Intelligence 67 (2018) 44–66

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Retest effects in cognitive ability tests: A meta-analysis

Jana Scharfen*, Judith Marie Peters, Heinz Holling

Institute of Psychology, Westfälische Wilhelms-Universität Münster, Fliednerstr. 21, 48149 Münster, Germany

ARTICLE INFO

Keywords:
Meta-analysis
Cognitive ability
Intelligence
Retest effect

ABSTRACT

Retest effects are referred to as the increase in test scores due to the repeated administration of cognitive ability tests. This meta-analysis attempts to update and extend previous meta-analyses by examining the size of retest effects and its determinants in a high number of cognitive ability tests for up to four test administrations. Strict inclusion and exclusion criteria were applied regarding study design, participant age and health status, and cognitive ability tests. An extensive literature search detected 174 samples from 122 studies, which resulted in 786 test outcomes and an overall sample size of 153,185. A comprehensive longitudinal multilevel meta-analysis revealed significant retest effects and no further score gains after the third test administration. Moderator analyses for multiple retests indicated that cognitive ability operation and content, equivalence of test forms, retest interval and participant age have a significant influence on the size of the retest effect. Implications for future research and retesting practice are discussed.

1. Retest effects in cognitive ability tests

Repeated administrations of cognitive ability tests occur frequently in selection settings, educational, and neuropsychological contexts. In fact, especially in personnel selection contexts, where cognitive ability tests find high acceptance and are often utilized as personnel decision-making tools (Ones, Viswesvaran, & Dilchert, 2005; Hülsheger, Maier, & Stumpp, 2007), retest effects have the potential to impede valid measurements. As long as causes, determinants and consequences of retest effects are not comprehended in detail, false decisions based on test results can be easily made (Randall & Villado, 2017).

A lot of research has been focusing on retest effects in cognitive ability tests (e.g., Arendasy & Sommer, 2017; Bartels, Wegryzyn, Wiedl, Ackermann, Ehrenreich, 2010; Freund & Holling, 2011; Lievens, Reeve, & Heggestad, 2007; Reeve & Lam, 2005; Villado, Randall, and Zimmer, 2016). To date, we have an approximate impression of the overall size of the effect thanks to important prior meta-analyses from Kulik, Kulik, and Bangert (1984), Hausknecht, Halpert, Di Paolo, and Moriarty Gerrard (2007), and Calamia, Markon and Tranel (2012). Hausknecht et al. (2007) made a contribution to the field by evaluating both coaching and retest effects. For 75 samples, the mere repetition of a test resulted in an improvement of almost a quarter of a standard deviation. From a clinical perspective, Calamia et al. (2012) analyzed retest effects in neuropsychological instruments for two test administrations and came to similar conclusions by giving estimates of practice effects for specific neuropsychological tests, however, effects were smaller in clinical than in healthy samples. Kulik et al. (1984) found a slightly larger effect of approximately one third of a standard deviation by analyzing 40 studies.

Most practical settings in which retesting takes place allow for more than two test administrations, or multiple retests. Findings in this field are relatively scarce, as only few studies explicitly focusing on retesting have administered more than three tests (Bartels et al., 2010). In the last years, retesting multiple times has gained more attention (e.g., Bartels et al., 2010; Puddey, Mercer, Andrich, & Styles, 2014). For three test administrations, Kulik et al. (1984) and Hausknecht et al. (2007) found increasing effects, which will be elaborated in more detail below. However, these prior findings have to be interpreted carefully as both analyses were based on a low number of samples and possible moderators of retest effects for more than two test administrations have not been examined meta-analytically. Updating and expanding results for multiple retest effects is of special interest, as due to theoretical deliberations a plateau effect would be expected (Donner & Hardy, 2015; Jaber & Glock, 2013; Newell & Rosenbloom, 1981). This is why it is important to investigate retest effects for more than two administrations meta-analytically.

The goals of the present meta-analysis are, on the one hand, to give an update of prior findings and expand results to multiple retests, and, on the other hand, to integrate methodological and theoretical developments from different perspectives into a more basic view on retest effects, including clinical and applied contexts of retesting. The three groups of causes of retest effects introduced by Lievens et al. (2007) will be elaborated on and hypotheses will be derived from a comprehensive theoretical framework with reference to this and to other important

* Corresponding author.
E-mail addresses: jana.scharfen@uni-muenster.de (J. Scharfen), holling@uni-muenster.de (H. Holling).

https://doi.org/10.1016/j.intell.2018.01.003
Received 18 July 2017; Received in revised form 11 January 2018; Accepted 16 January 2018

theoretical deliberations and prior research (e.g., Arendasy & Sommer, 2017; Freund & Holling, 2011; Randall & Villado, 2017; te Nijenhuis, van Vianen, & van der Flier, 2007). This approach will base this meta-analysis on a profound theoretical basis, possibly leading to new insights on underlying mechanisms that cause retest effects and pointing out relevant questions for future research. One major contribution of the current meta-analysis will be the differentiation between cognitive abilities according to the *Berlin Model of Intelligence Structure* (BIS; Jäger, 1982) and suggesting explanations for the disparity or similarity of retest effects between cognitive abilities.

Due to a forward and backward search of numerous important publications and an additional extensive literature search, it was possible to gather the potentially highest number of healthy and cognitively fully developed samples evaluated on this topic to date (122 studies and 174 samples), to extend findings to four test administrations and to investigate the influence of theoretically relevant determinants. Also, methodological shortcomings of prior works were addressed by, e.g., applying very strict inclusion and exclusion criteria, considering publication bias and not aggregating effect sizes for different cognitive ability domains if they were reported for the same sample, which was achieved by a comprehensive multilevel meta-analysis (e.g., Musekiwa, Manda, Mwambi, & Chen, 2016; Salanti, Higgins, Ades, & Ioannidis, 2008; Viechtbauer, 2010). All in all, this approach allowed us to expect new insights on the topic. Accordingly, this meta-analysis will update and expand prior findings, that is to say it will reliably summarize the current status of knowledge about the size of retest effects in cognitive ability tests and its determinants in healthy and cognitively fully developed samples for multiple retests on the basis of a profound theoretical framework.

1.1. The retest effect

The retest effect is defined as the change in test scores as a result of retaking the same or alternate cognitive ability test under comparable conditions (Lievens et al., 2007). It is also referred to as testing effect (Roediger & Butler, 2011), retest bias (Villado et al., 2016) or practice effect (Hausknecht et al., 2007). Although there is a broad acknowledgement of the existence of retest effects, not all of its determinants have been examined extensively, nor have the reasons for its occurrence and its impact on the psychometric quality of cognitive ability tests been fully understood (Lievens et al., 2007; Randall & Villado, 2017).

Three categories of causes of retest effects have been summarized by Lievens et al. (2007). Firstly, it is argued that the latent construct that is measured by the test could be enhanced by retesting, which leads to higher scores in repeated measurements. This explanation is seized in the research field of the so called testing effect. It is assumed that learning is enforced by test-taking, because retrieval practice might activate mnemonic enhancement (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006). Indeed, if the latent construct was enhanced by retesting, taking a test several times would not have an effect on the validity of the test. However, several studies contradict this view and imply that validity changes as a consequence of retesting (e.g., Hausknecht et al., 2002; Lievens et al., 2005; te Nijenhuis et al., 2007), which speaks against this first cause that might lead to retest effects. Generally, as cognitive ability is defined as a stable construct, an improvement due to retesting is seen critically.

Secondly, retest effects could be explained by the reduction of distorting and construct-irrelevant factors (Lievens et al., 2007; Matton, Vautier, & Raufaste, 2009; Freund & Holling, 2011). Participants' test anxiety, lack of understanding, or lack of familiarity are assumed to decrease when retested, which in turn leads to an increase in test scores. A variation in motivation might also affect the size of retest effects (Randall & Villado, 2017). It is argued that a person who is, e.g., less anxious about a test, or who understands the test fully, can probably show better results compared to when they were firstly confronted with the test. For example, it can be assumed that cognitive capacity is restricted when anxiety is high (Eysenck, Derakshan, Santos, & Calvo, 2007; Ng & Lee, 2015). When taking into account decreasing test anxiety due to repeated stimulus presentation, e.g., following the concept of habituation (Lader & Wing, 1964; Grissom & Bhatnagar, 2009), a higher amount of cognitive capacity might be available when retested. Studies investigating the causal relationship between these construct-irrelevant variables and retest effects, such as Anastasi (1981), Matton et al. (2009), Reeve and Lam (2005), and Reeve, Heggestad, & Lievens, (2009) find evidence that they contribute to causing the effect.

Lastly, the development and application of test-taking strategies or test-specific skills could also lead to an improvement of test scores (Lievens et al., 2007; te Nijehuis et al., 2007). Strategies are likely to be developed due to test taking, which might facilitate a better test performance when retested. This idea generally serves as a basis for several test coaching programs, as elaborated by, e.g., Allalouf and Ben-Shakar (1998) and Messick and Jungeblut (1981), and which are often based on strategies of test-wiseness (Millman, Bishop, & Ebel, 1965). Empirical evidence suggests strategies are indeed developed when retesting takes place and that the use of strategies increases test scores (Allalouf & Ben-Shakar, 1998; Arendasy & Sommer, 2017; Hayes, Petrov & Sederberg, 2015; Messick & Jungeblut, 1981).

These three causes of retest effects form a theoretical basis from which hypotheses of retest effects will be developed in the following. Generally, there is a high approval of retest effects in cognitive ability tests, as most of the causes elaborated above find theoretical and empirical support (e.g., Arendasy & Sommer, 2017; Matton et al., 2009; Lievens et al., 2007; te Nijehuis et al., 2007; Reeve and Lam, 2005, Reeve et al., 2009), and retest effects are a stable finding from previous primary and meta-analytic studies (Calamia et al., 2012; Hausknecht et al., 2007; Kulik et al., 1984).

Hypothesis 1. a. Retaking a cognitive ability test leads to higher test scores.

b. Retest effects between consecutive tests decrease with the number of test administrations.

1.2. Number of test administrations

Retest effects for three administrations have been summarized by Hausknecht et al. (2007) who found an effect of *Cohen's d* = 0.51 from the first to third test for 15 samples, without evaluating retest effects for further repetitions. Kulik et al. (1984) found retest effects of *Cohen's d* = 0.53 from the first to third test and of *Cohen's d* = 0.69 from the first to fourth test, assuming a linear improvement. Following theoretical assumptions above, a linear improvement seems implausible. Since 1984, several new studies investigating retest effects administering more than two tests appeared (e.g., Albers & Höft, 2009; Bartels et al., 2010; Dunlop, Morrison & Cordery, 2011; Puddey et al., 2014), whose results mostly describe a large score gain from first to second test and retest effects becoming smaller with the number of tests, rather suggesting a non-linear progression.

From a theoretical view, the widely acknowledged *power law of practice* describes the assumption that learning curves show diminishing gains over time (Donner & Hardy, 2015; Jaber & Glock, 2013; Newell & Rosenbloom, 1981). After a first phase of improvement, no further gains are observed. As Newell and Rosenbloom (1981) state, this theory should hold for “all types of mental ability” (p. 33), and thus it can be applied to retest effects in cognitive ability tests when multiple tests are administered. According to the *power law of practice*, retest effects will decrease with the number of test administrations.

It can be assumed that for all of the three causes of retest effects outlined above, their influence decreases when the participant is retested multiple times. Their role is expected to be greater in the first repetitions of the test, with mechanisms described above leveling off after a first or second test experience. For example, after test-specific strategies have been developed and applied within first test session

(Hayes et al., 2015; Messick & Jungeblut, 1981), there might be smaller gains in strategy-development in further tests. Similarly, a reduction of construct-irrelevant factors, such as anxiety and rule comprehension, has been found to take place from the first to second test (e.g., Matton et al., 2009). It can be assumed that after this first reduction, a smaller decrease of these factors takes place. For example, rule comprehension should be achieved almost fully after the first test, with smaller changes in the further sessions. This suggests a growth of test scores reaching a plateau after a few tests.

