen, 2003; Beglinger et al., 2005; Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010; Scharfen et al., 2018). In general, retest effects were found to be largest from the first to the second test administration. It is plausible to assume that the influence of the causes of retest effects mentioned above decreases with the number of test repetitions. For example, the highest gain in test-specific strategies should be developed over the first administrations and test anxiety should be reduced mostly after the first tests. This leads to retest effects leveling out after a few administrations. Thus it is hypothesized that when a test is administered more than twice, an increase in test scores can be observed beyond the second test administration, but retest effects reach a plateau after a few administrations (Hypothesis 2).

Moderators of retest effects

Studies in which cognitive or neuropsychological tests are repeatedly administered differ with respect to methodological characteristics such as the test-retest interval, test characteristics such as the modality of stimulus material or whether identical or alternate forms are used, and participant characteristics, such as participant age, intelligence, or neurological status (McCaffrey, Duff, & Westervelt, 2000). This allows for the investigation of their moderating impact on the size of retest effects and thus their reproducibility over several settings. Randall and Villado (2017), in their review on retest effects in cognitive ability tests in employment settings, mention further possible moderating variables for retest effects, like, e.g., motivation, personality or emotions. However, primary studies do seldomly provide information about these variables, which is why in this meta-analysis we focus on those variables, for which sufficient information is given in most primary studies and for which standardized measures exist to ensure comparability between samples.

Test-retest interval There is considerable evidence that retest score gains in cognitive and neuropsychological tests decline with larger test-retest intervals (Salthouse, Schroeder, & Ferrer, 2004; Hausknecht et al., 2007; Calamia et al., 2012; Scharfen et al., 2018). However, estimates on how long the test-retest interval must be in order to eliminate retest effects range between about two and thirteen years, differing both within and between studies (Salthouse et al., 2004; Calamia et al., 2012). When taking a test, participants might form memory traces either of specific items or test-taking strategies (e.g., Arendasy & Sommer, 2017), and if these memory traces decay over time, there will be a decrease in retest effects depending on the length of the test-retest interval. Thus, it is hypothesized that retest effects in working memory capacity effects decline with larger test-retest intervals (Hypothesis 3a).

Test form Recommendations on how to address practice effects include the use of alternate forms (McCaffrey et al., 2000), and support for this recommendation can be found in primary studies (Salthouse & Tucker-Drob, 2008) and meta-analytical research on retest effects (Calamia et al., 2012; Hausknecht et al., 2007; Scharfen et al., 2018). Assuming that retest effects are test-specific and emerge for instance due to the development of test-taking strategies, it seems likely that the use of alternate forms cannot fully eliminate retest effects, especially if test items only differ in non-salient surface features (Arendasy & Sommer, 2013; Matton, Vautier, & Raufaste, 2011). However, larger score gains in identical tests would be expected if retest effects were, at least in part, item-specific. Lievens et al. (2007) found that, when administering a cognitive ability test for the second time, item discrimination parameters were reduced and while items were less *g*-loaded, their correlation with long-term memory increased, possibly indicating an item-specific retest effect. Based on the strong research evidence for the moderating effect of the test form, it is hypothesized that retest effects in working memory capacity tests are larger when an identical test is used for retesting than when an alternate version is administered (Hypothesis 3b).

Test modality As described above, tests of working memory capacity contain either verbal-numerical or visuospatial stimuli. For memory tests, Benedict and Zgaljardic (1998) found that retest effects were more pronounced for a visuospatial memory test compared to a verbal memory test, a finding which the authors attribute to the novelty of the testing procedure, arguing that participants generally are more likely to be familiar with tasks containing verbal or numerical stimuli. A similar argument is used by Salthouse and Tucker-Drob (2008) to explain the finding that larger retest effects were found in spatial tests compared to other domains. Working memory capacity tests with verbal-numerical material often use digits or letters as stimuli, whereas visuospatial working memory tests often require participants to deal with abstract arrays in a form which most participants are unlikely to encounter in their everyday lives. Consistent with the second explanation of retest effects described by Lievens et al. (2007), enhanced familiarity with the test material might play a role when people score higher in a retest situation. If participants are already familiar with the test stimuli from their everyday life, there will not be a large increase in familiarity due to retesting, leading to a smaller retest effect. Therefore, it is hypothesized that retest effects are larger in visuospatial working

memory capacity tests than in verbal-numerical working memory capacity tests (Hypothesis 3c).

Participant age Concerning participant characteristics, research evidence suggests that retest effects decline with age. For example, Salthouse et al. (2004) and Salthouse (2010, 2015) found a linear effect of age on the size of retest effects in different measures of cognitive functioning. A study examining cognitive change in a large sample of adults with retest intervals ranging between one and eight years revealed that the influence of age on the size of retest effects was independent of the length of the test-retest interval, and rather exerted influence near the first test administration (Salthouse, 2011). The finding that retest effects decline with age has also been confirmed by meta-analytical research on neuropsychological tests (Calamia et al., 2012; Scharfen et al., 2018), and can be explained in terms of neurocognitive changes. Braver and Barch (2002) postulate that changes in the function of the dopamine system, which projects to the prefrontal cortex, lead to a decline in the ability to represent, maintain and update context information in healthy older adults.

Also, differences regarding the limits of working memory capacity between younger and older samples suggests higher plasticity and different ways to work on working memory tasks in younger participants (Kliegl, Maayr, & Krampe, 1994; Mayr & Kliegl, 1993; Oberauer & Kliegl, 2001; Towse, Hitch, & Hutton, 2000).

These changes might lead to diminished benefits from a first test administration in older adults. Thus, it is hypothesized that retest effects in working memory capacity tests decline with increasing participant age (Hypothesis 3d).

Test-specificity of retest effects Retest effects were found to vary not only across cognitive domains (Wilson, Li, Bienias, & Bennett, 2006; Salthouse & Tucker-Drob, 2008), but also across different tests within the same domain (Salthouse & Tucker-Drob, 2008; Calamia et al., 2012). If this observation is due to test-specific characteristics, it raises the question whether retest effects in working memory capacity tests vary across the different task paradigms described above. Some characteristics are shared by the different paradigms (e.g., both N-back and transformation span and running span paradigms often use digits or letters as stimuli), whereas others are not (e.g., N-back requires recognition while transformation span and running span require sequential recall). As a meta-analytical approach is not suited for disentangling the effects of these characteristics, it is difficult to hypothesize to what extent retest effects might vary across paradigms. Thus, retest effects for the different paradigms will be investigated on an exploratory level.

Method

Literature search

The databases PsycINFO, PSYNDEX, PsycARTICLES and PsycCRITIQUES were searched for all years of publication until December 2016 with a combination of the search terms test*, assess*, retest*, repeat*, repetit*, practice, retak*, train*, coach*, "working memory", "cognitive abilit*", and the names of the paradigms and specific tests as listed in Table 1. The search was limited to studies published in English or German language. The inital search yielded 12,379 studies that were screened for eligibility by reading titles and abstracts.

In addition, a forward and backward search on relevant meta-analyses of working memory and cognitive ability trainings and retest effects was conducted (Au et al., 2015; Ball, Edwards, & Ross, 2007; Brunoni & Vanderhasselt, 2014; Calamia et al., 2012; Hausknecht et al., 2007; Karch, Albers, Renner, Lichtenauer, & von Kries, 2013; Kelly et al., 2014; Kulik, Kulik, & Bangert, 1984; Lampit, Hallock, & Valenzuela, 2014; Melby-Lervåg & Hulme, 2013; Morrison & Chein, 2011; Powers, Brooks, Aldrich, Palladino, & Alfieri, 2013; Scharfen et al., 2018; Schuerger & Witt, 1989; Smith et al., 2010; Soveri et al., 2017; Toril, Reales, & Ballesteros, 2014; Wang et al., 2016; Zehnder, Martin, Altgassen, & Clare, 2009). This lead to additional 657 possibly relevant studies. All studies that seemed appropriate were checked for eligibility in detail by applying the strict inclusion and exclusion criteria described below. A total number of 68 studies including 95 samples were in accordance with the criteria. The most common reasons for exclusion were failure to fulfill either the retesting conditions or the definition of working memory capacity tests.

Inclusion and exclusion criteria

Studies had to fulfill the following criteria:

(a) Studies must be written either in English or in German, so that a high validity of coding could be ensured. Generally, the studies not fulfilling this criterion were already excluded during abstract screening.
(b) In line with the definition of retest effects (Heilbronner et al., 2010; Lievens et al., 2007), an identical or alternate but equally difficult test had to be administered at least twice within the same sample. If there was any systematic activity between the initial test and the retest, samples had to be excluded as this activity might deter the retest effect. With regard to intervention studies, this lead to an exclusion of

any groups, e.g., placebo control groups, in which other activities than retesting took place between pre and post measurements. We also excluded pre and post measurements of training samples, if activities other than retesting were administered as the training intervention. If the training itself consisted of retesting as defined above, this data from training samples was included. Pre and post measurements of passive control groups mostly fulfilled this criteria. The test-retest interval was not allowed to be longer than 7 years, because a true change in the latent variable due to cognitive decline could not be excluded and retest effects were found to diminish after 7 years following Calamia et al. (2012).

(c) The test performed must require either simultaneous storage and processing of stimuli or the relational integration of elements into structures to be defined as a working memory capacity test, as defined above (Conway et al., 2005; Oberauer et al., 2000, 2003). Studies were not included if the test only required the concurrent processing of stimuli or task sets, thereby excluding tests that measure executive functions without having a storage component.

(d) The test score must be measured in terms of accuracy, including outcome measures such as scores, hits minus false alarms, percentage of correct responses or errors, or number of errors. Tests were not included if the only reported outcome was a reaction time measure, as reaction times reductions due to retesting do not necessarily reflect an improvement in test performance and should be differentiated from score change according to Ackerman (1987) and Scharfen, Blum and Holling (2018).

(e) The mean scores of the first test administration must be below the maximum score reachable (e.g., accuracy below 100%) to ensure the absence of ceiling effects.

(f) Participants in the experiment must be mentally and physically healthy. Calamia et al. (2012) found clinical samples to show smaller retest effects. As for different kinds of clinical conditions, cognitive impairment can differ in its degree, differences in retest effects between different clinical conditions can be expected (Calamia et al., 2012). Also, within clinical samples, severity of cognitive impairment can vary. It is the goal of this meta-analysis to give consistent estimates of retest effects for a broad population and therefore clinical samples were excluded. Samples were excluded if one or more persons from the sample were reported to have any kind of clinical condition. If the study did not report any information on the health status of the samples, it was assumed to be intact. Also, participants had to be on average between 12 and 70 years old to guarantee a stable latent working memory capacity during the test-retest interval that prevents retest effects to be contaminated by a latent change due to cognitive development. Below 12 and beyond 70 years of age, cognitive ability and thus working memory capacity can be assumed change more rapidly compared to other periods in life (Cattell, 1987; Glisky, 2007; Shaffer & Kipp, 2010).