1.3. Moderators

There are numerous factors influencing retest effects that can be deduced from the three causes and that supposedly foster or reduce retest effects (Randall & Villado, 2017). In this meta-analytic study, hypotheses for those moderators for which data was reported most commonly in primary studies were formulated and investigated for several test administrations. Also, generalized methods of assessment of moderating variables had to exist that make it possible to give standardized values that are comparable between samples,

1.3.1. Cognitive ability operation

Hausknecht et al. (2007) differentiated between cognitive operations according to Carroll's three stratum model (Bors, 1993) and found no difference in retest effects between verbal, quantitative and analytical subfacets. Still, differences between cognitive abilities have been found by several primary (Dunlop et al., 2011; Maerlender, Masterson, James, Beckwith, & Brolinson, 2016; Puddey et al., 2014) and also meta-analytic studies (Calamia et al., 2012). Thus, a moderating effect of cognitive abilities on the retest effect seems plausible. This might not have been carried through by Hausknecht et al.'s results, because they have averaged estimates across operations if they were given for the same sample.

For the current meta-analysis, an approach of differentiation between cognitive operations was chosen that is distinct from those of prior meta-analyses. Cognitive abilities were categorized by means of the BIS (Jäger, 1982) because of its integrative approach. Following the BIS, cognitive abilities can be defined on two dimensions as illustrated in Fig. 1. The first dimension clusters cognitive abilities on an operational level. Four cognitive operations can be differentiated:processing speed, divergent thinking, memory and reasoning abilities. The second dimension focuses contents of cognitive ability tests: figural, numerical, and verbal contents can be distinguished from each other (Kubinger & Jäger, 2003; Jäger, 1982).

Within this model, processing speed is defined as a measure of concentration, working speed and ease of perception. This ability is especially important for easy tasks. Memory is defined as active memorizing and reproduction or recognizing. Divergent thinking means the flexible and innovative production of new ideas that considers many views and opinions and deploys them to find problem-centered solutions. Finally, reasoning is defined as the processing of complex information in difficult tasks that require the integration of pre-existing knowledge and formal logical thinking. The second dimension of the BIS incorporates the three content domains. Figural contents can be any kind of images, objects, abstract or spatial material. Numerical contents typically consist of numbers, and verbal contents can be letters, words, sentences or stories. All operations and contents can be combined with each other, leading to twelve different performance domains, which altogether represent general mental ability, or general intelligence (Jäger, 1982). Most cognitive ability tests can be categorized into one of these twelve performance domains. For example, Ravens matrices can be categorized as figural reasoning tests, whereas digit span tasks would be numerical memory tasks (see Table A2).

Differences in retest effects and their progression over several repetitions between the cognitive operations are hypothesized. Here, the complexity of the operations plays a key role. According to their

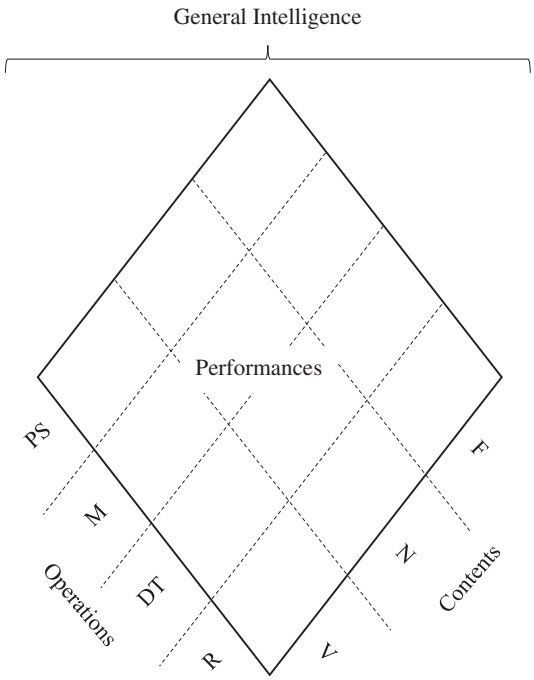

Fig. 1. Berlin Model of Intelligence Structure (BIS) according to Jäger (1982). PS = processing speed; M = memory; DT = divergent thinking; R = Reasoning; F = figural; N = numerical; V = verbal.

definitions (Jäger, 1982; Kubinger & Jäger, 2003), processing speed tasks consist of the lowest and reasoning tasks of the highest complexity, whereas memory and divergent thinking can be considered comparably complex. The complexity of a task is determined by the number of cognitive operations that have to be conducted in order to solve a task. When solving reasoning tasks, for example, several cognitive steps have to be passed (Carpenter, Just & Shell, 1990). Reasoning has been found to be highly correlated or even dependent on working memory (Buehner, Krumm & Pick, 2005; Kyllonen & Christal, 1990; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002), which stresses the complexity of this ability. In contrast, in processing speed tasks, one simply has to react fast without having to pass several cognitive processes.

For easier tasks, it is assumed that a lower reduction of distorting factors takes place, because there are less in number. For example, participants can generally be more anxious of difficult tasks than of easy tasks (e.g., Sarason & Palola, 1960). For example, when being confronted for the first time with a complex reasoning task, such as figural matrices or mental rotation, participants might be intimidated. When retested, this anxiety can be reduced, because they have learned in a previous test that they are in fact able to solve the task and understand how it works. Thus, in easy tasks, test anxiety cannot be reduced as much as in difficult tasks because it can be assumed to be lower from the beginning. The same applies, for example, for rule incomprehension. This suggests smaller retest effects in processing speed tests compared to more complex tasks, as found by, e.g., Maerlender et al. (2016). Randall and Villado (2017) also claim that factors such as anxiety play a smaller role in processing speed tasks. For the same reasons, when retested with processing speed tasks, a plateau is expected to be reached more quickly.

A similar argumentation can be put forward for the third cause of retest effects: Fewer possible solving strategies exist for easier tasks than for more complex tasks. Randall and Villado (2017) argue that speeded tests are less prone to retest effects and that distortion by the use of tests-specific skills is less probable (Arthur, Glaze, Villado, & Taylor, 2010). This can be explained by speeded tests allowing less strategy use than power tests of higher order operations, like reasoning.

For easier tasks, these mechanisms would lead to smaller retest effects that also arrive at a plateau more quickly.

Hypothesis 2. a. The size of the retest effect will differ between cognitive operations: Processing speed tests will exhibit smaller retest effects than memory and divergent thinking tests, which will show smaller effects than reasoning tests.

b. The number of administrations needed to reach a plateau differs between cognitive abilities: Retest effects in processing speed tasks will reach a plateau first, followed by divergent thinking and memory, and lastly by reasoning tasks.

1.3.2. Test form

A seemingly robust result from the previous meta-analyses of Calamia et al. (2012), Hausknecht et al. (2007), and Kulik et al. (1984), is the difference between retest effects depending on whether the same or an alternate test form was used for retesting. Cook and Campbell (1979) argued that taking the exact same test twice can lead to larger retest effects, because participants are likely to remember their answers and errors from the last test and use less capacity when recalling these answers than solving items anew (as cited in Hausknecht et al. 2007). Free capacity can, for example, be used to solve more difficult items within the same test that could not have been solved during a prior test. This procedure of recalling answers from a previous test can be considered a test-specific strategy, as suggested by the third group of causes of retest effects (Lievens et al., 2007; Nevo, 1976). It is inapplicable for alternate test forms, because items appearing differently are prevented from such item-specific memory effects. However, there is evidence for retest effects due to implicit memory effects as well (Hasegawa, 1997).

Also, according to the second group of causes of retest effects (Lievens et al., 2007), distorting and construct-irrelevant factors, such as test anxiety, rule comprehension and low familiarity with the first test, might decrease faster when taking an identical compared to an alternate test version. The recognition of identical items can increase familiarity with the test and lower test anxiety (Anasasi, 1981; Arendasy & Sommer, 2017). Because the rules of alternate test versions are the same, rule comprehension can be increased due to retesting in alternate forms as well. Though, when being presented with the exact same test, distorting factors would decrease to a higher degree, which would lead to smaller retest effects in alternate test forms.

Prior meta-analyses have focused on moderators of retest effects from the first to second test administration. Though, Kulik et al. (1984) differentiated between test forms when analyzing multiple retest effects. However, their analysis for more than two administrations was based on eleven samples only and several new studies using both identical and alternate test forms have been published since then. The influence of test equivalence on retest effects may seem a stable result, but it has not been meta-analyzed extensively and based on an efficiently high number of studies yet, whether the effect remains stable over more than two test administrations.

Hypothesis 3. Retest effects will be larger for identical than for alternate test forms.

1.3.3. Test-retest interval

A negative influence of the amount of time between tests on the size of retest effects has been found by, e.g. Calamia et al. (2012), Hausknecht et al. (2007) and Salthouse, Schroeder and Ferrer (2004). Regarding the specific test-retest interval after which retest effects diminish, results vary between two and thirteen years. The longer the test-retest interval, the less test-specific skills (e.g., strategies, rules of the test, correct answers to items), can be recalled. For example, the test-retest interval between the measurements directly corresponds with item-specific memory effects, which can lead to larger retest effects as argued above. Consequently, they would be larger if the retest interval is shorter (Hausknecht et al., 2007; Salthouse et al., 2004). A longer test-retest interval between administrations might also lead to a decrease in familiarity and rule comprehension, which are examples for the second cause of retest effects (Lievens et al., 2007), because participants are less likely to remember the test in order to feel familiar with it, its rules and their comprehension thereof. Even if the latent cognitive ability was advanced by retesting, as suggested by the first cause according to Lievens et al. (2007), time would also have a negative effect on the retest effect, because it is not probable that retesting has sustainable consequences on this stable latent construct. Again, this moderator has not been investigated in the case of retesting multiple times by prior meta-analyses. For the current meta-analysis, the effect of test-retest interval on the retest effect will therefore be investigated for more than two test administrations as well.

Hypothesis 4. Retest effects will decline with a longer test-retest interval between administrations.

1.3.4. Age

Evidence is mixed with regard to whether age has an influence on retest effects. For example, Calamia et al. (2012), Lo, Humphreys, Byrne and Pachana (2012), Schleicher, Van Iddekinge, Morgeson, and Campion (2010), van Eersel et al. (2015), and Van Iddekinge, Morgeson, Schleicher, and Campion (2011) found a lowering effect of age on retest effects, whereas Sanches de Oliveira, Trezza, Busse and Jocob-Filho (2014) and Bartels et al. (2010) did not. The moderating effect of age on the retest effect was not investigated by Hausknecht et al. (2007) and Kulik et al. (1984) and none of the prior meta-analyses tested its influence for more than two test administrations.

Generally, fluid intelligence decreases with age (Cattell, 1987) and this influences the ability to learn from test experiences (Van Iddekinge et al., 2011). Braver and Barch (2002) found an age-related decrease in the ability to represent, maintain and update context information in healthy adults. It might, thus, be easier for younger people to develop test-specific strategies and maintain them from one administration to the other, than for older people. More specifically, item-specific memory effects, which can account for higher scores in subsequent tests as outlined above, could be smaller in older participants. Declining memory ability with age might therefore lead to smaller retest effects.

Hypothesis 5. Retest effects will decline with age.

1.3.5. Intelligence

Although Kulik et al. (1984), Arendasy and Sommer (2013), and Randall, Villado, and Zimmer (2016) found a positive relationship between general intelligence and retest effects, results are inconsistent, as, e.g., Bartels et al. (2010) did not find any association. Neither of the more recent meta-analyses investigated the influence of general intelligence on the size of the retest effect (Hausknecht et al., 2007; Calamia et al., 2012), and for multiple retests, no meta-analytic results were reported.