(g) Effect size calculation must be possible with the information provided in the study. If a study initially failed to fulfill this criterion, corresponding authors were contacted. Six out of fifteen authors provided the necessary data.

Coding of variables

A coding scheme was developed that incorporated all relevant study characteristics and moderators. Coding was conducted by one of the authors. Coding of all studies was carefully double-checked and outliers were inspected carefully for coding errors to ensure correctness.

First, general information about the study and its source, such as the year of publication, the type of publication, the country in which the study was conducted, and whether the study was published in a peer-reviewed journal or not, were coded.

The number of test administrations and the time elapsed between the administrations (test-retest interval, coded in hours) were coded individually for each administration.

Further, it was coded whether the sample was a control group or an experimental group (e.g, a training group of which training data were included in the meta-analysis). Sample size was coded for each administration. The mean and standard deviation of the sample's age, and the percentage of male participants were noted for the first administration. Dropout rate was recorded as the total dropout rate throughout the whole study.

The name of the working memory capacity test administered was coded as mentioned in the study. Test characteristics included test content (verbal-numerical or visuospatial-figural), test paradigm (N-back, complex span, memory updating, running span, sorting span, coordination, or other), outcome measure, and internal consistency of the test measured by Cronbach's α. Table 13 in the Appendix summarizes which tests were assigned to which paradigm. For each but the first test administration, it was coded whether the same or an alternate form of the test, relative to the previous administration, was used. In addition, it was coded whether the participants received practice on the task, and if so, the number of practice items was recorded. Often, more than one test was administered, resulting in multiple effect sizes for a majority of studies.

Meta-analytic strategy

Effect size calculation Effect sizes were calculated as the standardized mean difference between the scores of two test administrations in terms of Hedges' *g* (Hedges, 1981), thereby correcting for sample size bias. Despite investigating the effects of the repeated administration of a test within the same group, the standardized mean difference was chosen as an effect size metric in preference to the standardized mean change (for calculation formulas, see Borenstein, 2009, p. 226–227). Calculating the standardized mean change would require the correlation between the respecting test administrations, which was not often reported in the studies included in the present meta-analysis. In this case, using the standardized mean change is not recommended (Morris & DeShon, 2002). Morris and DeShon further argue that standardized mean changes should only be used if the pre-post correlations can be assumed to be homogeneous for all outcomes, which is unlikely in our case, especially due to the large variation of the time interval between test administrations, because test-retest correlations decrease with increasing test-retest intervals (Calamia, Markon, & Tranel, 2013). Moreover, there is a bias of standardized mean changes towards an overestimation of the true population value (Dunlap, Cortina, Vaslow, & Burke, 1996; Morris, 2000; Hunter & Schmidt, 2004). Since other meta-analyses of retest effects also report the standardized mean difference (Calamia et al., 2012; Hausknecht et al., 2007), it will be possible to compare the results of the present meta-analysis to prior research.

Effect size estimates were obtained for all tests administered within a sample and for each combination of test administrations (e.g., first and second administration, first and third administration, and second and third administration if the test was administered three times). If multiple outcome measures were reported for one test, the most common outcome measure was chosen for effect size calculation. Error outcomes were reversed prior to effect size calculation. Effect sizes were coded as positive if there was a gain in retest scores.

In most cases, effect sizes were computed from the means and standard deviations of the respective test scores using the *metafor* package for *R* (R Core Team, 2017; Viechtbauer, 2010). If means and standard deviations were not given, effect sizes were calculated from *t*-values using the formulas given in Borenstein (2009, p. 228).

Multilevel meta-analysis As mentioned above, multiple effect sizes were obtained for a majority of samples. As described by Olkin and Gleser (2009), failing to account for dependence among effect sizes will produce biased estimates. Therefore, a multilevel approach to meta-analysis was chosen. Specifically, a three-level meta-analytic model was applied, taking into account the sampling variance of the estimated effect sizes (level one), the variance of effect sizes of different outcomes within a sample (level two), and the variance of effect sizes between samples (level three) (for a detailed description, see van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2015). Main advantages of this approach are that it does not require any knowledge about or estimation of the correlations between the different outcomes within a sample, and that it allows for a variation in the number of effect sizes per sample (van den Noortgate et al., 2015).

The same multilevel approach was used both for the investigation of an overall effect and in the analysis of potential moderators. All analyses were conducted using the rma.mv() function of the metafor package (Viechtbauer, 2010). Homogeneity was assessed by *Q*-test for all models. However, if *Q*-tests indicated homogeneity, random effects models and moderator analyses were applied due to theoretical assumptions, because the studies included differed considerably in their methods and the tests used for the assessment of working memory capacity. For directed hypotheses, one-sided *p*-values, for explorative analyses, two-sided *p*-values, using an alpha level of .05 are reported. Both theoretically assumed and, if significant, exploratory moderators were examined for collinearity. Publication bias was assessed using funnel plots (Sterne & Egger, 2001).

Results

Description of studies

A total number of 68 studies was included in the present meta-analysis, including 95 samples and a total sample size of $N = 3{,}281$. The total number of outcomes was 234.

A reference list containing the studies included in the meta-analysis is provided in Online Supplement 1. Online Supplement 2 lists all studies and the included samples, including effect sizes and coded characteristics of the most important methodological, sample, and test characteristics.

Of the 68 studies, only two were not published in a peer-reviewed journal (Buschkühl, 2007; Lange, 2013). Studies were published between the years 1987 and 2016, and the median year of publication was 2010. Most of the studies were conducted in the United States, in Germany, and Australia.

Test administration characteristics The number of test administrations per sample ranged between 2 and 20, with a mean of 3 and a median of 2. Table 2 provides a summary of the number of studies, samples and outcomes, sample sizes and test-retest interval in weeks per number of test administrations. For the first five test administrations, the median sample size was noticeably smaller than the mean

Table 2 Test administration characteristics

	2	3	4	5	6	7
No. of...						
studies	68	23	13	9	4	4
samples	95	29	16	12	6	6
outcomes	234	66	27	22	10	10
Sample sizes						
N	3,281	1,121	642	451	103	103
M	35	39	40	38	17	17
Mdn	19	20	19	18	15	15
Test-retest interval[a]						
M	7.65	16.80	9.74	7.67	5.00	9.41
SD	12.75	27.80	16.20	14.59	10.15	20.58
Mdn	3.00	4.00	1.80	0.90	1.00	1.14
Year of publication						
Mdn	2010	2009	2009	2010	2010	2010
Min	1987	1989	1989	1989	2000	2000
Max	2016	2016	2015	2015	2012	2012

Note. Statistics of the test-retest interval were calculated on the level of samples. *N* = Total sample size; *Mdn* = Median; *Min* = Minimum; *Max* = Maximum

[a]Test-retest interval in weeks from the first test administration to the respective test administration

sample size, indicating that a large number of small samples was included in the present meta-analysis. Note that test-retest interval was highest for the third test administration. This is because many studies included in this meta-analyses were intervention studies of which control groups were coded. In such studies, the third test administration represents follow-up measurements that are most likely to be conducted a few months after post-measurements. With regard to the small number of samples with more than seven test administrations, the analyses were carried out only for the first seven test administrations.

Sample characteristics Sample characteristics were calculated based on the coded statistics for all samples with the respective number of, or a larger number of test administrations. Means and standard deviations of dropout rate, percentage of male participants, and mean age of participants, as well as the numbers of control, training or other samples can be found in Table 3. The majority of samples for less than four test administrations were control samples or other samples, such as samples in methodological studies examining test-retest reliability. For more than three test administrations, training samples composed a relatively large portion of all the samples included.

Test characteristics Table 4 contains the numbers of outcomes for the variables test form, test modality and paradigm for each test administration. For a large number of studies, test form could not be coded because the information was not given in the study. For the remaining samples, alternate forms were reported more often than identical forms for two and three test administrations. Most of the tests used were verbal-numerical tests. The most common test paradigm were the N-back and memory updating paradigms, followed by complex span. In samples with more than three test administrations, mostly N-back tasks were used.

Retest effect sizes

Results up to the third test administration that will be reported in the following were controlled for test-retest interval. As mentioned above, eligible studies showed a long test-retest interval between the first and third test

Table 3 Sample characteristics

		No. of test administrations					
Variable	Type of statistic	2	3	4	5	6	7
Mean age	*M*	33.01	39.42	46.28	46.14	33.81	33.81
	SD	16.13	19.43	18.18	17.86	10.20	10.20
Gender (in % male)	*M*	54.91	38.82	37.74	36.71	41.11	41.11
	SD	29.71	16.68	13.49	12.38	15.66	15.66
Dropouts (in %)	*M*	11.82	14.31	14.08	16.53	0	0
	SD	13.93	13.05	11.46	10.64	0	0
Sample type	Control count	57	13	6	4	1	1
	Training count	4	4	3	3	3	3
	Other count	34	12	7	5	2	2

Note. All statistics were calculated on the level of samples. All means and standard deviations are weighted by the number of participants. Age was coded for the first test administration. Gender was coded as the percentage of male participants excluding dropouts

Table 4 Number of outcomes for each test characteristic per number of test administrations

	No. of test administrations					
Variable	2	3	4	5	6	7
Test form						
identical	59	13	12	6	2	2
alternate	63	24	8	6	5	5
NA	112	29	7	10	3	3
Test modality						
verbal-numerical	182	53	19	17	7	7
visuospatial-figural	43	9	5	2	0	0
both	7	3	3	3	3	3
Paradigm						
N-back	68	24	13	11	8	8
Complex span	45	12	0	0	0	0
Memory updating	69	10	10	9	0	0
Running span	11	5	0	0	0	0
Transformation span	30	9	3	2	2	2
Coordination	10	5	1	0	0	0
Other	1	1	0	0	0	0

Note. NA = not available from the study

administration and the meta-analytic effect sizes based on the eligible studies were therefore not comparable between administrations. Further, for the first three test administrations, test-retest interval had a significant influence on retest effects as will be elaborated below. Moreover, test-retest interval was observed to be highly associated with other moderators, which stresses the necessity to control for this moderator when conducting further analyses. For this purpose, for all analyses up to the third test administration, test-retest interval was set to 7.65 weeks, which was the mean interval between first and second test administration. For analyses concerning fourth to seventh test administrations, test-retest interval was not a significant moderator and test-retest intervals were comparably long to those between earlier administrations, which is why in these analyses, it was not controlled for.

The 234 effect sizes obtained for the retest effects between the first and second test administration ranged between $g = -0.47$ and $g = 1.22$, with a mean of 0.31 and a standard deviation of 0.30. Applying the multilevel model yielded a weighted average of $g = 0.28$ and an overall standard deviation of $\sqrt{\tau^2_{between} + \tau^2_{within}} = \sqrt{0.009 + 0.002} = 0.11$, with test-retest interval set to 7.65 weeks. This result can be understood as a first confirmation of the hypothesis that there are retest effects in working memory capacity tests (Hypothesis 1). For a summary of the results, including the 95% confidence interval and the within- and between-samples variances, see Table 5.