Randall & Villado (2017) argue that higher general intelligence provides testees with the ability to learn more from the first test than less intelligent people. General intelligence can thus also be seen as representing learning ability (Randall et al., 2016; Guthke & Beckmann, 2001), and the ability to develop and maintain test-specific skills and strategies, and therefore enhancing retest effects from one test to the next. Additionally, item-specific memory effects can be expected to be larger for more intelligent people, because memory can be defined as a facet of intelligence, e.g., according to Jäger (1982).

Hypothesis 6. Retest effect will increase with higher intelligence.

1.4. Methodology of prior meta-analyses

Main results of prior meta-analyses (Calamia et al., 2012; Hausknecht et al., 2007; Kulik et al., 1984) have been referred to above.

With respect to the influential findings from these prior works, a few shortcomings have to be admitted. From a methodological perspective, it has to be mentioned that publication bias was not addressed in any of the analyses. Also, Kulik et al. (1984) and Hausknecht et al. (2007) aggregated effect sizes for different cognitive ability domains if they were reported for the same sample into one effect size, which made a comparison between cognitive abilities difficult. In contrast, Calamia et al. (2012) reported retest effects for specific neuropsychological tests and did not cluster them according to a comprehensive model of intelligence. Unfortunately, Calamia et al. (2012) and Hausknecht et al. (2007) did not check the influence of intelligence on the size of retest effects, although Kulik et al. (1984) found it to be significant. Importantly, these prior works applied rather loose inclusion and exclusion criteria concerning sample characteristics, allowed activities between measurements and varying difficulties of test forms, which questions the reliability of their results. Further, new studies have been published in the last years that have not been meta-analyzed yet and not all of the previous meta-analyses considered studies from earlier ones, which signals a lack of completeness. All in all, these methodological limitations and the high number of new studies questions the reliability and generalizability of the currently available meta-analyses on the topic.

For these reasons, the current meta-analysis builds on previous ones by extending the focus of retesting to multiple test administrations and thereby investigating a non-linear development of retest effects. Also, a high number of moderator analyses are conducted and cognitive ability tests are differentiated on a highly resolved level. By applying strict eligibility criteria, longitudinal meta-analytic methods and including a high number of new studies, this meta-analysis gives a reliable and comprehensive summary on the current knowledge about retest effects in cognitive ability tests.

2. Methods

2.1. Inclusion and exclusion criteria

Studies were identified as eligible if they met the following inclusion and exclusion criteria. Inclusion criteria were the following: (a) The same or an alternate but equally difficult version of a cognitive ability test was administered. Hausknecht et al. (2007) included studies that used different test forms for retesting that were actually not equally difficult (e.g., Lane, Penn, & Fischer, 1966; Spielberger, 1959; Woehlke & Wilder, 1963). This may distort the gain in test scores that are exclusively due to retesting, which is why, in this analysis, studies were excluded if the eligible study reported that alternate test forms were not equally difficult. (b) The test was administered at least twice within the same sample under comparable conditions. (c) The test used in the study had to measure at least one cognitive ability as defined in the BIS by Jäger (1982) or general intelligence. (d) The samples' mean age had to be between 12 and 70 years. In contrast to Hausknecht et al. (2007), who did not set up age restrictions, we included only subjects, for which the latent cognitive ability can be considered relatively stable within the test-retest interval. As change in the latent variable is one of the three causes of retest effects (Lievens et al., 2007), it was controlled for differences between mechanisms that cause retest effects between samples in order to be able to conclude about the causes without making differences between age groups. For example, it is more probable to observe retest effects in a 10-year old caused by latent change due to cognitive development, compared to a 40-year-old, for whom other mechanisms are more probable to be causing retest effects. Thus, different mechanisms that lead to retest effects might take place in age groups for which a change in the latent variable can be expected due to reasons other than retesting. A change in the latent variable due to cognitive development is more probable to occur in age groups above 70 and below 12 than in others (Cattell, 1987; Holling, Preckel, & Vock, 2004; Shaffer & Kipp, 2010). Below the age of 12 and above the age of 70, retest effects might therefore reflect changes in the latent ability that are due to cognitive development, and not due to retesting. Including these age groups would therefore have made a theoretical discussion very difficult. (e) Furthermore, participants had to be neither mentally nor physically challenged. (f) Finally, the effect size had to be given or its calculation had to be possible with the given information.

Studies were excluded if the following exclusion criteria applied: (a) If any systematic activity, such as a video game intervention, test coaching, or a medical treatment, was administered in the samples between test administrations samples were excluded. A systematic activity between tests can deter the score gain that is exclusively due to retesting and thus could contaminate the retest effect. (b) Neuropsychological tests that differentiate in the clinically relevant area of cognitive ability only, e.g., Mini-Mental State Examination, were excluded. The relevant population for this meta-analysis, which are healthy and cognitively fully developed samples, achieve maximum scores in these tests (as they are mostly administered for the purpose of diagnoses in clinical samples), which prohibits the observation of score gains. (c) Tasks that have been shown to mainly measure working memory or other executive functions, such as *N*-back tasks, complex span or inhibition tasks, were excluded because they measure different cognitive functions than those defined by the BIS. (d) General knowledge or vocabulary tasks (e.g., WISC information and vocabulary subtests, SAT-V) were excluded due to the BIS definition of intelligence and their high dependence on education and culture. Also, different retesting mechanisms can be assumed for these kinds of tests (e.g., Halford, Cowan & Andrews, 2007). (e) In addition, studies that exclusively report reaction times were not included. Ackerman (1987) suggests considering reaction times and score outcomes in separate analyses, and previous studies have found differences between practice effects regarding scores compared to reaction times (e.g., Enge et al., 2014). (f) Finally, outcomes that showed ceiling effects were excluded (e.g., Bartels et al., 2010: figure and line copy and list recognition tasks from the RBANS).

2.2. Literature search

For a flow chart of the literature search and study selection process following Moher, Liberati, Tetzlaff, Altman, & Group (2009), see Fig. 2. Because Hausknecht et al. (2007) ended their search in 2005, we conducted an extensive database literature search for studies published between 2005 and December 2016. Abstracts and key words of studies were searched for the terms test*, assess*, cognit*, intelligen*, aptitude test, achievement test, IQ, reasoning, logical thinking, analytical thinking, inductive thinking, processing speed, mental speed, divergent thinking, memory, recall, retest*, repeat*, repetit*, practice, retak*, train*, and coach*. The databases PsycARTICLES, Academic Search Premier, Business Source Premier, PsycINFO and PSYNDEX and – to obtain a higher number of unpublished studies - ProQuest Dissertation & Theses were used. The search was limited to studies published in the English or German language. This search yielded m = 34,313 studies that were screened for eligibility by scanning titles and abstracts.

In addition, we conducted a forward and backward search based on meta-analyses focusing on retest effects and on training or coaching effects in cognitive ability tests, which allowed us to find relevant studies published before 2005 (Au et al., 2015; Ball, Edwards, & Ross, 2007; Becker, 1990; DerSimonian & Laird, 1983; Hausknecht et al., 2007; Karch, Albers, Renner, Lichtenauer, & Kries, 2013; Kelly et al., 2014; Klauer, 2014; Klauer & Phye, 2008; Kulik et al., 1984; Lampit, Hallock, & Valenzuela, 2014; Powers, Brooks, Aldrich, Palladino, & Alfieri, 2013; Schuerger & Witt, 1989; Scott, Leritz, & Mumford, 2004; te Nijenhuis et al., 2007; Toril, Reales, & Ballesteros, 2014; Uttal et al., 2013; Wang et al., 2016; Zehnder, Martin, Altgassen, & Clare, 2009). Training and coaching intervention studies were taken into account because of passive control groups fitting our criteria. This search revealed an additional m = 657 potentially relevant studies.

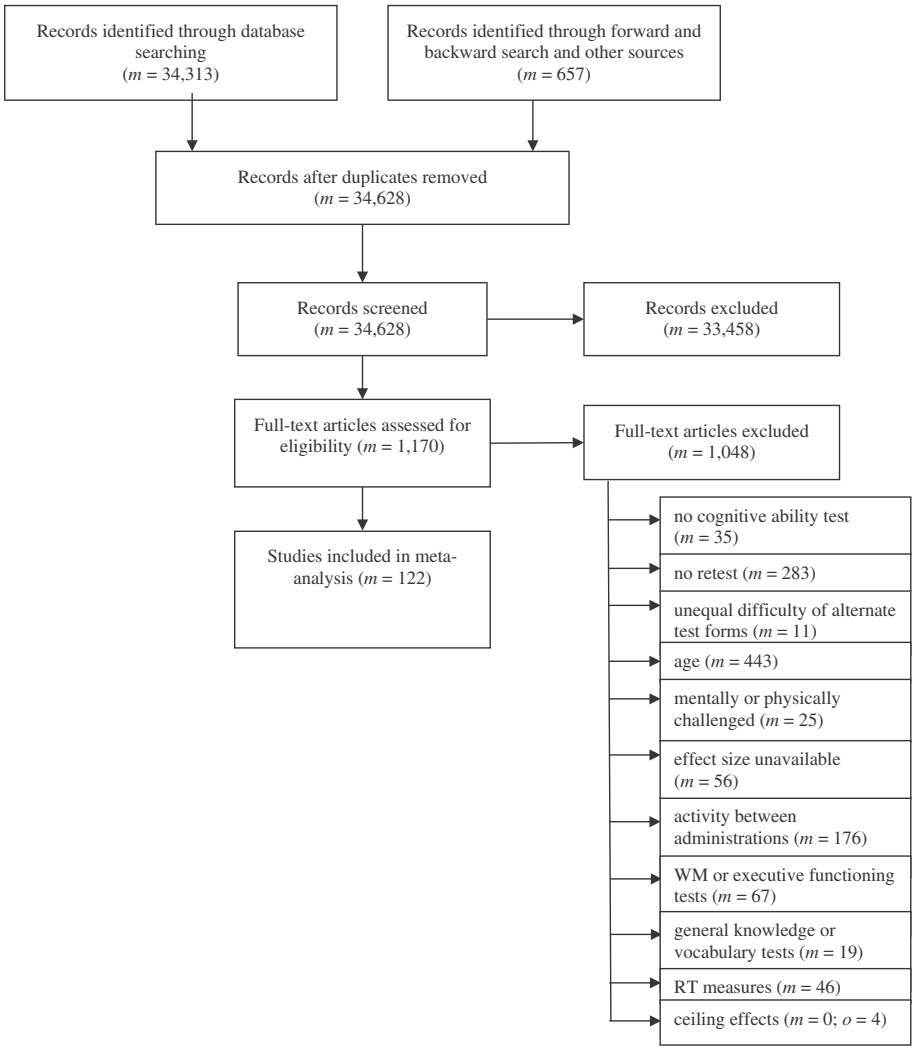

Fig. 2. PRISMA Flow Diagram (Moher et al., 2009) illustrating the search and selection process. m = number of studies; WM = working memory; RT = reaction time; o = number of outcomes.

In all, $m = 34{,}628$ studies were screened, of which $m = 1170$ were checked for eligibility in detail by applying the inclusion and exclusion criteria described above. Finally, $m = 122$ eligible studies containing $k = 174$ samples were identified. Reasons for exclusion and estimates of the excluded number of studies per criteria are illustrated in Fig. 2. The most common reasons for exclusion were the age of the sample (inclusion criteria (e)), retesting did not take place as defined in inclusion criteria (b), and activity between administrations, such as active control groups of intervention studies (exclusion criteria (a)). Note that studies could fail to fulfill more than one criteria and given values for reasons for exclusion are estimates of the number of studies. This is because the exclusion of studies was based on a failure to fulfill one or two criteria, and no further detailed inspection of other unfulfilled criteria in already excluded studies took place.

2.3. Coding procedure

A very definite coding scheme incorporating all relevant information about the studies was utilized, which is provided in the Appendix (Table A1). All studies were coded by two of the authors and finally every single study was carefully double-checked for coding errors. Any cases of insecurity and inconsistencies were discussed and corrected. It was not applicable to evaluate neither study quality (Cooper, 2017) nor risk of bias (Moher et al., 2009), because the strict inclusion and exclusion criteria already set the quality of included studies to a very high standard. For this meta-analysis, no protocol had been registered (Moher et al., 2009).

Several variables were coded, such as if the study had been published or not, and the study context (field vs. research) etc. These variables primarily served to describe characteristics of eligible studies, samples and tests used in retesting studies and represent categories of information that was given by most eligibly studies. For these variables, no hypotheses were developed a priori and no significant differences in retest effects due to these variables were expected. However, as retest effects are not yet fully understood, and also as these variables might give insight into further moderating variables, we exploratively tested their influence as control variables. For example, Hausknecht et al. (2007) analyzed differences between the field and research context and did not find a significant moderating effect, which is why in this analysis, it was included as a control variable but without a corresponding hypothesis.

As mentioned above, cognitive ability tests were categorized by means of the BIS (Jäger, 1982; Fig. 1). Cognitive ability operations were coded as processing speed, memory, divergent thinking, reasoning tasks, or, if the test measured more than one of these operations, as general intelligence. Contents used by the cognitive ability test were coded as figural, numerical, verbal, or, for the case that more than one content was used, it was coded as *several*. See Table A2 in the Appendix for a categorization of tests into the BIS.

In longitudinal meta-analyses as the current one, contrast variables are used to specify the comparison of test administrations and later included in the model (e.g., Salanti et al., 2008). In this case, we were able to analyze retest effects up to four test administrations. Thus, three contrast variables

Table 1
Test, study and sample characteristics.

No. of administrations	Level	Variable	*M*	*SD*	*Min*	*Max*	NA (%)
2	Outcome	$SMCR_{1.2}$	0.29	0.37	− 0.84	1.99	0.00
	(*o* = 566)	Test form (% alternate)	27.56				0.00
		Test-retest interval (weeks)	38.60	71.04	0.00	312.90	3.01
		Rule explanation (%)	15.28				74.55
		Practice items (%)	28.89				68.19
	Study	Published (%)	95.08				0.00
	(*k* = 174)	Year of publication	2000	20.36	1921	2016	0.00
		Post hoc analyses (%)	11.48				0.00
		Field studies (%)	13.93				0.00
		No. of tests	2.43	1.12	2.00	10.00	0.00
	Sample	Samples size	890.6	5137.98	5.0	61,500.00	0.00
	(*m* = 122)	Age (years)	28.92	16.31	12.30	67.35	4.94
		Intelligence (*z*)	0.69	0.57	− 0.92	1.80	72.44
		Gender (% male)	44.70	17.01	0.00	100.00	13.4
		Control group (%)	43.26				0.00
		Dropouts (%)	26.46	43.32	0.00	96.86	28.6
3	Outcome	$SMCR_{1.3}$	0.33	0.44	− 0.66	1.97	
	(*o* = 181)	Test form (% alternate)	27.78				0.00
		Test-retest interval (t1.3, weeks)	95.33	150.67	0.00	388.50	0.00
		Rule explanation (%)	26.87				61.75
		Practice items (%)	30.86				54.64
	Study	Published (%)	96.77				0.00
	(*k* = 31)	Year of publication	2002	22.12	1921	2016	0.00
		Post hoc analyses (%)	9.68				0.00
		Field studies (%)	16.13				0.00
		No. of tests	3.74	1.65	3.00	10.00	0.00
	Sample	Sample size	479.50	1337.35	9.00	6369.00	0.00
	(*m* = 50)	Age (years)	24.12	13.79	12.30	67.35	1.10
		Intelligence (*z*)	0.94	0.38	0.27	1.80	65.75
		Gender (% male)	53.60	22.39	0.00	100.00	7.18
		Control group (%)	0.91	10.45			0.00
		Dropouts (%)	7.03	18.38	0.00	96.47	3.31
4	Outcome	$SMCR_{1.4}$	0.49	0.52	− 0.40	2.45	
	(*o* = 39)	Test form (% alternate)	17.95				0.00
		Test-retest interval (t1.4, weeks)	25.36	21.66	0.00	51.79	0.00
		Rule explanation (%)	50.00				53.84
		Practice items (%)	50.00				53.84
	Study	Published (%)	90.00				0.00
	(*k* = 10)	Year of publication	2000	19.61	1949	2014	0.00
		Post hoc analyses (%)	10.00				0.00
		Field studies (%)	30.00				0.00
		no. of tests	5.4	2.17	4.00	10.00	0.00
	Sample	Samples size	118.50	157.20	9.00	545.00	0.00
	(*m* = 12)	Age (years)	43.17	19.03	15.41	64.80	2.56
		Intelligence (*z*)	1.23	0.52	1.00	1.80	51.28
		Gender (% male)	83.51	0.28	0.00	100	5.13
		Control group (%)	4.75				0.00
		Dropouts (%)	18.39	12.86	0.00	60.87	5.13

Note. *o* = number of outcomes; *m* = number of studies; *k* = number of samples; *Min* = minimum; *Max* = maximum; NA (%) = proportion of unavailable data; *SMCR* = standardized mean change with raw score standardization.

(t2, t3, t4) were coded, which described the comparisons between administrations that the data referred to. For example, the identifying variable t2 was coded as 1 if the data reported compared the first to the second test, and it was coded as 0 for any other comparison (t3 = 1 for first to third test, t3 = 0 for other comparisons; t4 = 1 for first to fourth test, t4 = 0 for other comparisons).

Average intelligence of the sample was coded as a *z*-standardized estimate to reach comparability between samples. If standardized estimates were not reported by eligible studies but standardization was possible, e.g., if values reported were given on the IQ scale, they were transformed into *z*-scale metric. In most studies, however, it was not possible to standardize values if they were not given, as reflected in the high percentage of missing values for this moderator (72%, see Table 1).

2.4. Extraction of effect sizes

This meta-analysis and the calculation of effect sizes were conducted by means of the program *R* (Version 3.4.2; R Core Team, 2015) using the package *metafor* (Viechtbauer, 2010). As outcomes from the same sample were compared between two test administrations, bias corrected standardized mean change with raw score standardization (*SMCR*) was chosen as the meta-analytic effect size (Becker, 1988; Gibbons, Hedeker, & Davis, 1993; Viechtbauer, 2010). The following formula was used to calculate effect sizes for all outcomes: $SMCR_{1.t} = J \times \frac{m_t - m_1}{sd_1}$, where J is the correction coefficient accounting for bias, m_t is the mean score at the respective t = second, third or fourth test, m_1 is the mean score at the first test, and sd_1 is the standard deviation of the mean score of the first test. Note that sd_1 was used, because a pooled standard deviation would have led to inconsistent results between comparisons of test administrations and resulting effect sizes can be considered less affected by bias compared to a standardization using the pooled standard deviation (Morris & DeShon, 2002). Also, retesting might have an influence not only on the size of the test scores but also on their variance (Ackerman, 1987), which was controlled using the standardization by sd_1. The sampling variance was computed as $var(SMCR_{1.t}) = \frac{2 \times (1 - r_{1.t})}{n_t} + \frac{SMCR^2_{1.t}}{2 \times n_t}$, where $r_{1.t}$ is the

Pearson's correlation between outcomes of the first and the respective t administration, and n_t is the sample size at the respective t administration. These formulas according to Becker (1988) are used by the *escalc*() function in metafor, when *SMCR* is chosen as the effect size (Viechtbauer, 2010). Generally, if several outcome values were reported for a sample, effect sizes were not aggregated but calculated for each outcome.

To be able to calculate $var(SMCR_{1.t})$, if $r_{1.t}$ was not given by the eligible study, it had to be estimated. For $o = 93$ outcomes, test-retest correlations were reported by eligible studies. Thus, for the rest of the outcomes, $r_{1.t}$ had to be estimated. According to Calamia, Markon, and Tranel (2013), who meta-analyzed test-retest correlations of neuropsychological tests, $r_{1.t}$ decreases with the length of the test-retest interval, increases with participant age, and is smaller for alternate than for identical test forms. A linear model predicting *Fisher-z* transformed correlations ($z_{r1.t}$) from relevant variables ($z_{r1.t} \sim \text{time}_{1.t}$ + test form + age) was calculated based on the 93 given correlations. Then, this model was used to predict missing $z_{r1.t}$, which were, as a last step, retransformed into Pearson's correlations. Resulting $r_{1.t}$ had a mean of $M_{r1.t} = 0.66$ and a standard deviation of $SD_{r1.t} = 0.09$.

Finally, the following corrections for unreliability of cognitive ability tests were applied for effect sizes and variances by the formula of Hunter and Schmidt (2004): $SMCR_{1.t\_corrected} = SMCR_{1.t\_uncorrected} \times \frac{1}{\sqrt{rel}}$, using the reported Cronbach's α of cognitive ability tests as an indicator of reliability (*rel*). Sampling variances of the effect sizes were corrected by $var(SMCR_{1.t})_{corrected} = var(SMCR_{1.t})_{uncorrected} \times \frac{1}{rel}$. For 89% of the outcomes, α was not reported by the studies. In these cases, it was estimated as $rel = 0.83$, which was the mean of all reported values.

2.5. Analysis

A comprehensive random effects meta-analysis was conducted. The following meta-analytic model was applied:

$$y_{ijc} = \mu_c + u_{ic} + w_{ijc} + \varepsilon_{ijc}$$

where, for each comparison of administrations c (1.2, 1.3, or 1.4), y_{ijc} is the *j*th effect size from the *i*th study, μ_c is the true *SMCR*, u_{ic} is a random effect at the level of studies, w_{ijc} is a random effect at the level of samples, and ε_{ijc} is the sampling error (e.g., Bryk & Raudenbush, 1988; Konstantopoulos, 2011).

In longitudinal meta-analyses, effects can show autoregressive correlations, meaning the correlation between true effects becomes smaller with the number of administrations. This can be modeled as autoregressive or heteroscedastic random effects (Ishak, Platt, Joseph, Hanley, & Caro, 2007; Trikalinos & Olkin, 2012; Musekiwa et al., 2016; Viechtbauer, 2010) and was resolved by specifications for the variance structure in the meta-analytical model. The chosen specification of the model admits autoregressive and heteroscedastic variances, which allows decreasing and different amounts of variance between comparisons. When tested against models with autoregressive structures only and models without specifications of the variance structure by model comparisons, the best fit was indicated for the model applying both heteroscedastic and autoregressive structures ($p < 0.001$). This model was used for all further analyses.

Further, covariances between sampling errors are not accounted for by multilevel modeling, which leads to unreliable standard errors. Thus, cluster-robust methods as described by Hedges, Tipton, and Johnson (2010) were applied, using studies as clusters.

According to the hypotheses, linear hypotheses were specified by the use of the contrast variables t2, t3, and t4 to test retest effects between administrations (1.2, 1.3, 1.4, 2.3, 3.4) and also between these comparisons of administrations (1.2 vs 2.3, 2.3 vs 3.4; e.g., Salanti et al., 2008)). Subgroup analyses and meta-regressions were conducted by adding moderators into the model by interaction terms of the moderator and contrast variables indicating the comparison of administrations. Moderators were tested by one separate model each. Control variables, for which no significant effects were expected, were analyzed exploratively and analogously to moderators. As meta-analysis relies on data from eligible studies, moderators can be confounded, if characteristics of eligible studies happen to co-occur. Associations between moderators were calculated to become aware of possible confounding. Note that from associations, no causal relationship between moderators can be deduced, as they only describe co-occurrences of characteristics of eligible studies.

Generally, for directed hypotheses, one-sided *p*-values will be reported. A significance level of $\alpha = 0.05$ was applied and type-1 error was controlled for by Bonferroni-Holm corrections (Holm, 1979; Holland & DiPonzio Copenhaver, 1988). The α-level was adjusted for significance tests that were related to the same moderator, or related to connected hypotheses (such as Hypothesis 1a and b).