Table 5 Retest effects between the first and second test administration

g	SE	95% CI	p	$\tau^2_{between}$	τ^2_{within}	τ
0.284	0.023	[0.238, 0.329]	<.001	0.009	0.002	0.105

Note. g = Hedges' g for a test-retest interval of 7.65 weeks; SE = standard error; CI = confidence interval; p = p-value indicating whether the effect size differs from zero; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of g, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$

The same multilevel model was applied for retest effects between the first and later test administrations (see Table 6). For all test administrations, effect sizes were significantly above zero, and they were generally larger for later test administrations. Hence, results support Hypothesis 1 suggesting retest effects in working memory capacity tests also with regard to further test repetitions.

To examine if retest effects reach a plateau after a few test administrations, effect sizes between consecutive test administrations were analyzed in the same way as the effect sizes between the first and later test administrations. A summary of all results can be found in Table 7, confirming the expectation that an increase in test scores can be observed beyond the second test administration, but that retest effects reach a plateau after a few test administrations (Hypothesis 2). Results suggest that retest effects diminish after the fourth test administration. Note that differences between results for consecutive and those with reference to the first test are not directly comparable, because analyses are based on different outcomes. For example, the difference between the meta-analytic effect sizes of g = 0.28 for first to second test administration and g = 0.51 for the first to third test administration of 0.51 - 0.28 = 0.23 does not correspond to g = 0.18 for second to third administration, although test-retest interval is held constant. This is because when comparing the first to second test, the analysis is based on 234 outcomes, whereas both of the other analyses are based on 66 outcomes only.

Moderator analyses

Multilevel analyses were applied individually for each moderator. Table 8 gives an overview of the results for moderating effects between the first and second test administration. As expected (Hypothesis 3a), the length of the test-retest interval (in weeks) was found to impact significantly on the size of retest effects ($\beta_{TRinterval} = -0.005, p < .01$). As explained above, for further moderator analyses, test-retest interval was controlled for.

Retest effects were found slightly smaller for alternate ($g_{alternate} = 0.30$) compared to identical ($g_{identical} = 0.32$) test forms. Though, this differences was not significant (p

Table 6 Retest effects between the first and a later test administration

Administration	*k (s)*	*g*	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
3^{rd}	66 (29)	0.515	0.045	[0.427; 0.603]	<.001	0.001	0.018	0.135
4^{th}	27 (16)	0.809	0.092	[0.628,0.989]	<.001	0.066	0.000	0.256
5^{th}	22 (12)	0.967	0.135	[0.703,1.232]	<.001	0.132	0.000	0.363
6^{th}	10 (6)	0.930	0.190	[0.557,1.302]	<.001	0.098	0.000	0.313
7^{th}	10 (6)	1.013	0.206	[0.609, 1.416]	<.001	0.132	0.000	0.363

Note. g = Hedges' g (for first to third test administration, results are reported for a test-retest interval of 7.65 weeks); k = number of outcomes; s = number of samples; *SE* = standard error; CI = confidence interval; p = p-value indicating whether the effect sizes differ from zero; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of g, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$

Table 7 Retest effects between consecutive test administrations

Administration	*k (s)*	*g*	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
2^{nd} to 3^{rd}	66 (29)	0.178	0.038	[0.040, 0.191]	<.001	0.000	0.000	0.000
3^{rd} to 4^{th}	27 (16)	0.206	0.058	[0.093, 0.319]	.001	0.000	0.000	0.000
4^{th} to 5^{th}	22 (12)	0.099	0.060	[−0.018, 0.216]	.096	0.000	0.000	0.000
5^{th} to 6^{th}	10 (6)	0.070	0.128	[−0.182, 0.321]	.588	0.000	0.000	0.000
6^{th} to 7^{th}	10 (6)	0.093	0.128	[−0.159, 0.345]	.468	0.000	0.000	0.000

Note. g = Hedges' g (for second to third and third to fourth test administration, results are reported for a test-retest interval of 7.65 weeks); k = number of outcomes; s = number of samples; *SE* = standard error; CI = confidence interval; p = p-value indicating whether the effect sizes differ from zero, for second to third and third to fourth test administration, p-values indicate if intercepts from moderator analysis with test-retest interval differ from zero; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of g, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$

Table 8 Moderator analyses: first to second test administration

Variable	*k*	Coefficient	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
TR interval						0.009	0.002	0.105
Intercept	234	0.322	0.028	[0.268, 0.377]				
Slope		-0.005	0.002	[−0.008, −0.002]	$< .001^a$			
Test form						0.011	0.000	0.107
identical	59	0.315	0.054	[0.209, 0.421]				
alternate	63	0.301	0.051	[0.201, 0.401]	$.428^a$			
Test modality						0.011	0.000	0.104
verbal-num.	183	0.260	0.027	[0.208, 0.313]				
visuospatial	43	0.331	0.042	[0.249, 0.414]	$.061^a$			
Age						0.007	0.004	0.101
Intercept	214	0.354	0.054	[0.246, 0.463]				
Slope		-0.001	0.001	[−0.004, 0.001]	$.152^a$			
Year						0.006	0.002	0.090
Intercept	234	14.432	5.600	[3.463, 25.401]				
Slope		−0.007	0.003	[−0.013, −0.002]	.012			

Note. Coefficient = Hedges' g for categorical moderators (for test form and test modality, g is given for a test-retest interval of 7.65 weeks), meta-regression β-weights for metric moderators (controlled for test-retest interval); k = number of outcomes; *SE* = standard error; CI = confidence interval; p = p-value for factors refer to the differences between effect sizes for the respective factor levels; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of g, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$; TR interval = test-retest interval in weeks; verbal-num. = verbal-numerical; Year = Publication year; ^{a} one-sided p-value

Table 9 Moderator analyses: first to third test administration

Variable	*k*	Coefficient	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
TR interval						0.001	0.018	0.135
Intercept	66	0.539	0.050	[0.441, 0.637]				
Slope		−0.003	0.002	[−0.006, −0.000]	.015[a]			
Test modality						0.007	0.018	0.157
verbal-num.	53	0.494	0.063	[0.370, 0.617]				
visuospatial	9	0.543	0.085	[0.376, 0.617]	.317[a]			
Age						0.007	0.019	0.162
Intercept	61	0.637	0.010	[0.450, 0.824]				
Slope		−0.003	0.003	[−0.008, 0.002]	.107[a]			

Note. Coefficient = Hedges' *g* for categorical moderators (for test form and test modality, *g* is given for a test-retest interval of 7.65 weeks), meta-regression β-weights for metric moderators (controlled for test-retest interval); *k* = number of outcomes; *SE* = standard error; CI = confidence interval; *p* = *p*-values for factors refer to the differences between effect sizes for the respective factor levels; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of *g*, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$; TR interval = test-retest interval in weeks, verbal-num. = verbal-numerical; [a] one-sided *p*-value

Table 10 Retest effects for the different task paradigms: first to second test administration

Paradigm	*k*	*g*	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
N-back	68	0.261	0.036	[0.190, 0.332]	<.001	0.005	0.002	0.084
Complex span	45	0.282	0.043	[0.197, 0.367]	<.001			
Memory updating	70	0.397	0.045	[0.309, 0.486]	<.001			
Running span	11	0.236	0.112	[0.018, 0.455]	.013			
Transformation span	30	0.124	0.055	[0.016, 0.232]	.004			
Coordination	10	0.410	0.092	[0.231, 0.590]	<.001			

Note. g = Hedges' *g* for a test-retest interval of 7.65 weeks; *k* = number of outcomes; *SE* = standard error; CI = confidence interval; *p* = *p*-value indicating whether the effect sizes differ from zero; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of *g*, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$

Table 11 Retest effects for the different task paradigms: first to third test administration

Paradigm	*k*	*g*	*SE*	95% CI	*p*	$\tau^2_{between}$	τ^2_{within}	τ
N-back	24	0.511	0.074	[0.366, 0.657]	<.001	0.000	0.019	0.138
Complex span	12	0.424	0.077	[0.273, 0.576]	<.001			
Memory updating	10	0.781	0.116	[0.553, 1.008]	<.001			
Running span	5	0.302	0.216	[-0.121, 0.725]	.148			
Transformation span	9	0.302	0.133	[0.041, 0.563]	.020			
Coordination	5	0.718	0.180	[0.365, 1.070]	<.001			

Note. g = Hedges' *g* for a test-retest interval of 7.65 weeks; *k* = number of outcomes; *SE* = standard error, CI = confidence interval; *p* = *p*-value indicating whether the effect sizes differ from zero; $\tau^2_{between}$ = between-sample variance; τ^2_{within} = within-sample variance; τ = overall standard deviation of *g*, as $\sqrt{\tau^2_{between} + \tau^2_{within}}$

Table 12 Associations between the moderators

	(1)	(2)	(3)	(4)	(5)
(1) TR interval		−.04	.22	−.18	.23
(2) Test form	(.27)		.19	−.11	.65
(3) Test modality	(−.11)	(.09)		−.14	.22
(4) Age	(.32)	(−.01)	(−.14)		−.08
(5) Year	(.26)	(−.43)	(−.00)	(.09)	(.14)
(6) Paradigm	.02	.79	.29	.01	.14
	(.25)	(.59)	(.36)	(.16)	(.03)

Note. Values above the diagonal represent correlations between the respecting variables for all samples with at least two test administrations. Below the diagonal, correlations for the samples with three or more test administrations are displayed in brackets. In the last two rows, η and Cramer's V values, and, in brackets, for samples with three or more test administrations are displayed

TR interval = test-retest interval in weeks, Year = year of publication

= 0.43). Test modality did not have a significant influence on the size of the retest effect, as visuospatial contents ($g_{visuospatial} = 0.33$) showed only slightly larger retest effects than verbal-numerical contents ($g_{verbal-numerical} = 0.26$), but this difference was not significant ($p = 0.06$). Age had a slight negative but non-significant influence on the size of the retest effect ($\beta_{age} = -0.001$; $p = 0.15$). Thus, Hypotheses 3b, 3c, and 3d were not supported for retest effects from the first to second test administration.

Moderator analyses for later test administrations yielded similar patterns (see Table 9). For test-retest interval, participant age, and test modality, moderator analyses were carried out up to the third test administration. For later test administrations, either the distribution (for continuous variables) or the number of samples per factor level (for factors) did not allow to carry out meaningful moderator analyses. Due to the small number of identical test forms in samples with more than two test administrations (cf. Table 4), no further moderator analyses were carried out for this variable.

Test-retest interval (in weeks) had a significant impact on the size of retest effects from the first to the third test administration ($\beta_{TRinterval} = -0.003$, $p = 0.02$), again speaking in favor of Hypothesis 3a. Also from the first to the third test administration, retest effects were not significantly larger for visuospatial tests ($g_{visuospatial} = 0.54$) than for verbal-numerical tests ($g_{verbal-numerical} = 0.49$; $p = .32$), speaking against Hypothesis 3c. With respect to participant age, retest effects from the first to the third test administration slightly decreased with age ($\beta_{age} = -0.003$), but the effect was not significant ($p = .11$). Thus, results fail to support Hypothesis 3d.

Differences between paradigms

Subsequently, it was analyzed whether retest effects differ between the test paradigms (Table 10). For all paradigms, after controlling for test-retest interval, effect sizes were significantly positive. Comparing the paradigms to each other, retest effects in transformation span tasks from first to second test were significantly smaller than those for N-back ($p < .01$), complex span ($p = .04$), updating ($p < .01$), and coordination tasks ($p = .01$). Further, retest effects in updating tasks were significantly larger than retest effects in N-back tasks ($p = .02$).

From first to third test, larger retest effects were observed in memory updating tasks compared to complex span ($p = .01$) and transformation span tasks ($p = .01$). All other effect sizes were not significantly different (Table 11).

Further explorative analyses

Explorative analyses yielded that year of publication was negatively associated with the size of retest effects for first to second test administration ($\beta_{year} = -0.007$, $p = .01$), after controlling for test-retest interval. All other coded variables were not significantly associated with the size of retest effects.

Collinearity

All of the theoretically assumed moderators and the ones which were significant in explorative analyses were examined for collinearity. Coefficients that correspond to the scale level of moderators were calculated (Table 12). As mentioned above, test-retest interval was observed to be associated with most of the variables, which emphasizes that it was necessary to control for this moderator when analyzing additional ones. Also, paradigms were strongly associated with test form. For example, most outcomes from N-back, complex span, running span and coordination tasks were alternate test forms, whereas in updating tasks, mostly identical test forms were used for retests.

Publication bias

Funnel plots, indicating publication bias by plotting effect sizes against their standard errors, were inspected for publication bias. A funnel plot for the main analysis comparing first to second test administrations is shown in Fig. 1. For analyses comparing first to later test administrations, funnel plots are depicted in Fig. 2, and those for analyses comparing consecutive test administrations are shown in Fig. 3. The funnel plots do not show signs of publication bias, as they appear mostly symmetric.

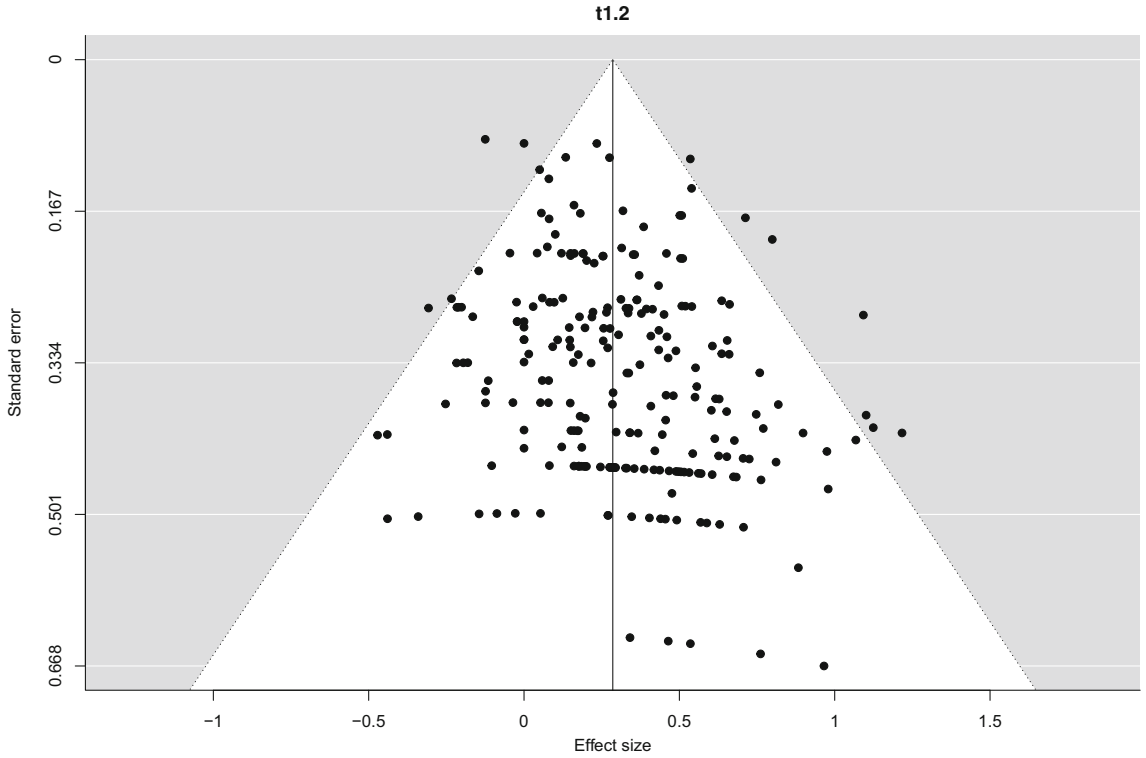

Fig. 1 Funnel plot for retest effects between the first and second test administration

Discussion

Interpretation of results and theoretical implications

It was the goal of the present meta-analysis to show that the commonly known phenomenon of retest effects in cognitive ability tests is reproducible for working memory capacity tests. Further, effect sizes resulting from outcomes for multiple test administrations were investigated. Finally, possible moderators of retest effects for multiple test administrations of working memory capacity tests were examined. Concluding from the results of main and moderator analysis, retest effects were shown to be replicable in working memory capacity tests. Scores increased up to the fourth test administration until they reached a plateau, where, with a test-retest interval of 7.65 weeks, a large average increase of three quarters of a standard deviation was attained. Test-retest interval had a significant impact on the size of the retest effect and differences in retest effects between paradigms of working memory capacity tests were found. Interestingly, year of publication had a significant influence on retest effects as well. However, age did not influence the effect significantly. It is especially noteworthy that the use of alternate test forms did not reduce the effect size.

Retest effects Results of the present multilevel meta-analysis show that participants score on average slightly more than a quarter of a standard deviation, or 4.26 IQ points higher, when taking a working memory capacity test for the second time after a test-retest interval of 7.65 weeks (Note that for the ease of interpretation, score gains in working memory capacity tests are transformed from Hedges' g to the IQ scale, although this might not correspond to its initial assignment). A prediction of retest scores based on this main result for the example of an initial score of 100 IQ points would lead, on average, to a score of 104.26 IQ points when retesting for the first time. Although this is regarded as a small effect following Cohen's conventions (Cohen, 1988), it is noteworthy that this effect results from a simple repetition of a test, which is much less effortful than, e.g., an intervention program and should thus be taken seriously. The result is consistent with prior meta-analytic research, which found retest effects of a similar magnitude in cognitive ability tests (Hausknecht et al., 2007; Scharfen et al., 2018) and common neuropsychological measures (Calamia et al., 2012). The observed retest effects can be explained by three causes, summarized by Lievens et al. (2007), which can be assumed to apply to working memory capacity tests as one special type of cognitive ability tests. Taking into account the overall standard deviation of the true effect from the multilevel model ($\tau = 0.11$), one can assume that 95% of the true

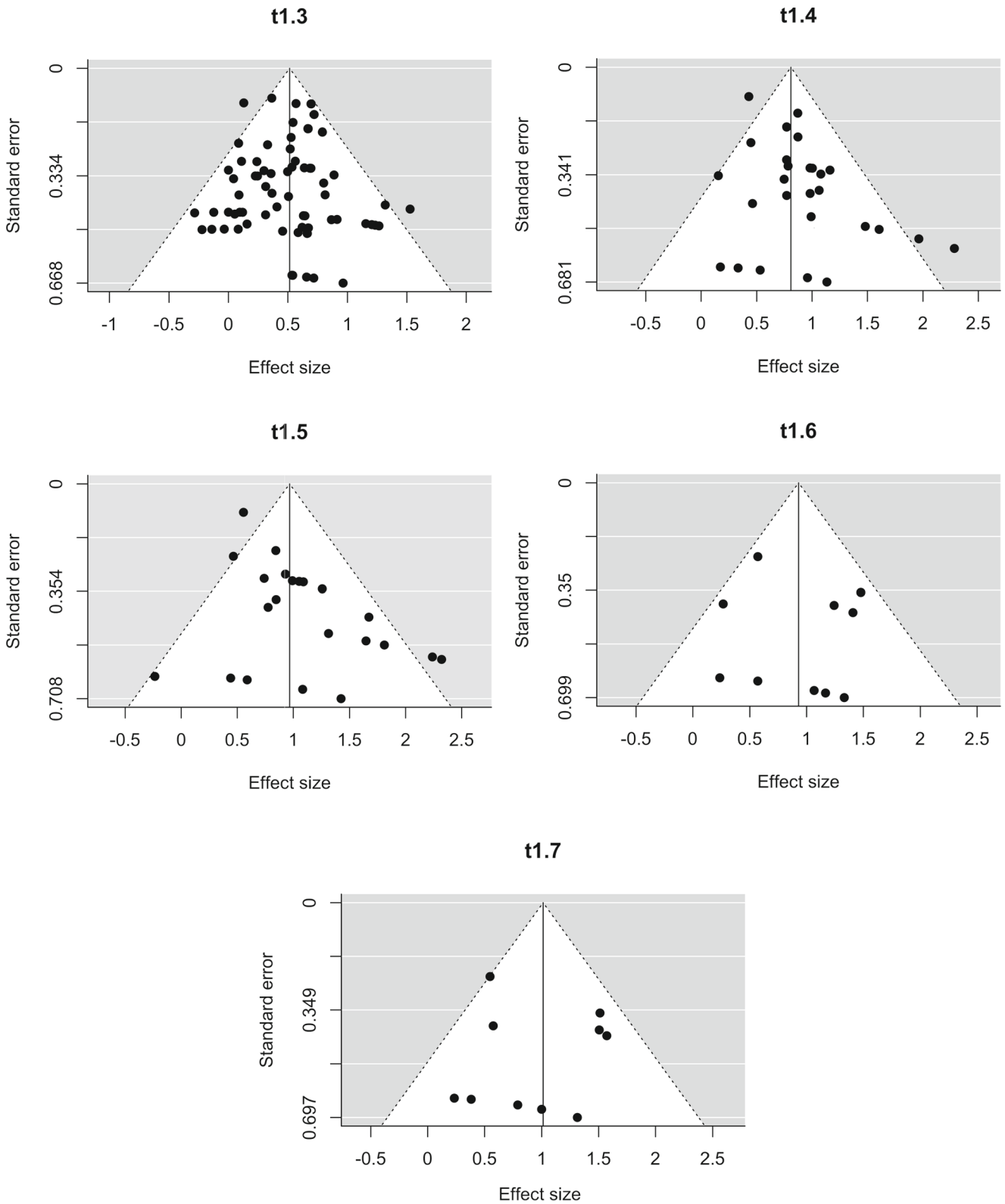

Fig. 2 Funnel plots for retest effects between the first and respective test administrations

values lie within a prediction interval from 0.07 to 0.50. Moderator and explorative analyses were able to reduce the variance and explain differences between outcomes partly.