To date, methods that indicate publication bias for complex multilevel models have not yet been derived. Nonetheless, funnel plots were inspected to obtain an impression of a possible publication bias (Cooper, Hedges, & Valentine, 2009). The funnel plot has to be interpreted carefully, because it neglects multilevel structures. Nonetheless, a possible relationship between effect sizes and standard errors can be monitored. In addition, it was tested if published studies showed larger effects than unpublished ones.

3. Results

3.1. Sample, study, and outcome characteristics

This meta-analysis contains $k = 174$ samples from $m = 122$ studies, $o = 786$ outcomes and an overall sample size of $N = 153{,}185$. Table 1 provides detailed descriptive information about sample, study and outcome characteristics for up to four test administrations.

Eligible studies appeared between 1921 and 2016 and only 5% were unpublished studies. It is worth noting that the overall sample consists of people of comparably low age (e.g., $M_{1.2} = 28.92$ years) and of general intelligence above average (e.g., $M_{1.2} = 0.69$ standard deviations). Identical test forms were administered almost twice as often as alternate ones, and information on rule explanation and practice items was provided only rarely by eligible studies. Also, standardized general intelligence of the sample was unavailable from most of the studies (72%). Interestingly, almost half of the samples were control groups from intervention studies. A few studies with large sample sizes were included in the analysis, which explains the high standard deviation of this variable.

Cognitive ability tests that were used by the studies and the categorization of outcomes into the BIS model (Jäger, 1982) are presented in the Appendix (Table A2). For the $o = 786$ outcomes, 137 different tests or test batteries were observed. The five tests most commonly administered were the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1958; 12.90%), the Wechsler Memory Scale (WMS; Wechsler, 1987; 8.30%), different versions of Raven's Matrices (Raven, 1936; Raven, 1962a; Raven, 1962b; 5.10%), the Differential Aptitude Test (DAT; Bennett, Seashore, & Wesman, 1990; 3.36%) and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, Tierney, Mohr, & Chase, 1998; 3.00%). All eligible studies, along with their main features, tests and resulting effect sizes, are available in Online Supplement 1.

3.2. The retest effect

Table 1 encloses descriptive values for observed effects sizes. One outcome was excluded from the analysis because it was identified as an outlier by influence diagnostics and visual data inspection ($SMCR_{1.2} = -3.06$; Buschkühl, 2007). The main results changed only marginally due to the exclusion, as this study contained $n = 8$ participants only.

Table 2
Retest effects.

Comp	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	*p*	α_{adj}	τ	ρ
1.2	122	172	566	153,185	0.327	0.024	[0.278, 0.375]	< 0.001*	0.007	0.317	0.998
1.3	31	50	181	23,977	0.495	0.045	[0.406, 0.585]	< 0.001*	0.008	0.402	
1.4	10	12	39	1422	0.516	0.058	[0.401, 0.631]	< 0.001*	0.010	0.474	
2.3					0.169	0.035	[0.010, 0.239]	< 0.001*	0.013		
3.4					0.021	0.021	[−0.020, 0.062]	0.318	0.050		
1.2 vs 2.3					0.158	0.040	[0.078, 0.238]	< 0.001*	0.017		
2.3 vs 3.4					0.148	0.031	[0.087, 0.209]	< 0.001*	0.025		

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; ρ = autocorrelation among true effects; $SMCR_{1.2 \text{ vs } 2.3} > 0$ indicates retest effects from first to second administration being larger than those from second to third.
* $p < \alpha_{adj}$.

Results from main analyses without the inclusion of moderators are presented in Table 2. It was possible to derive effect sizes for retest effects up to the fourth test administration, as only $m = 3$ studies conducted more than four test administrations.

Hypothesis 1a was supported: Significant retest effects of $SMCR_{1.2} = 0.33$ ($p < 0.001$) for the first test repetition and further increasing effects of $SMCR_{1.3} = 0.50$ ($p < 0.001$) and $SMCR_{1.4} = 0.52$ ($p < 0.001$) were found. This implies that retaking cognitive ability tests possibly leads to higher test scores. Nonetheless, the high standard deviation of the true effect (τ) indicates high heterogeneity of effects. To illustrate, for the effect $SMCR_{1.2} = 0.33$, the standard deviation of the true effect $\tau = 0.32$ leads to a prediction interval of $0.33 \pm 1.96 \times 0.32 = [-0.30, 0.96]$.

Also, gains between consecutive administrations were investigated, as they were expected to decrease with the number of tests (Hypothesis 1b). Between the second and the third administration, a significant gain of $SMCR_{2.3} = 0.17$ ($p < 0.001$) was observed. This gain was significantly smaller than the one between first and second administrations ($SMCR_{1.2} = 0.33$; $p < 0.001$). Between the third and the fourth administration, an effect of $SMCR_{3.4} = 0.02$ ($p = 0.318$) indicated no further gain. This effect was significantly smaller than $SMCR_{2.3} = 0.16$ ($p < 0.001$). Thus, Hypothesis 1b could be confirmed, as retest effects between consecutive tests decreased significantly with the number of administrations.

3.3. Moderator analyses

Results from moderator analyses are provided in Tables 4 to 9. Estimates for associations between moderators and significant control variables are displayed in Table 3.

3.3.1. Cognitive operations

Table 4 provides results from the moderator analysis regarding differences between cognitive operations. Differences of retest effects between cognitive operations were not significant with one exception comparing the third and fourth administration: Retest effects in processing speed tasks, which showed no gain in test scores ($SMCR_{3.4} = -0.07$), were smaller than in memory tasks ($SMCR_{3.4} = 0.11$; $p < 0.001$). Support for Hypothesis 2a, predicting processing speed tests to show the smallest retest effects followed by memory, divergent thinking and lastly reasoning tasks, was thus scarce. Note that results for divergent thinking tasks for more than three test administrations have to be interpreted carefully, as only $o = 2$ outcomes were observed.

Hypothesis 2b, assuming differences between the operations regarding the number of administrations necessary to reach a plateau, was not supported. Table 4 presents the according results in the rows indicated as 1.2 vs 2.3, and 2.3 vs 3.4. Here, gains between consecutive administrations were compared between operations. None of these

Table 3
Associations between moderators.

No. of administrations	Variable	Association					
		1.	2.	3.	4.	5.	6.
2	1. Operations	–					
	2. Test form	0.29	–				
	3. Test-retest interval (weeks)	0.01	−0.25***	–			
	4. Age (years)	0.21	−0.18***	**0.51***	–		
	5. Intelligence (z)	0.07	0.02	**−0.46***	0.05	–	
	6. Test contents	**0.45**	0.23	0.00	0.01	0.07	–
	7. Year of publication	0.29	−0.06	0.03	0.20***	−0.14	0.20
3	1. Operations	–					
	2. Test form	0.38	–				
	3. Test-retest interval (weeks)	0.12	**−0.41***	–			
	4. Age (years)	0.13	−0.26***	**0.71***	–		
	5. Intelligence (z)	0.07	**0.60***	**−0.79***	−0.29*	–	
	6. Test contents	0.40	0.26	0.10	0.06	0.12	–
4	1. Operations	–					
	2. Test form	**0.53**	–				
	3. Test-retest interval (weeks)	0.27	−0.31	–			
	4. Age (years)	0.13	0.13	**0.69***	–		
	5. Intelligence (z)	0.11	**0.71***	0.05	**−0.97***	–	
	6. Test contents	**0.53**	**0.62**	0.27	0.39	0.72	–

Note. Pearson's and pointbiserial correlations, η and *Cramers V* were computed; high values (correlations > |0.40| and $p < 0.001$; η > |0.40|; *Cramers V* > |0.40|) are printed boldly; for Pearson's and pointbiserial correlations.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

Table 4
Moderator analysis: subgroup analysis for cognitive ability operations.

Comp	Operation	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	*p*	Δ Operations						τ
										M		DT		R		
										p	α_{adj}	*p*	α_{adj}	*p*	α_{adj}	
1.2	PS	22	19	98	1497	0.367	0.044	[0.280, 0.454]	< 0.001	0.118	0.002	0.085	0.002	0.473	0.010	0.314
	M	26	35	191	21,479	0.283	0.045	[0.194, 0.373]	< 0.001					0.233	0.004	
	DT	11	17	34	2625	0.255	0.058	[0.141, 0.370]	< 0.001					0.140	0.002	
	R	50	68	207	8399	0.326	0.039	[0.248, 0.404]	< 0.001							
	G	16	30	35	119,185	0.416	0.079	[0.259, 0.573]	< 0.001							
1.3	PS	3	3	37	127	0.509	0.081	[0.348, 0.671]	< 0.001	0.451	0.008	0.194	0.003	0.338	0.005	0.402
	M	7	12	79	1232	0.422	0.081	[0.263, 0.582]	< 0.001					0.167	0.003	
	DT	3	9	13	2361	0.356	0.099	[0.161, 0.552]	< 0.001					0.092	0.002	
	R	14	18	41	1635	0.549	0.109	[0.333, 0.765]	< 0.001							
	G	4	8	11	18,622	0.678	0.089	[0.502, 0.854]	< 0.001							
1.4	PS	1	1	10	45	0.441	0.092	[0.260, 0.625]	< 0.001	0.248	0.004	0.385	0.006	0.074	0.002	0.439
	M	2	3	16	586	0.531	0.088	[0.356, 0.706]	< 0.001					0.167	0.002	
	DT	1	1	2	36	0.482	0.112	[0.260, 0.704]	< 0.001					0.092	0.002	
	R	3	3	6	191	0.710	0.163	[0.388, 1.032]	< 0.001							
	G	3	3	5	564	0.665	0.097	[0.474, 0.857]	< 0.001							
2.3	PS					0.142	0.061	[0.021, 0.263]	0.022	0.976	0.050	0.678	0.017	0.244	0.004	
	M					0.139	0.059	[0.023, 0.256]	0.020					0.229	0.003	
	DT					0.101	0.080	[−0.057, 0.259]	0.208					0.172	0.003	
	R					0.223	0.101	[0.023, 0.423]	0.029							
	G					0.262	0.030	[0.203, 0.321]	< 0.001							
3.4	PS					−0.068	0.030	[−0.128, −0.008]	0.028	< 0.001*	0.001	0.005	0.001	0.039	0.002	
	M					0.109	0.025	[0.060, 0.157]	< 0.001					0.322	0.005	
	DT					0.126	0.072	[−0.017, 0.268]	0.084					0.404	0.007	
	R					0.160	0.117	[−0.072, 0.393]	0.173							
	G					−0.012	0.013	[−0.037, 0.013]	0.327							
1.2 vs 2.3	PS					0.226	0.069	[0.090, 0.361]	0.001	0.223	0.003	0.097	0.002	0.388	0.006	
	M					0.144	0.067	[0.011, 0.277]	0.035					0.167	0.002	
	DT					0.154	0.098	[−0.041, 0.349]	0.120					0.092	0.002	
	R					0.103	0.108	[−0.110, 0.317]	0.339							
	G					0.154	0.080	[−0.005, 0.313]	0.058							
2.3 vs 3.4	PS					0.210	0.061	[0.089, 0.331]	0.001	0.028	0.002	0.035	0.002	0.194	0.003	
	M					0.030	0.061	[−0.091, 0.152]	0.621					0.848	0.025	
	DT					−0.025	0.123	[−0.268, 0.219]	0.842					0.676	0.013	
	R					0.062	0.158	[−0.252, 0.375]	0.698							
	G					0.274	0.032	[0.210, 0.339]	< 0.001							

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; Δ = difference between; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; PS = processing speed; M = memory; DT = divergent thinking; R = reasoning; G = general intelligence.
* $p < \alpha_{adj}$.

comparisons was significant, suggesting that a plateau is reached equally fast for all cognitive ability operations.

As depicted in Table 3, cognitive ability operations were associated with cognitive ability contents for studies with two and four administrations: Certain operations are more often tested by certain contents. For example, reasoning tests contained figural contents more often ($o = 143$) than divergent thinking tasks ($o = 3$), which in turn more commonly used verbal contents ($o = 26$). For studies with four administrations, the use of alternate test forms depended on the cognitive ability operation, so that identical test forms were used more often in all kinds of cognitive ability tasks except for memory tasks, which used alternate and identical test forms equally often.