For further test administrations, retest effects were found to increase until the fourth test administration. On average, participants score slightly more than half a

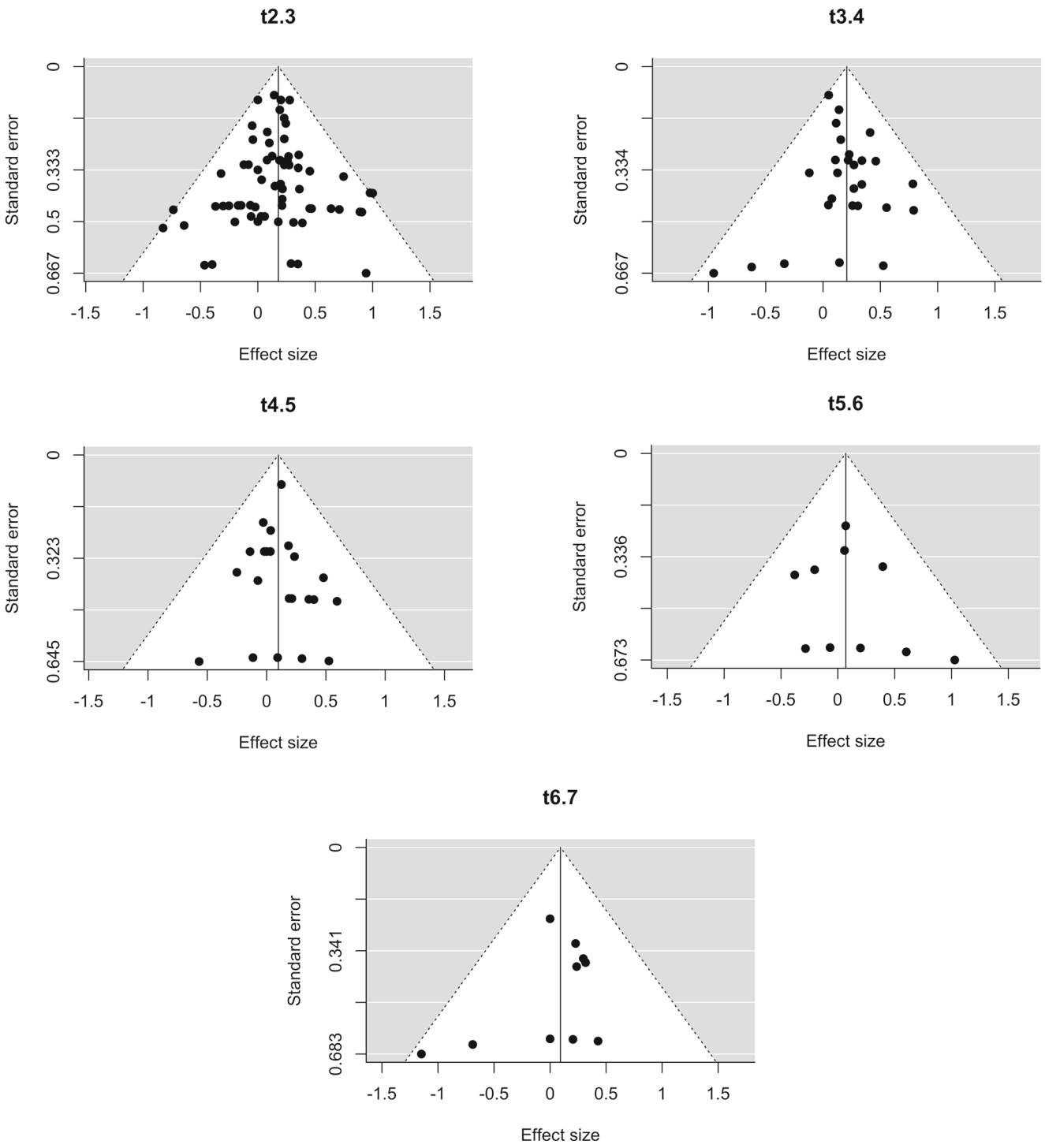

Fig. 3 Funnel plots for retest effects between consecutive test administrations

standard deviation or 7.73 IQ points higher in a third test administration and more than three quarters of a standard deviation or 12.14 IQ points higher in a fourth test administration, relative to the first test administration. These effects correspond to medium to large effect sizes according to Cohen (1988). Note, however, that results with reference to more than four test administrations have to be interpreted carefully as less than ten studies were observed. Although retest effects relative to the first test administration increased up to the seventh test administration, analyses of retest effects between consecutive test administrations suggested that this increase is no longer significant after the fourth test administration. To illustrate, based on these results, for a person with an initial score of 100 IQ points,

it would be predicted that retesting with a working memory capacity test four times would lead to a score of 112.14 IQ points. After four tests, retest effects seem to increase no further. This finding can be explained by a decrease of the influence of the causes of retest effects with the number of test repetitions (Lievens et al., 2007; Scharfen et al., 2018). The factors that lead to an initial increase in retest scores might play a smaller role in later administrations. For example, test anxiety might be reduced during the initial test administrations and then, no further reduction of this factor might take place, leading to no further increase in retest effects when retesting multiple times. Comparing these results to those from meta-analytic evidence from Scharfen et al. (2018), who focused on a wide range of cognitive ability tests, a plateau seems to be reached a little bit later in working memory compared to other cognitive ability tests. This might be due to the complexity of working memory tasks. It has been argued that more complex tasks lead to larger retest effects, because more test-specific strategies can be developed compared to easier tasks for which strategies do not apply (Scharfen et al., 2018). A more complex task, such as a 3-back task, might allow for more strategies to apply compared to a simple reaction time task. Also, construct-irrelevant factors might be reduced to a higher degree, as, e.g., test anxiety can be very high in a first working memory task but then be reduced over further tests, leading to increased retest scores. In easy tasks, which might not threaten a participant as much as complex tasks, less reduction of text anxiety can take place. As working memory capacity tasks can be considered rather complex, this might be one reason why a plateau is reached later compared to all kinds of more or less complex cognitive ability tasks.

Moderator analysis and explorative analysis Moderator analyses for retest effects from the first to the second and from the first to the third test administration suggest that the length of the test-retest interval is critical for the size of retest effects. The finding that the length of the test-retest interval is a moderator of the size of retest effects is consistent with the assumption that participants form memory traces of test-related contents or procedures. As these decay over time (e.g., Arendasy & Sommer, 2017), smaller retest effects result from a longer test-retest interval. This finding is also in line with prior research (Calamia et al., 2012; Salthouse et al., 2004; Hausknecht et al., 2007; Scharfen et al., 2018). Illustrating this result, a person who has scored 100 IQ points on a first working memory capacity test, would be predicted to score 104.83 IQ points the second time if they take the second test directly after the first (corresponding to $\beta(intercept)_{TRinterval} = 0.32$), provided all other conditions are held constant. If they would instead take the second test ten weeks later, a second score of 102.55 IQ points would be predicted. A second test after half a year would lead to an average score of 101.43 IQ points. A full elimination of retest effects would be expected after on average $0.32/0.005 = 64$ weeks, thus, approximately sixteen months, which is a relatively short delay compared to the findings from Salthouse et al. (2004) and Calamia et al. (2012). Note that this value applies for an intercept of 0.32 that averages all kinds of test paradigms. There are differences between paradigms concerning the time needed to eliminate retest effects, as these are significant moderators of the overall effect as well. For example, for coordination tasks, 21 months, and, in contrast, for transformation tasks, 6 months would be necessary to prevent retest effects.

None of the other theoretically expected moderators had a significant impact on the size of retest effects. Surprisingly, alternate test forms did not yield significantly smaller retest effects than identical test forms in the present meta-analysis. This runs counter to what would be expected in terms of the causes of retest effects (Lievens et al., 2007), and prior research findings (e.g., Calamia et al., 2012; Hausknecht et al., 2007). On the other hand, it is a common research finding that retest effects in alternate test forms which only differ in non-salient surface features are comparable to retest effects in identical test forms (Arendasy & Sommer, 2013; Matton et al., 2011; Morley, Bridgeman, & Lawless, 2004). In the present meta-analysis, the use of alternate test forms was often observed in studies using N-back tasks and complex span tasks. For both task paradigms, it seems reasonable to assume that alternate test forms were constructed using the same set of stimuli (e.g., digits, letters, or dot locations in a matrix), but varying their composition and order. Moreover, it has been suggested that the development of test-taking strategies and the accommodation to novel task demands plays a larger role than memory effects in complex tasks (Basso, Bornstein, & Lang, 1999; Beglinger et al., 2005; Thorgusen, Suchy, Chelune, & Baucom, 2016), and both N-back tasks and complex span tasks can be regarded as attentionally demanding. Note, however, that in a lot of studies it was not reported whether identical or alternate test forms were used. Nonetheless, the relatively robust finding that test form equivalence serves as a moderator of retest effects from the area of cognitive ability tests (Arendasy & Sommer, 2017; Freund & Holling, 2011) could not be shown to be reproducible for working memory capacity tests, which might be due to working memory task characteristics.

Concerning the modality of the test, retest effects were equally large in visuospatial compared to verbal-numerical tests for all test administrations. This finding contradicts evidence from Benedict and Zgaljardic (1998) and Salthouse and Tucker-Drob (2008), suggesting smaller effects in verbal-numerical tests because of higher familiarity. Indeed, it was assumed that verbal-numerical test contents would be more familiar to participants than visuospatial test

contents. However, neither Benedict and Zgaljardic (1998) nor Salthouse and Tucker-Drob (2008) nor Scharfen et al. (2018), who did find an effect of task modality, analyzed working memory capacity tests. The moderating effect of task modality might thus not hold for working memory capcity tests. Familiarity of task content might thus play a minor role in working memory retesting compared to other cognitive ability tests. Again, it might be relevant that working memory tasks can be considered very complex when compared to all kinds of cognitive ability tests.

When analyzing the effect of age on retest effects, tendencies were found towards the expected direction. Both from the first to the second and from the first to the third test administration, retest effects were found to decrease with age, although insignificantly. While 20-year-old participants would be expected to show retest effects of on average 0.30 standard deviations when retested after 7.65 weeks, retest effects in 65-year-old participants would be predicted to be about 0.24 standard deviations, given that all other conditions are equal. For retest effects from the first to the third test administration, the impact of age was found to be equally large. However, the moderating effect of age on retest effects was not significant and this moderator did not contribute to a reduction of the standard deviation of the true effect ($\tau = 0.10$). It is thus questionable if a prediction of retest effects based on age is efficient. The missing moderating effect of age on retest effects is in line with some prior studies (Bartels et al., 2010; de Oliveira, Trezza, Busse, & Filho, 2014), whereas it contradicts other evidence that did find a lowering effect of age on retest effects (Calamia et al., 2012; Scharfen et al., 2018; Schleicher, Iddekinge, Morgeson, & Campion, 2010; Iddekinge, Morgeson, Schleicher, & Campion, 2011). Whereas Scharfen et al. (2018), Schleicher et al. (2010), and Iddekinge et al. (2011) investigated other cognitive ability tasks, working memory tasks were one of many tasks evaluated by Bartels et al. (2010), de Oliveira et al. (2014), and Calamia et al. (2012). In all of these studies, effects of age on retest effects were thus averaged over several kinds of cognitive ability tasks, and working memory has not been examined individually. This speaks for the assumption that age might not be of as much relevance in working memory retesting as it is when explaining retest effects in other cognitive ability domains. One explanation for this finding is that differences in the ability to represent, maintain and update context information (Braver & Barch, 2002) might not exclusively be explained by age differences in the observed sample. At least within the restricted range of 12 to 70 years, age does not seem to differentiate between participants regarding this ability. It has to be noted that in this meta-analysis, overall participant age might not be representative for the population, as mainly younger, and a few or older samples were observed and only very few samples between 40 and 60 years were included in the analysis. In former meta-analyses, age might have been more representative and results from this analysis might thus be restricted in this regard.

The sizes of retest effects for the different task paradigms investigated in the present meta-analysis were quite homogeneous, with a few exceptions: Tasks categorized as belonging to the transformation span paradigm, such as digit span backward and letter number sequencing tasks, showed smaller retest effects than most of the other paradigms for first to second and first to third test. Transformation span tasks were mostly digit span backward tasks and letter number sequencing tasks, which both mostly stem from different versions of the WAIS. The WAIS is one of the most commonly administered intelligence tests and participants might thus have been familiarized with this kind of task more compared to other working memory paradigms that they have never been confronted with before. A higher familiarity with the task that had been developed before the eligible study took place might be responsible for the lower retest effect in transformation span. When comparing first to third test administration, memory updating tasks showed significantly larger effects than complex and transformation span tasks. Note that for memory updating tasks, mostly identical test forms were administered in a retest, whereas in other paradigms, alternate test forms were more common. This might have led to an overestimation of the retest effect in memory updating tasks. However, outcomes categorized as memory updating tasks often stemmed from the Paced Auditory Serial Addition Test (PASAT). In his review, Tombaugh (2006) states that the PASAT is extremely prone to retest effects. Tombaugh (2006) argues that the PASAT shows large retest effects because of its high complexity and the possibilities to develop solving strategies which result from this complexity.

Further explorative analyses imply that retest effects for first to second test administration were smaller for studies with a later year of publication. Per decade, the effect from first to second administration decreased on average by 0.07 standard deviations (after controlling for test-retest interval), which can be considered a small effect. Ideally, the effect of year of publication could be explained by a higher awareness of the occurrence of retest effects and the development of tests which are more resistant to them. Also, a higher familiarity with cognitive ability tests that evolved over the last decades could explain retest effects becoming smaller over the years.

Limitations

In a few cases, these results have to be interpreted cautiously due to limitations, which will now be discussed.

Firstly, the fact that only a small number of samples with more than four test administrations was obtained

for inclusion in the present meta-analysis is critical for the interpretation of retest effects for more than four test administrations. Less than ten studies were observed and thus results have to be interpreted carefully.

The limited number of studies had consequences for the investigation of moderators: Analyses for most of the moderators could only be carried out for up to three test administrations (and, concerning the use of identical or alternate test forms, were even restricted to two administrations), as the distribution of moderator characteristics did not allow for further investigations.

Also, concerning the analysis of potential moderators of retest effects, it should further be noted that the moderators examined represent a selection of variables on which information could be obtained from the studies included in the analyses. In order to achieve a satisfactory explanation of retest effects, future studies and meta-analyses should consider further moderators, as for example general mental ability. In this regard, it will be of relevance to evaluate the relationship between baseline performance and the size of the retest effect, which is only possible if raw data is available from a sufficient number of primary studies. Further moderators might be level of education, motivation or health status (see Randall and Villado (2017), for an overview of further possibly relevant moderators). Although the moderating effect of health status on retest effects has already been investigated for a selection of working memory capacity tests (Calamia et al., 2012), a reconsideration and expansion to a broader range of paradigms seems appropriate in the light of the diverging results between them in healthy samples. Furthermore, Schmidt et al. (2013) suggest to correct effect sizes for reliability of the test. In this meta-analysis, we did not conduct this correction, because Cronbach's α was given for five outcomes only. Sensitivity analyses were conducted to observe differences in the main results that would have resulted from unreliability corrections. Reliabilities were set to 0.70, 0.80, and 0.90 and led to effect sizes of $g_{1.2} = [0.29, 0.34]$, $g_{1.3} = [0.54, 0.62]$ and $g_{1.4} = [0.85, 0.97]$. It becomes obvious that, when a correction for unreliability is applied, even higher effect sizes result. For the current analysis, it was chosen not to apply this correction because the estimation of the reliability for all tests for which it was not reported would have led to imprecise correction of effect sizes.

On a theoretical level, moderators were mainly derived from the three causes of retest effects that were summarized for cognitive ability tests in general (Lievens et al., 2007). These three categories of causes have important implications for the validity of retest scores. Importantly, our results do not allow for a conclusion about the validity of retest scores. For practitioners, it is especially relevant how retesting affects psychometric properties, as results are inconsistent with regard to whether initial or retest scores can be considered more valid (Hausknecht et al., 2007; Freund & Holling, 2011; te Nijenhuis, van Vianen, & van der Flier, 2007). The present meta-analysis is, however, limited to a summary of results on the size of retest effects in working memory capacity tests and its determinants.

Finally, a large number of samples included in the present meta-analysis was small. For example, the median sample size of studies with at least two test administrations indicated that 50% of the samples had a sample size below or equal 20. This characteristic is not considered to be a limitation concerning the present meta-analysis, as the effect size metric used includes a correction of small sample bias. Still, it raises methodological concerns about the studies which the analyses were based on, because the statistical power of an experiment depends on the sample size (Cohen, 1992). Also, in some of the smaller studies, e.g., Verhaeghen, Cerella, and Basak (2004), negative retest effects were observed. Yet, this observation is in line with those made by Redick (2015), who found control groups of working memory trainings to have negative pre to post changes.

Future research

Altogether, the results clearly indicate that retest effects cannot be ignored when a working memory capacity test is administered twice or more. Thus, research on effective and sensitive methods to account for retest effects is needed. Also, our results call for studies that examine causes and determinants of retest effects in working memory capacity tests in order to gain a better comprehension of the effect's origins.

For a substantial investigation of retest effects for multiple test administrations, a larger number of studies in which participants are assessed with a working memory capacity test more than twice would be needed. A larger body of research on the repeated administration of working memory capacity tests might also enable the examination of moderators for more than two or three test administrations, which was not possible in the present meta-analysis.

The results of the moderator analysis of retest effects for identical versus alternate test forms clearly call for further research on the use of alternate test forms in working memory capacity tests. In particular, it should be investigated whether the size of retest effects in alternate tests depends on how alternate forms are constructed.

Furthermore, it is important that future studies, in which a test is repeatedly administered, provide more information on whether identical or alternate test forms are used. If alternate test forms are used, they should include a description of how they were constructed. In the long run, this would enable the investigation of the moderating effect of the use of different kinds of alternate test forms on a meta-analytical level.

It is noteworthy that some of the moderators that have been found to moderate retest effects in a wide range of

cognitive ability tests (Calamia et al., 2012; Hausknecht et al., 2007; Scharfen et al., 2018) did not moderate working memory retest effects: Test form, age, and task modality did not have a significant influence on retest effects. Also, a plateau is being reached somewhat later compared to other cognitive ability tasks. As discussed above, this could be due to different mechanisms underlying retest effects in working memory compared to other cognitive ability tasks, or also due to the high complexity of working memory capacity tests. Generally, not much is known about the actual cognitive mechanisms that cause retest effects (Randall & Villado, 2017). Future research might thus focus on these underlying mechanisms causing the increase in test scores between two tests and compare these mechanisms between different cognitive ability tests.

When reviewing the literature, it became apparent that only seldomly causes of retest effects are directly investigated (Randall & Villado, 2017) and that results on consequences of retesting, such as validity changes, are mixed (Freund & Holling, 2011; te Nijenhuis et al., 2007). This applied irrespective of whether general cognitive ability or working memory retest effects were investigated. Future research should therefore focus on causes, determinants and consequences of retest effects in both cognitive ability and working memory capacity tests.

Practical implications

In line with the recommendations of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments (Heilbronner et al., 2010), the findings of the present meta-analysis clearly suggest that retest effects in working memory capacity tests must be taken into account when they are repeatedly administered in practical or research settings. Furthermore, the results provide, at least in part, clues on how to, and how not to, adequately address retest effects in working memory capacity tests.

Revisiting scenarios from the introduction, the results stress that practitioners and researchers must be aware that the examination of disease progression, cognitive decline or the impact of a cognitive ability training can be distorted by retest effects. When a diagnosis or treatment decision relies on the results of the repeated administration of a working memory capacity test, it must be acknowledged that the scores of the second test administration might overestimate the latent working memory capacity.

On an experimental level, the fact that retest effects in working memory capacity tests exist stresses the importance of adequate control groups, for example in research on the effects of working memory training (Green et al., 2014; Melby-Lervåg & Hulme, 2013). The present meta-analysis only included samples with no activity between the test administrations, and thus, shows that a research design with an inactive control group can be informative when trying to account for retest effects. In fact, evaluation studies of working memory trainings can be challenged to show efficacy of their programs by the retest effects found in this analysis: If retest effects exist in a control group, comparing an intervention group to this control group might lead to the finding that intervention effects are low. The intervention group itself might also be affected by retest effects from pre to post assessment. Even if alternate test forms of working memory capacity are used, retest effects influence the sizes of the intervention effect. It might thus be very difficult to show high intervention efficacy when working memory capacity tests are used as criterion measures.