3.3.2. Test form

Alternate test forms showed significantly smaller retest effects than identical test forms comparing first to second and first to third test administrations (see Table 5). The effects significantly differed by $\Delta SMCR_{1.2} = 0.15$ ($p < 0.001$) and $\Delta SMCR_{1.3} = 0.20$ ($p < 0.001$), which speaks in favor of Hypothesis 3. For further comparisons, the difference between test forms was not significant. It is worth noting that the difference in retest effects between identical and alternate test forms was significantly smaller for first to fourth administrations when compared to the difference in first to second ($p = 0.009$) and first to third administrations ($p = 0.025$). This suggests that the moderating effect of test form diminishes with the number of test administrations.

The use of alternate versus identical test forms was associated with participants' general intelligence in studies with three and four administrations: Samples with higher intelligence were more often tested with alternate test forms. In studies administering three tests, alternate test forms were administered more often in studies with shorter test-retest intervals. Further confounds for studies with four administrations were observed with regard to cognitive ability operations (as described above) and contents. In, e.g., figural tests, identical test forms were used most often, whereas in verbal tests, alternate test forms were more common.

3.3.3. Test-retest interval

Table 6 contains results from meta-regressions analyzing the influence of the test-retest interval on retest effects. The amount of time between administrations had a significant influence on the retest effect comparing first to second, first to fourth, and third to fourth administrations, as assumed by Hypothesis 4. The retest effect was observed to decrease by $\beta_{1.2} = -0.0008$ per week (with an intercept of $SMCR_{1.2} = 0.34$). To illustrate, increasing the test-retest interval by 50 weeks would lead to a decrease of the retest effect by $0.0008 \times 50 = 0.04$ standard deviations. Hence, this can be considered a weak effect. Another interpretation would imply the disappearance of retest effects for one test repetition after about 426 weeks, or 8.19 years. Comparing first to third, and second to third administrations, the length of the test-retest interval did not seem to have

Table 5
Moderator analysis: subgroup analysis for test form.

Comp	Test form	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	*p*	Δ Test forms		τ
										p	α_{adj}	
1.2	I	79	116	409	146,173	0.372	0.028	[0.317, 0.427]	< 0.001	< 0.001*	0.010	0.309
	A	43	56	156	7012	0.226	0.031	[0.165, 0.288]	< 0.001			
1.3	I	23	38	130	23,165	0.548	0.045	[0.458, 0.637]	< 0.001	0.008*	0.013	0.394
	A	8	12	50	812	0.347	0.084	[0.180, 0.514]	< 0.001			
1.4	I	8	10	28	1375	0.563	0.058	[0.449, 0.677]	< 0.001	0.354	0.050	0.471
	A	2	2	11	47	0.547	0.051	[0.447, 0.648]	< 0.001			
2.3	I					0.175	0.034	[0.108, 0.243]	< 0.001	0.234	0.025	
	A					0.120	0.080	[−0.038, 0.279]	0.136			
3.4	I					0.015	0.018	[−0.021, 0.052]	0.402	0.024	0.017	
	A					0.201	0.078	[0.047, 0.355]	0.011			

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; Δ = difference between; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; I = identical; A = alternate.
* $p < \alpha_{adj}$.

Table 6
Moderator analysis: meta-regressions for test-retest interval (weeks).

Comp	Regression component	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	Significance			τ
									p (Int)	*p* (β)	α_{adj}	
1.2	Int	117	167	549	150,538	0.3406	0.0294	[0.2824, 0.3988]	< 0.001			0.3146
	β					−0.0008	0.0003	[−0.0013, −0.0002]		0.0059*	0.017	
1.3	Int	31	50	180	23,977	0.5126	0.0577	[0.3983, 0.6269]	< 0.001			0.3986
	β					−0.0005	0.0003	[−0.0010, 0.0001]		0.0451	0.025	
1.4	Int	10	12	39	1422	0.6462	0.0696	[0.5084, 0.7841]	< 0.001			0.4592
	β					−0.0030	0.0008	[−0.0046, −0.0015]		< 0.001*	0.010	
2.3	Int					0.1720	0.0484	[0.0761, 0.2679]	< 0.001			
	β					0.0003	0.0003	[−0.0003, 0.0008]		0.3012	0.050	
3.4	Int					0.1336	0.0354	[0.0635, 0.2038]	< 0.001			
	β					−0.0025	0.0007	[−0.0040, −0.0010]		< 0.001*	0.012	

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; Int = Intercept; β = β-weight.
* $p < \alpha_{adj}$.

an influence. Therefore, Hypothesis 4 was supported only partly. Interestingly, $\beta_{1.4}$ differed significantly from $\beta_{1.2}$ ($p = 0.003$), $\beta_{1.3}$ ($p = 0.001$), and $\beta_{2.3}$ ($p = 0.002$), suggesting a larger influence of the test-retest interval for the fourth test administration than for earlier ones.

Time interval was associated with participant age, so that studies with older samples tended to have longer test-retest intervals. Studies with two and three administrations administered longer test-retest intervals more often in samples with lower general intelligence. Finally, studies with three administrations used shorter test-retest intervals more often in alternate test forms than in identical ones.

3.3.4. Age

A significant negative influence of age on retest effects was found only when comparing third and fourth administrations ($\beta_{2.3} = -0.003$; $p = 0.008$; see Table 7). To illustrate, repeating a test four times leads to a decrease in the gain from third to fourth administration of $0.003 \times 10 = 0.03$ per decade of age. Thus, Hypothesis 5 was supported only partly. However, there was a trend toward a negative effect of age on the size of the retest effect for all comparisons of administrations.

As described above, age and retest interval are confounded, as

Table 7
Moderator analysis: meta-regressions for age (years).

Comp	Regression component	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	Significance			τ
									p (Int)	*p* (β)	α_{adj}	
1.2	Int	107	152	538	87,611	0.398	0.053	[0.293, 0.503]	< 0.001			0.332
	β					−0.003	0.001	[−0.006, 0.001]		0.018	0.017	
1.3	Int	29	48	179	23,903	0.624	0.118	[0.390, 0.858]	< 0.001			0.408
	β					−0.005	0.003	[−0.010, 0.001]		0.041	0.025	
1.4	Int	9	11	38	1398	0.776	0.147	[0.486, 1.068]	< 0.001			0.473
	β					−0.007	0.003	[−0.014, −0.001]		0.014	0.013	
2.3	Int					0.226	0.099	[0.031, 0.422]	0.024			
	β					−0.002	0.002	[−0.006, 0.003]		0.215	0.05	
3.4	Int					0.153	0.059	[0.035, 0.270]	0.011			
	β					−0.003	0.001	[−0.005, −0.001]		0.008*	0.010	

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; Int = Intercept; β = β-weight.
* $p < \alpha_{adj}$.

Table 8
Moderator analysis: meta-regressions for intelligence (z-scaled).

Comp	Regression component	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	Significance			τ
									p (Int)	*p* (β)	α_{adj}	
1.2	Int	23	37	156	2142	0.278	0.041	[0.192, 0.364]	< 0.001			0.335
	β					0.084	0.055	[−0.032, 0.201]		0.073	0.010	
1.3	Int	6	11	62	1194	0.450	0.096	[0.248, 0.652]	< 0.001			0.322
	β					0.053	0.122	[−0.205, 0.311]		0.335	0.025	
1.4	Int	3	3	19	105	0.367	0.184	[−0.021, 0.756]	0.062			0.305
	β					0.159	0.113	[−0.080, 0.397]		0.089	0.013	
2.3	Int					0.172	0.091	[−0.020, 0.364]	0.076			
	β					−0.031	0.125	[−0.295, 0.233]		0.403	0.050	
3.4	Int					−0.082	0.108	[−0.437, 0.272]	0.630			
	β					0.106	0.108	[−0.123, 0.334]		0.171	0.017	

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; * = $p < \alpha_{adj}$; Int = Intercept; β = β-weight.

eligible studies with older samples often use longer test-retest intervals. In studies with four administrations, a negative relationship of age and general intelligence was found.

3.3.5. Intelligence

Table 8 contains results from a meta-regression analyzing the moderating effect of intelligence on retest effects. Hypothesis 6 was not supported, because general intelligence was not a significant moderator of the size of retest effects for any of the comparisons. Nonetheless, regression weights were directed in the expected, positive direction, although effects were small and not significant.

General intelligence of the samples was associated with time, test form and participant age. In studies with three and four administrations, alternate test forms were more often administered in more intelligent samples compared to identical forms. Shorter test-retest intervals were used in more intelligent samples in studies with two administrations. Younger samples of studies with four administrations tended to be more intelligent.

3.3.6. Test contents

We did not hypothesize differences between tests that used different kinds of contents. Yet, it was coded whether the cognitive ability test used figural, numerical, verbal, or several contents, and differences of retest effects between contents of the cognitive ability tasks were analyzed exploratively (Table 9). Comparing first to second administration, numerical tasks ($SMCR_{1.2}$ = 0.18) showed smaller retest effects than tasks using several kinds of contents ($SMCR_{1.2}$ = 0.41, $p < 0.001$), and tasks using verbal contents ($SMCR_{1.2}$ = 0.33, $p < 0.001$). Then,

Table 9
Explorative analysis: subgroup analysis for test contents.

Comp	Test content	*m*	*k*	*o*	*N*	*SMCR*	*SE*	95% CI	*p*	Δ Test contents						τ
										N		v		S		
										p	α_{adj}	*p*	α_{adj}	*p*	α_{adj}	
1.2	F	44	68	242	26,631	0.321	0.034	[0.254, 0.388]	< 0.001	0.003	0.002	0.824	0.017	0.032	0.002	0.315
	N	17	21	58	3304	0.175	0.038	[0.101, 0.249]	< 0.001			0.001*	0.002	< 0.001*	0.002	
	V	31	40	149	1680	0.331	0.040	[0.253, 0.410]	< 0.001					0.123	0.003	
	S	30	43	116	121,840	0.406	0.039	[0.329, 0.483]	< 0.001							
1.3	F	14	26	77	4242	0.518	0.054	[0.411, 0.625]	< 0.001	0.958	0.050	0.411	0.004	0.547	0.005	0.402
	N	2	2	10	173	0.523	0.099	[0.326, 0.720]	< 0.001			0.524	0.005	0.699	0.006	
	V	10	14	59	1215	0.443	0.087	[0.270, 0.615]	< 0.001					0.234	0.003	
	S	5	8	35	18,347	0.560	0.060	[0.441, 0.679]	< 0.001							
1.4	F	7	8	20	628	0.528	0.070	[0.390, 0.667]	< 0.001	–	–	0.868	0.025	0.758	0.013	0.439
	V	1	2	11	581	0.547	0.094	[0.361 +, 0.732]	< 0.001	–	–			0.715	0.008	
	S	2	2	7	213	0.504	0.077	[0.350, 0.657]	< 0.001	–	–					
2.3	F					0.197	0.046	[0.105, 0.289]	< 0.001	0.110	0.003	0.275	0.003	0.469	0.004	
	N					0.348	0.087	[0.176, 0.521]	< 0.001			0.033	0.002	0.019	0.002	
	V					0.111	0.071	[−0.029, 0.252]	0.120					0.606	0.006	
	S					0.154	0.041	[0.072, 0.235]	< 0.001							
3.4	F					0.011	0.038	[−0.065, 0.087]	0.781	–	–	0.043	0.002	0.018	0.002	
	V					0.104	0.012	[0.081, 0.127]	< 0.001	–	–			< 0.001*	0.002	
	S					−0.056	0.030	[−0.115, 0.002]	0.058	–	–					
1.2 vs 2.3	F					0.124	0.061	[0.004, 0.245]	0.044	0.007	0.002	0.295	0.004	0.080	0.003	
	N					−0.173	0.090	[−0.352, 0.005]	0.057			< 0.001*	0.002	< 0.001*	0.002	
	V					0.220	0.075	[0.072, 0.369]	0.004					0.722	0.010	
	S					0.253	0.053	[0.149, 0.357]	< 0.001							
2.3 vs 3.4	F					0.186	0.061	[0.065, 0.308]	0.003	–	–	0.043	0.002	0.702	0.007	
	V					0.007	0.066	[−0.124, 0.138]	0.915	–	–			0.012	0.002	
	S					0.210	0.037	[0.136, 0.283]	< 0.001	–	–					

Note. Comp = comparison of administrations; *m* = number of studies; *k* = number of samples; *o* = number of outcomes; *N* = total sample size; *SMCR* = standardized mean change; *SE* = standard error; CI = confidence interval; Δ = difference between; α_{adj} = alpha adjusted according to Bonferroni-Holm; τ = standard deviation of the true effect; F = figural; N = numerical; V = verbal; S = several; no numerical contents were used for four test administrations.

* $p < \alpha_{adj}$.

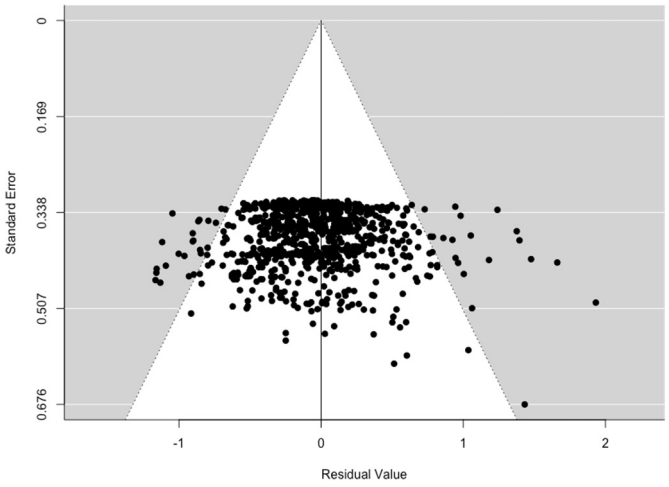

Fig. 3. Funnel plot for inspection of publication bias. Residuals of the main model without moderators are plotted against their standard errors. The symmetry of the plot suggests absence of publication bias.

comparing second to third administration, numerical tasks showed a larger gain ($SMCR_{2.3} = 0.35$) than comparing first to second administration ($SMCR_{1.2} = 0.18$). Although this difference between $SCMR_{1.2}$ and $SMCR_{2.3}$ was not significantly different from zero ($SMCR_{1.2vs2.3} = -0.17$, $p = 0.057$), it was still significantly different from the one observed in tests using verbal ($SMCR_{1.2vs2.3} = 0.22$, $p < 0.001$) and several contents ($SMCR_{1.2vs2.3} = 0.25$, $p < 0.001$). This is because for tests using verbal and several contents, larger gains were observed from first to second, whereas for those using numerical contents larger gains were observed from second to third administrations. Moreover, comparing third to fourth administration, tests using several contents ($SMCR_{3.4} = -0.06$) showed smaller retest effects than those using verbal contents ($SMCR_{3.4} = 0.10, p < 0.001$). Note that no numerical contents were used for more than three test administrations.

As mentioned above, cognitive ability task contents were associated with operations in studies administering two and in those with four administrations. For example, in studies administering four tests, tasks with several contents were processing speed tasks more often than they were, e.g., reasoning or memory tasks. Also, for studies with four administrations, identical test forms were used for figural tests and tests using several contents more often, whereas alternate forms were more commonly used for tests with verbal contents.

3.4. Publication bias

A funnel plot is presented in Fig. 3. Although methods to find publication bias have not been reported for complex meta-analytical models like the present one, a funnel plot can give an impression of the relationship between effect sizes and standard errors. For the present meta-analysis, no systematic relationship was indicated by the funnel plot.

Effect sizes from published studies were compared to those from unpublished ones by moderator analysis. For more than two administrations, two or less unpublished outcomes were observed, which is why the moderating influence of publication of the study was tested only when comparing first to second administration. Here, $m = 6$ eligible studies ($o = 26$, $k = 13$) were compared to $m = 116$ studies ($o = 539$, $k = 159$). Published studies ($SMCR_{1.2} = 0.33$) and unpublished studies ($SMCR_{1.2} = 0.19$) did not differ significantly in their retest effect sizes ($p = 0.095$).

With regard to the studies analyzed by this meta-analysis, it can be questioned whether publication bias was critical. This analysis was based on a high number of control groups from intervention studies (43%), whose goal it was to show that an intervention was effective. Thus, the studies did not seek to show large retest effects in control groups, and therefore publication bias with regard to the topic of this meta-analysis might not be critical here. This is underlined by the finding that effect sizes in control groups did not differ significantly from those in non-control groups. To sum up, it seems unlikely that this meta-analysis might be affected by publication bias.

4. Discussion

It was the goal of this multilevel meta-analysis to contribute to the field of retest effects and its determinants by updating and expanding previous meta-analyses applying a high number of new studies, methodological improvements and an extended focus on retesting multiple times in order to gain new insights by hypotheses based on comprehensive theoretical deliberations. Effect sizes for up to four test administrations were analyzed, expecting a non-linear growth of retest effects with the number of administrations. The influence of several moderators was investigated. Results from the high number of test repetitions and from moderator analyses for up to four administrations are of special interest; for most of the variables, this is, to the best of our knowledge, the first time these have ever been examined. In the following, key results will be summarized, limitations of this work will be pointed out and, finally, implications for practice and future research will be discussed.

4.1. Interpretation of results and theoretical implications

For the mere repetition of a cognitive ability test, an improvement of a third of a standard deviation was found. This is in line with the comparably large findings of the prior works of Hausknecht et al. (2007) and Calamia et al. (2012). To illustrate, a person that achieved an IQ score of 100 in a first test, would be predicted to attain 105 IQ points when retested with a second test. An increase in test scores of half a standard deviation for the third and fourth test were revealed. Taking the example from above, the person would be predicted to score 107 IQ points in a third and 108 in a fourth test. Nonetheless, in the light of the context of simply repeating a cognitive ability test, which occurs fairly often, this effect should not be neglected.

The observed retest effects can be explained by three groups of causes, as elaborated by Lievens et al. (2007): Firstly, the latent cognitive ability could be enhanced by retesting (Roediger & Butler, 2011), although this cause seems rather improbable (e.g., Hausknecht et al., 2002; Lievens et al., 2005; te Nijenhuis et al., 2007). Secondly, the reduction of distorting factors, such as motivation, test anxiety and unfamiliarity, might lead to an increase in test scores (Anastasi, 1981; Matton et al., 2009; Reeve & Lam, 2007; Reeve et al., 2009). Finally, the development of strategies and test-specific skills might lead to better test results as well (Allalouf & Ben-Shakar, 1998; Lievens et al., 2007; te Nijehuis et al., 2007).

According to the results, a plateau seems to be reached after the third test administration, as is often observed in primary studies (e.g., Albers & Höft, 2009; Bartels et al., 2010; Dunlop et al., 2011; Puddey et al., 2014). However, these findings contradict the findings from Kulik et al. (1984), who claimed a linear improvement with the number of administrations. Our findings confirm assumptions according to the *power law of practice* (Donner & Hardy, 2015; Jaber & Glock, 2013; Newell & Rosenbloom, 1981) describing diminishing gains over time after first improvements, which can be applied to retest effects in cognitive ability tests (Newell & Rosenbloom, 1981). The plateau can further be explained by a reduced influence of the causes of retest effects (Lievens et al., 2007) with the number of repetitions. The reduction of distorting factors probably takes place within the first test sessions, meaning the reduction of, e.g., situational test anxiety, is highest between the first administrations, then levels off afterwards so that no further substantial change takes place. The same idea applies for test-specific strategies or skills, as they might improve most within the first

sessions and not afterwards. For example, if a participant develops a strategy that helps him solve an item efficiently, it is most likely that he develops this strategy within the first or second test and then uses this strategy for later tests. Therefore, a decreasing influence of these two causes of retest effects with the number of test repetitions can explain that a plateau is reached after the third administration.

4.1.1. Cognitive operations

Cognitive ability tests were categorized according to the BIS (Jäger, 1982; Fig. 1), which differentiates between four cognitive operations: processing speed, memory, divergent thinking, and reasoning. Only one significant difference between cognitive ability operations was observed: Processing speed tasks showed no gain between third and fourth test, whereas a significantly larger gain was observed in memory tasks. Here, the low number of outcomes for which the analysis of multiple retests was carried out has to be considered. For example, reasoning operations were observed in six outcomes only, leading to broad confidence intervals and high standard errors suggesting a low statistical power of this analysis. Also, confounding of moderators has to be considered. Cognitive ability operations were associated with test contents and test forms, which makes it difficult to segregate results and deduce conclusions concerning which moderator led to the differences in retest effects.

However, when retesting only once, it is likely that retest effects are equally large between cognitive operations. When administering processing speed tasks multiple times, smaller retest effects can possibly be expected, reaching a plateau after three tests. When retesting multiple times using memory and possibly also reasoning tasks, larger retest effects have to be taken into account and it might take more administrations to reach a plateau. These results stress the importance of a differentiation between cognitive ability operations, especially when retesting multiple times.

Yet, results only partly support the assumption that larger retest effects occur for tests with higher complexity (e.g., Randall & Villado, 2017), as differences were only observed when retesting four times. It seems that the complexity of the task only plays a role when tests are administered multiple times. One possible reason for this finding could be that for tests using different cognitive ability operations, similar mechanisms that cause retest effects take place and that these mechanisms change in a similar way when retesting only a few times: For each operation, the contribution of distorting factors and test-specific skills and strategies that cause retest effects seems to decrease in a comparable way. Only after a few repetitions, does there seem to be differences in the degree to which the causes have an influence on the effects. For processing speed tasks, the causes might have less influence more quickly. Taking rule incomprehension as one distorting factor that is reduced by retesting, after the first test administration, rule comprehension might increase similarly for all tasks. Then, after a few more administrations, rule comprehension is already fully achieved for processing speed tasks, whereas for reasoning tasks, this might take longer to obtain. Hence, results could be explained in such a way that similar mechanisms that cause retest effects take place between operations when retesting only a few times, whereas differences might become obvious only when retesting multiple times.

Another explanation why cognitive ability operations might not differ in their retest effects when retesting only a few times can be put forward when integrating our theoretical approach with a different view on intelligence and how cognitive ability operations can be categorized. According to Carroll's (1993) three-stratum hierarchical factor model, processing speed and memory are lower cognitive functions, or less *g*-loaded, whereas reasoning operations are declared to be higher order functions, or higher *g*-loaded. The relationship between *g*-loadedness of a test and its retest effects have been investigated by, e.g., te Nijenhuis et al. (2007) and Olenick, Bhatia, & Ryan (2016), who are mainly arguing for smaller retest effects in higher *g*-loaded tasks. They claim that, because *g* is a stable latent ability, highly *g*-loaded tests are harder to deter, which leads to smaller retest effects (Olenick et al., 2016). As becomes obvious, this argumentation contradicts the one that is put forward in the derivation of the hypotheses of the current meta-analysis: According to this view, operations that are less *g*-loaded, like processing speed, were expected to show larger retest effects, whereas we expected them to show smaller effects. Because the BIS does not differentiate between higher and lower order functions (Jäger, 1982), this differentiation was not made when deriving hypotheses. Nonetheless, both theoretical approaches taken together might explain the absence of differences when retesting only a few times: On the one hand, distorting factors and test-specific strategies could be developed faster in less complex or less *g*-loaded tasks, on the other hand, performance is easier to deter. These two mechanisms might counterbalance each other and therefore no differences between operations could be observed.