Suggestions on how to control for retest effects (Arendasy & Sommer, 2017; McCaffrey et al., 2000; Heilbronner et al., 2010; Green et al., 2014) include the use of one or two baseline evaluations prior to the actual first test administration in order to familiarize the participants with the test procedure. However, analyses of retest effects for further test administrations revealed that score gains must be expected at least up to the fourth test administration. Thus, it is questionable whether the effort of administering the test a couple of times without using the data of the initial test administrations would pay off. Freund and Holling (2011) further suggest that familiarity with the test should always be assessed in order to be aware of possible retest effects. Although this recommendation does not prevent retest effects, it though enables practitioners to be aware of their appearance. Importantly, the use of alternate test forms was not found to be an effective method to reduce the size of retest effects in the present meta-analysis. The use of alternate test form is often recommended to prevent retest effects (Arendasy & Sommer, 2017). For working memory capacity tests, however, this recommendation does not seem to apply, which has to be taken care of. Administering multiple working memory tasks at different time points might be another way to reduce retest effects, although evidence is scarce in this regard.

Another strategy to address retest effects for which this analysis found high support would be to prolong the test-retest interval. Moderator analysis revealed that this can indeed be an effective strategy, if applicable in the respective setting. Following the results from this meta-analysis, an interval of sixteen months between first and second test would be sufficient to eliminate retest effects in most settings. It is noteworthy that for test paradigms showing larger retest effects, like coordination tasks, the interval needed to eliminate retest effects can be longer. Thus, according to the results, a retest interval of at least 21 months should be applied to prevent retest effects in all kinds of working memory capacity tests when retesting only once.

Following the results, it might also be an efficient method to use transformation span tasks, because, compared to other paradigms, these were found to show lower retest effects.

Other approaches to deal with retest effects rely on statistical methods, for example reliable change indices (e.g., Chelune, Naugle, Lüders, Sedlak, & Awad, 1993; Knight, McMahon, Skeaff, & Green, 2007) or regression-based approaches (e.g., Salthouse & Tucker-Drob, 2008; Cysique et al., 2011). These approaches can be applied on a group level, for example when investigating cognitive decline. Based on the results of the present meta-analysis, an application of these methods seems justified at first sight, but as Salthouse and Tucker-Drob (2008) state, "methods to correct for retest effects can only be strongly justified, and eventually improved upon, after retest effects are fully characterized and understood" (p. 2). Regression-based approaches, for example, require a profound knowledge about which variables moderate the size of retest effects, and as outlined above, the results of the moderator and explorative analyses indicate that retest effects are a complex phenomenon which depends on a variety of different conditions, which have not yet been investigated exhaustively. The promising approach of using latent variable models has gained attention in working memory training evaluations and retest effect analyses (Freund & Holling, 2011; Guye et al., 2017; Matton et al., 2011).

Conclusions

By investigating retest effects in a comprehensive sample of working memory capacity tests for multiple test administrations, the present multilevel meta-analysis confirms and extends the knowledge obtained from prior research on retest effects and finds evidence for their reproducibility for working memory capacity tests. Retest effects in working memory capacity tests were found to be of a similar magnitude as retest effects in a broader range of cognitive ability tests. Further analyses yielded that retest effects increase for multiple test administrations, but reach a plateau after four test administrations. These findings have implications for both practical and research settings. For practitioners, it is important to be aware of retest effects in working memory capacity tests and of the finding that they are comparably high as in other cognitive ability tests. Overestimation of the latent working memory capacity can easily result from multiple retesting or even high familiarity with similar tests, as the use of alternate test forms did not reduce the effect. A diagnosis can thus not be reliable if it is based on retest scores without taking into account retest effects. Because retest effects vary enormously, especially with regard to multiple test repetitions, a general correction is difficult to deduce. This is endorsed by results from moderator analyses, which could explain some but not all of the variance of the effect. This analysis was able to reveal important moderating variables that were able to explain the effect partly, such as test-retest interval and test paradigm. Moderator analyses indicated that effects decrease with an increasing test-retest interval. According to our results, retest effects can thus be prevented by using a test-retest time interval of at least 21 months between administrations. Also, specific kinds of tests might show smaller retest effects. In any case, when a test is administered, the testee should always be asked if they are familiar with this kind of test in order to be aware of possible retest effects (Freund & Holling, 2011), and using a test-retest interval of at least 21 months, which can be considered the safest method to prohibit retest effects for practitioners. Future research should clearly address more effective methods to predict retest effects precisely and should examine causes, determinants and consequences of retesting in order to develop more reliable methods to prevent them.

Acknowledgements We thank David Darby, Susanne M. Jaeggi, Thomas W. Kaminski, Shu-Chen Li, Florian Schmiedek, Nash Unsworth, and Barbara A. Wilson for providing data on eligible studies. This work was partly supported by grant HO 1286/6-4 of the Deutsche Forschungsgemeinschaft.

Appendix

Table 13 Assignment of tests to paradigms

Paradigm	Test
N-back	0-back, 1-back, 2-back, 3-back, 4-back, 5-back, 2-6-back, N-back (figural/numerical), 1-back/2-back + processing (figural/numerical), Continuous performance (ANAM), Dual N-back (affective/neutral)
Complex span	Computation span, Counting span, Dot matrix, Operation span, Reading span, Rotation span, Symmetry span
Memory updating	Memory updating (spatial/numerical), Mental counters, PASAT, Rapid visual information processing, Self-ordered pointing test, Updating (auditory-verbal/visuospatia/dual-modality)
Running span	Letter memory running span, Number running span, Running letter span, Running memory span, Running span
Transformation span	Alpha span, Animal span, Block-tapping backward, Digit span backward, Letter number sequencing, Spatial span backward, Spatial WM backward, Word Suppression Test (neutral/affective),
Coordination	ATClab, Coordination numerical, Coordination verbal, Dot test, Groton Maze Learning Test (CogState), Spatial ST, Visual Learning (CogState), Verbal WM
Other	N-back and odd-one out (composite), Concussion Sentinel (composite)

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, *102*(1), 3–27. https://doi.org/10.1037/0033-2909.102.1.3

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: the same or different constructs? *Psychological Bulletin*, *131*(1), 30–60. https://doi.org/10.1037/0033-2909.131.1.30

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*(10), 1086–1093. https://doi.org/10.1037/0003-066X.36.10.1086

Arendasy, M. E., & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence*, *41*(3), 181–192. https://doi.org/10.1016/j.intell.2013.02.004

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence*, *62*, 89–98. https://doi.org/10.1016/j.intell.2017.03.003

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkühl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *22*(2), 366–377. https://doi.org/10.3758/s13423-014-0699-x

Au, J., Buschkühl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing evidence that working memory training is not effective: A reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin & Review*, *23*(1), 331–337. https://doi.org/10.3758/s13423-015-0967-4

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Ball, K., Edwards, J. D., & Ross, L. A. (2007). The impact of speed of processing training on cognitive and everyday functions. *Journals of Gerontology: Series B*, *62B*(I), 19–31.

Baltes, P. B., & Kliegl, R. (1992). Further testing of limits of cognitive plasticity: Negative age differences in a mnemonic skill are robust. *Developmental Psychology*, *28*(1), 121–125.

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), 118. https://doi.org/10.1186/1471-2202-11-118

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283–292. https://doi.org/10.1076/clin.13.3.283.1743

Beckmann, B., Holling, H., & Kuhn, J. T. (2007). Reliability of verbal-numerical working memory tasks. *Personality and Individual Differences*, *43*(4), 703–714. https://doi.org/10.1016/j.paid.2007.01.011

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., & Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, *20*(4), 517–529. https://doi.org/10.1016/j.acn.2004.12.003

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 339–352. https://doi.org/10.1076/jcen.20.3.339.822

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York: Russell Sage Foundation.

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, *26*(7), 809–817. https://doi.org/10.1016/S0149-7634(02)00067-2

Brunoni, A. R., & Vanderhasselt, M. A. (2014). Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: A systematic review and meta-analysis. *Brain and Cognition*, *86*, 1–9. https://doi.org/10.1016/j.bandc.2014.01.008

Buschkühl, M. (2007). *Arbeitsgedächtnistraining: Untersuchungen mit jungen und älteren Erwachsenen [Working memory trainings: Studies on younger and older adults]* (Doctoral Dissertation). University of Bern, Bern.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Cattell, R. B. (1987). *Intelligence: Its structure growth and action*. Amsterdam: North-Holland.

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hilsdale: Erlbaum.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society*, *9*(3), 419–428. https://doi.org/10.1017/S1355617703930074

Conway, A. R. A., & Kane, M. J. (2001). Capacity, control and conflict: an individual differences perspective on attentional capture. In C. L. Folk, & B. S. Gibson (Eds.), *Attraction, disctraction and action: Multiple perspectives on attentional capture* (pp. 349–372). New York: Elsevier Science. https://doi.org/10.1016/S0166-4115(01)80016-9

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cowan, N. (1999). An embedded-processes model of working memory. In A. E. Miyake, & P. E. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix, & H. Hagendorf (Eds.), *Human memory and cognitive capabilities* (pp. 409–422). Amsterdam: Elsevier Science.

Cysique, L. A., Franklin, D., Abramson, I., Ellis, R. J., Letendre, S., Collier, A., ... HNRC group (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology*, *33*(5), 505–522. https://doi.org/10.1080/13803395.2010.535504

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

de Oliveira, R. S., Trezza, B. M., Busse, A. L., & Filho, W. J. (2014). Learning effect of computerized cognitive tests in older adults. *Einstein (São Paulo)*, *12*(2), 149–153. https://doi.org/10.1590/s1679-45082014ao2954

D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Experimental Brain Research*, *133*(1), 3–11. https://doi.org/10.1007/s002210000395

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? *Memory & Cognition*, *42*(6), 854–862. https://doi.org/10.3758/s13421-014-0410-5

Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger III (Ed.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, DC: American Psychological Association. https://doi.org/10.1037/10394-016

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*(4), 233–243. https://doi.org/10.1016/j.intell.2011.02.009

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*(1), 136–158. https://doi.org/10.1016/j.jml.2004.03.008

Glisky, E. L. (2007). Brain aging: Models, methods, and mechanisms. In D. Riddle (Ed.), *Changes in cognitive function in human aging* (pp. 3–21). Boca Raton: CRC Press.

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Assessment & Disease Monitoring*, *1*(1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

González, H. M., Tarraf, W., Bowen, M. E., Johnson-Jennings, M. D., & Fisher, G. G. (2013). What do parents have to do with my cognitive reserve? Life course perspectives on twelve-year cognitive decline. *Neuroepidemiology*, *41*(2), 101–109. https://doi.org/10.1159/000350723

Green, C. S., Strobach, T., & Schubert, T. (2014). On methodological standards in training and transfer experiments. *Psychological Research*, *78*(6), 756–772. https://doi.org/10.1007/s00426-013-0535-3

Guye, S., Simoni, C. D., & von Bastian, C. C. (2017). Do individual differences predict change in cognitive training performance? A latent growth curve modeling approach. *Journal of Cognitive Enhancement*, *1*(4), 374–393. https://doi.org/10.1007/s41465-017-0049-9

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*(8), 1267–1278. https://doi.org/10.1080/13854046.2010.526785

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research finding* (2nd ed.). Thousand Oaks: Sage Publications.

Iddekinge, C. H. V., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*, *96*(5), 941–955. https://doi.org/10.1037/a0023562

Jaeggi, S. M., Buschkühl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport: Praeger Publishers/Greenwood Publishing Group.

Jolles, D. D., Grol, M. J., van Buchem, M. A., Rombouts, S. A., & Crone, E. A. (2010). Practice effects in the brain: Changes in cerebral activation after working memory practice depend on task demands. *NeuroImage*, *52*(2), 658–668. https://doi.org/10.1016/j.neuroimage.2010.04.028

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671. https://doi.org/10.3758/BF03196323

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science*, *25*(11), 2027–2037. https://doi.org/10.1177/0956797614548725

Karch, D., Albers, L., Renner, G., Lichtenauer, N., & von Kries, R. (2013). The efficacy of cognitive training programs in children and adolescents. *Deutsches Ärzteblatt International*, *110*(39), 543–652.

Kelly, M. E., Loughrey, D., Lawlor, B. A., Robertson, I. H., Walsh, C., & Brennan, S. (2014). The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: A systematic review and meta-analysis. *Ageing Research Reviews*, *15*, 28–43. https://doi.org/10.1016/j.arr.2014.02.004

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Kliegl, R., & Baltes, P. B. (1987). Theory-guided analysis of mechanisms of development and aging through testing-the-limits and research on expertise. In C. Schooler (Ed.), *Cognitive functioning and social structure over the life course* (p. 95–119). Norwood: Ablex Publishing.

Kliegl, R., Smith, J., & Baltes, P. B. (1989). Testing-the-limits and the study of adult age differences in cognitive plasticity of a mnemonic skill. *Developmental Psychology*, *25*(2), 247–256. https://doi.org/10.1037/0012-1649.25.2.247

Kliegl, R., Maayr, U., & Krampe, R. (1994). Time-accuracy functions for determining process and person differences: An application to cognitive aging. *Cognitive Psychology*, *26*(2), 134–164. https://doi.org/10.1006/cogp.1994.1005

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2007). Reliable change index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology*, *22*(4), 513–518. https://doi.org/10.1016/j.acn.2007.03.005

Kulik, J. A., Kulik, C.-l. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.2307/1162453

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Lampit, A., Hallock, H., & Valenzuela, M. (2014). Computerized cognitive training in cognitively healthy older adults: A systematic review and meta-analysis of effect modifiers. *PLoS Medicine*, *11*(11), e1001756. https://doi.org/10.1371/journal.pmed.1001756

Lange, S. (2013). *Transfer von kognitivem Training in den Alltag bei älteren Erwachsenen [Transfer of cognitive training into everyday lives of elderly adults]* (Unpublished doctoral dissertation). University of Magdeburg.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*(4), 981–1007. https://doi.org/10.1111/j.1744-6570.2005.00713.x

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672–1682. https://doi.org/10.1037/0021-9010.92.6.1672

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: 'Sinking shafts at a few critical points'. *Annual Review of Psychology*, *51*, 405–444. https://doi.org/10.1146/annurev.psych.51.1.405

Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, *37*(4), 412–421. https://doi.org/10.1016/j.intell.2009.03.011

Matton, N., Vautier, S., & Raufaste, É. (2011). Test-specificity of the advantage of retaking cognitive ability tests. *International Journal of Selection and Assessment*, *19*(1), 11–17. https://doi.org/10.1111/j.1468-2389.2011.00530.x

Mayr, U., & Kliegl, R. (1993). Sequential and coordinative complexity: Age-based processing limitations in figural transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1297–1320. https://doi.org/10.1037/0278-7393.19.6.1297

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, *58*(2), 480–494. https://doi.org/10.1016/j.jml.2007.04.004

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments*. New York: Springer Science & Business Media.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. https://doi.org/10.1037/a0028228

Melby-Lervåg, M., & Hulme, C. (2016). There is no convincing evidence that working memory training is effective: A reply to Au et al. (2014) and Karbach and Verhaeghen (2014). *Psychonomic Bulletin & Review*, *23*(1), 324–330. https://doi.org/10.3758/s13423-015-0862-z

Morley, M. E., Bridgeman, B., & Lawless, R. R. (2004). Transfer between variants of quantitative items. *ETS Research Report Series*, *2004*(2), 1–27. https://doi.org/10.1002/j.2333-8504.2004.tb01963.x

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta–analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 17–29. https://doi.org/10.1348/000711000159150

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105. https://doi.org/10.1037/1082-989X.7.1.105

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46–60. https://doi.org/10.3758/s13423-010-0034-0

Oberauer, K. (1993). Die Koordination kognitiver Operationen—eine Studie über die Beziehung zwischen Intelligenz und 'working memory.' [The coordination of cognitive operations: A study on the relation between intelligence and 'working memory']. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, *201*(1), 57–84.

Oberauer, K. (2009). Design for a working memory. In B. H. Ross (Ed.), *The psychology of learning* (Vol. 51, pp. 45–100). New York: Academic Press. https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K., & Kliegl, R. (2001). Beyond resources: Formal models of complexity effects and age differences in working memory. *European Journal of Cognitive Psychology*, *13*(1-2), 187–215. https://doi.org/10.1080/09541440042000278

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626. https://doi.org/10.1016/j.jml.2006.08.009

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045. https://doi.org/10.1016/S0191-8869(99)00251-2

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence—their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Olkin, I., & Gleser, L. (2009). Stochastically dependent effect sizes. In H. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York: Russell Sage Foundation.

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2015). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, *57*(3), 137–146. https://doi.org/10.1037/h0046137

Powers, K. L., Brooks, P. J., Aldrich, N. J., Palladino, M. A., & Alfieri, L. (2013). Effects of video-game play on information processing: A meta-analytic investigation. *Psychonomic Bulletin & Review*, *20*(6), 1055–1079. https://doi.org/10.3758/s13423-013-0418-z

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna: Austria. http://www.R-project.org

Randall, J. G., & Villado, A. J. (2017). Take two: Sources and deterrents of score change in employment retesting. *Human Resource Management Review*, *27*, 536–553.

Redick, T. S. (2015). Working memory training and interpreting interactions in intelligence interventions. *Intelligence*, *50*, 14–20. https://doi.org/10.1016/j.intell.2015.01.014

Roediger, H. L. III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, *24*(5), 563–572. https://doi.org/10.1037/a0019026

Salthouse, T. A. (2011). Effects of age on time-dependent cognitive change. *Psychological Science*, *22*(5), 682–688. https://doi.org/10.1177/0956797611404900

Salthouse, T. A. (2015). Test experience effects in longitudinal comparisons of adult cognitive functioning. *Developmental Psychology*, *51*(9), 1262–1270. https://doi.org/10.1037/dev0000030

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging*, *6*(1), 118. https://doi.org/10.1037/0882-7974.6.1.118

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology*, *40*(5), 813–822. https://doi.org/10.1037/0012-1649.40.5.813

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*(6), 800–811. https://doi.org/10.1037/a0013091

Scharfen, J., Blum, D., & Holling, H. (2018). Response time reduction due to retesting in mental speed tests: A meta-analysis. *Journal of Intelligence*, *6*(6). https://doi.org/10.3390/jintelligence6010006

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schleicher, D. J., Iddekinge, C. H. V., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology*, *95*(4), 603–617. https://doi.org/10.1037/a0018920

Schmidt, P. J., Keenan, P. A., Schenkel, L. A., Berlin, K., Gibson, C., & Rubinow, D. R. (2013). Cognitive performance in healthy women during induced hypogonadism and ovarian steroid addback. *Archives of Women's Mental Health*, *16*(1), 47–58. https://doi.org/10.1007/s00737-012-0316-9

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.01475

Schuerger, J., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, *45*(2), 294–302.

Shaffer, D. R., & Kipp, K. (2010). *Developmental psychology: Childhood and adolescence* (8th ed.). Belmont: Thomson Brooks/Cole Publishing Co.

Shing, Y. L., Schmiedek, F., Lövdén, M., & Lindenberger, U. (2012). Memory updating practice across 100 days in the COGITO study. *Psychology and Aging*, *27*(2), 451–461. https://doi.org/10.1037/a0025568

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628–654. https://doi.org/10.1037/a0027473

Smith, P. J., Blumenthal, J. A., Hoffman, B. M., Cooper, H., Strauman, T. A., Welsh-Bohmer, K., ... Sherwood, A. (2010). Aerobic exercise and neurocognitive performance: A meta-analytic review of randomized controlled trials. *Psychosomatic Medicine*, *72*(3), 239–252. https://doi.org/10.1097/PSY.0b013e3181d14633

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A mulitlevel meta-analysis of N-back training studies. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-016-1217-0

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, *35*(3), 283–300. https://doi.org/10.1016/j.intell.2006.07.006

Thorgusen, S. R., Suchy, Y., Chelune, G. J., & Baucom, B. R. (2016). Neuropsychological practice effects in the context of cognitive decline: Contributions from learning and task novelty. *Journal of the International Neuropsychological Society*, *22*(4), 453–466. https://doi.org/10.1017/S1355617715001332

Tombaugh, T. N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Archives of Clinical Neuropsychology*, *21*(1), 53–76. https://doi.org/10.1016/j.acn.2005.07.006

Toril, P., Reales, J. M., & Ballesteros, S. (2014). Video game training enhances cognition of older adults: A meta-analytic study. *Psychology and Aging*, *29*(3), 706–716. https://doi.org/10.1037/a0037507

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory & Cognition*, *28*(3), 341–348. https://doi.org/10.3758/bf03198549

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language*, *49*(4), 446–468. https://doi.org/10.1016/S0749-596X(03)00095-0

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Verhaeghen, P., Cerella, J., & Basak, C. (2004). A working memory workout: How to expand the focus of serial attention from one to four items in 10 hours or less. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1322–1337. https://doi.org/10.1037/0278-7393.30.6.1322

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03.

Wang, P., Liu, H. H., Zhu, X. T., Meng, T., Li, H. J., & Zuo, X. N. (2016). Action video game training for healthy adults: A meta-analytic study. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.00907

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV)*. San Antonio: Pearson Assessments.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. https://doi.org/10.3389/fpsyg.2013.00433

Wilson, R. S., Li, Y., Bienias, L., & Bennett, D. A. (2006). Cognitive decline in old age: Separating retest effects from the effects of growing older. *Psychology and Aging*, *21*(4), 774–789. https://doi.org/10.1037/0882-7974.21.4.774

Zehnder, F., Martin, M., Altgassen, M., & Clare, L. (2009). Memory training effects in old age as markers of plasticity: A meta-analysis. *Restorative Neurology and Neuroscience*, *27*(5), 507–520. https://doi.org/10.3233/RNN-2009-0491