4.1.2. Test Form

Substantial differences between alternate and identical test forms were observed, which is in line with findings from, e.g., Arendasy and Sommer (2017), Hausknecht et al. (2007), Freund and Holling (2011), and Kulik et al. (1984). Test-specific strategies and skills seem to be easier to develop for identical test forms (Cook & Campbell, 1979; Nevo, 1976) and distorting factors like rule incomprehension and situational test anxiety might be more easily reduced by identical test forms more easily (Anastasi, 1981; Arendasy & Sommer, 2017).

Interestingly, the moderating effect of the equivalence of test forms remained significant only until the third test administration. On the one hand, the confounding of test forms and general intelligence of the samples might be important here, as studies administering four tests used alternate test forms in samples with higher IQ more often than identical test forms. This might have led to an overestimation of the retest effect in alternate test forms for the fourth test administration. On the other hand, this finding can be explained by the assumption that the influence of the mechanisms that cause retest effects (Lievens et al., 2007) might decrease with the number of test repetitions. This would lead to smaller differences in retest effects that are due to the equivalence of test forms with the number of administrations. For example, when repeating a test four times, a test-specific strategy might be fully developed regardless of whether identical or alternate test forms were administered. It might take a few administrations to develop strategies when alternate test forms are used, but then, this might be achieved in an equal way compared to identical test forms. Also, the reduction of distorting factors and construct-irrelevant variables might, at first, proceed faster when identical test forms are administered, but after one or two tests, an equal level might be reached also when alternate test forms are used. Thus, a reason for these findings could be that only for the first and maybe second tests, differences between mechanisms that cause retest effects exist when alternate versus identical test forms are administered.

4.1.3. Test-retest interval

The length of the test-retest interval had a significant influence on retest effects for up to four test administrations: The longer the test-retest interval between administrations, the smaller the retest effects. This finding is in line with previous research (Calamia et al., 2012; Hausknecht et al., 2007; Salthouse et al., 2004), though no significant influence was found for the third test administration. By inspecting Table 1, it becomes obvious that for the third test administration, longer test-retest intervals were used by eligible studies than for the second or fourth ones, which might underestimate the influence of this moderator for the third test. Also, the confounding between test-retest interval and test forms for studies with three administrations can explain why no effect was found for the third administration. Alternate test forms, which can cause smaller effects, were administered with shorter test-retest intervals than those administered for identical test forms. This might have led to an underestimation of effect sizes with short test-

retest intervals for the third test administration.

Nonetheless, the effect of the test-retest interval on the retest effect can be explained by an enforcement of the causes of retest effects (Lievens et al., 2007) by shorter intervals between tests. The longer the test-retest interval, the less test-specific skills (e.g., strategies, rules of the test, correct answers to items), can be recalled. Because of a longer test-retest interval, construct-irrelevant factors, such as familiarity with the test, can decrease as well. Although the influence of the rest-retest interval on the size of the retest effect is rather small, it increases with the number of administrations, meaning the more administrations are conducted, the higher the influence of the test-retest interval. Thus, results imply that a general conclusion that the influence of mechanisms that cause retest effects decreases with the number of retests, might not be valid. Indeed, these results imply that a differentiation between these mechanisms is important with regard to the length of the test-retest interval. For example, the length of the test-retest interval might affect item-specific memory effects more strongly and sustainably than it does, e.g., test anxiety. If test-retest intervals are short, item-specific memory effects would maybe not decrease with the number of test repetitions, because the recognition of correct answers to test items very much depends on the test-retest interval. In order to make use of this strategy, the whole item and the exact solution has to be recalled, which is harder in the case of a long test-retest interval. This could lead to larger retest effects, even when retesting multiple times with short test-retest intervals.

4.1.4. Age

Results only partially support the moderating effect of participant age on the retest effect, although meta-analytic regression weights were directed in the expected direction. Thus, our results did not confirm those from primary (e.g., Van Iddekinge et al., 2011; Schleicher et al., 2010) and meta-analytic studies (Calamia et al., 2012) which found age to lower retest effects. Note that for the second and third administrations, the mean age of the total sample was low (see Table 1). Thus, results from the fourth test administration might be more representative for a general population of healthy adults in this regard.

For up to three test administrations, age did not seem to moderate retest effects. The development of test-specific strategies and skills might therefore be independent of age for the up to three test administrations. The absence of the influence of age on the retest effect can also be due to an equal development of distorting and construct-irrelevant factors (Lievens et al., 2007), which can be seen independently from age. These variables might have changed in an equal way between age groups, leading to similar retest effects.

Though, with the number of retests, participant age seems to become more important, as the gain between third and fourth test was moderated by age. As expected, older participants showed slightly smaller retest effects. This can possibly be explained by younger participants establishing test-specific strategies more quickly than older participants. Younger participants were shown to have a higher capacity to maintain and update information (Braver & Barch, 2002) and can be expected to learn from prior test experience more than older participants (Van Iddekinge et al., 2011). This might, over several test administrations, lead to a faster development of strategies.

4.1.5. Intelligence

Intelligence did not influence the retest effect significantly, although the direction of the meta-analytic β weight was directed into the expected direction. Due to the high number of missing values (72%) that resulted from eligible studies not reporting a standardized value for their participants' intelligence, only 156 outcomes could be included in the meta-regression analyzing the influence of intelligence on retest effects. This resulted in large confidence intervals for β weights, indicating that this analysis might lack statistical power. Also, average intelligence of the total sample was above population average (see Table 1). In these samples, there might be not enough variability present to investigate the influence of this variable.

On the other hand, these results can be interpreted as intelligence not influencing the size of retest effects, as also found by, e.g., Bartels et al. (2010) and Coyle (2005), which contradicts findings from, e.g., Randall et al. (2016), and Salthouse and Tucker-Drob (2008). For all levels of intelligence, similar retest effects might be possible, because mechanisms causing retest effects might apply for people with all levels of intelligence in a similar way. Test-specific strategies and skills might increase and distorting factors (Lievens et al., 2007) might decrease independently from participants' intelligence.

4.1.6. Test contents

Contents of cognitive ability tests were categorized according the BIS (Jäger, 1982; Fig. 1). The use of figural, numerical, verbal contents and of more than one of these within one test (several) were distinguished. An explorative analysis using cognitive ability contents as a control variable was conducted and differences between tests using different contents were found. This result was not in line with most prior works on the topic, as Hausknecht et al. (2007), for example, did not find differences in retest effects between quantitative and verbal tests, although some studies argue in favor of an important role that the type of content takes (Benedict & Zgaljardic, 1998; Salthouse & Tucker-Drob, 2008; Villado et al., 2016). In the current meta-analysis, for retest effects from the first to second administration, tests using numerical contents showed significantly smaller effects than those using verbal and several contents. Comparing gains from the first to second and second to third administration between contents, tests with numerical contents seemed to compensate this by showing larger gains from second to third test than tests using several or verbal contents. Unfortunately, for more than three tests, we did not find any study that used numerical contents. Benedict and Zgaljardic (1998) and Salthouse and Tucker-Drob (2008) argue that differences between contents can be plausible because some contents are more familiar than others. A lower familiarity with the content might lead to smaller retest effects, because it might be harder to memorize items. Villado et al. (2016) argue that unfamiliar item-types or contents might demand more cognitive resources than familiar ones and that for this reason, they lead to smaller retest effects. It is argued that numerical contents are remembered less well because numbers do not have a meaning when presented alone (e.g., Venkatesan, Lancaster, & Kendall, 1986). In contrast, verbal contents necessarily contain semantic information, which leads to higher memorability of verbal stimuli (Huber, 1980). For numerical contents, which might be harder to memorize than verbal and complex kinds of contents, the influence of the distorting factor of unfamiliarity (Freund & Holling, 2011; Lievens et al., 2007) could be lower than for other contents in a second test: There seemed to be less increase in familiarity of the items compared to other contents. When retesting three times, however, familiarity of numerical items was even higher, leading to retest effects of similar size to those in items using different contents. Thus, the reduction of distorting factors like unfamiliarity seems to follow different mechanisms between contents. Familiarity and memorability of the stimuli might play a key role in this regard (Benedict & Zgaljardic, 1998; Freund & Holling, 2011; Salthouse & Tucker-Drob, 2008).

From the third to fourth administration, tests using several contents showed no further gain, whereas retest effects in those using verbal contents still seemed to increase after three administrations. Before that, no differences of retest effects were observed between these contents. This could be due to a confounding between contents and cognitive operations, because tasks using several contents were mostly processing speed tasks in the fourth administration, for which smaller effects were found as well. Thus, the effect for tests using several contents might be underestimated. With regard to familiarity of item types (Benedict & Zgaljardic 1998; Salthouse & Tucker-Drob, 2008; Villado et al., 2016), verbal contents might be even easier to memorize than several different ones (e.g., Childers & Viswanathan, 2000, Huber, 1980; Venkatesan et al., 1986).

4.2. Limitations

A few shortcomings have to be admitted that meta-analysis cannot resolve. Firstly, this analysis was limited to those moderators for which enough information was given by the primary studies. Randall and Villado (2017) suggest numerous factors that might have an influence on retest effects but for which evidence is scarce. Meta-analysis depends on the given information from primary studies, which is why we did not analyze moderators such as personality variables, motivation, emotions, although results would have surely given interesting theoretical insights. Next, only very few studies reported any information on variables such as general intelligence, feedback, practice items or rule explanation, which made it difficult to analyze and interpret these as moderators. Finally, the confounding of moderators hinders a straightforward interpretation of moderator effects in some cases. It is not possible to conclude if, e.g., test-retest interval or age was the moderator that contributed to the size of the retest effects most because shorter test-retest intervals were administered more commonly in younger samples. Such confounding of moderators represents the characteristics of eligible studies that co-occur more often than others. It is important to note that confounding does not represent a causal relationship between these variables.

The variance of the random effects indicated a high heterogeneity of retest effects. This is of special interest because none of the prior meta-analyses examined the variance of the effects. Moderator analyses were able to explain retest effects only to a low degree. This means, there could be alternative factors that contribute to the size of the effects that have not yet been considered. Examining those studies with large and small effect sizes in detail may give hints on further moderator variables that explain some of the variance. Studies that have large effect sizes are, e.g., Pereira, Costa, and Cerqueira (2015; $SMCR_{1.2} = 1.70$), Lievens et al. (2007; $SMCR_{1.2} = 1.57$), and Albers and Höft (2009; $SMCR_{1.4} = 2.45$). Albers and Höft's study took place in a selection setting with a highly motivated, young, and therefore selective sample. Ten speeded tests of spatial ability were administered directly after each other. Thus, the large effects of this study can to a certain degree be explained by the moderators suggested by this meta-analysis. Nonetheless, motivation of the sample and time limit of the test could be further moderators that have not been investigated meta-analytically yet. These variables have also been suggested by Randall and Villado (2016) and are, however, characteristics of other studies with large effect sizes as well (e.g., Bartels et al., 2010; Lievens et al., 2007). Negative effect sizes, on the other hand, can be observed in studies such as Robinson, Leach, Owen-Lych and Siinram-Lea (2013; $SMCR_{1.2} = -0.84$), Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott (2007; $SMCR_{1.2} = -0.72$), and Colom et al. (2012; $SMCR_{1.2} = -0.60$). Interestingly, for the following test repetitions, effect sizes increased again for most of these observations. Negative effects could possibly be explained by a low motivation of samples, as might be the case in laboratory setting rather than selection settings. However, context was not a significant control variable in the current analysis, which suggests that motivation should be analyzed separately from the study context.

4.3. Impact

4.3.1. Practice

A gain in test scores of almost a third standard deviation for a mere repetition of a cognitive ability test without any further intervention between measurements has an impact on how we should deal with retesting and test experience in the future. Indeed, the increase of score gains up to the third test underlines the importance. Applicants, for example, could reach much higher scores when they take a test another time. Clinical patients could reach much higher scores if they answer the same neuropsychological test battery twice, without actually being recovered, but because they have taken the test before. In fact, this score gain might even change a diagnosis because a cutoff-value could be exceeded that had not been reached before. We cannot be sure if a score from a repeated or initial test is more representative for the person's latent ability (Lievens et al., 2005; te Nijenhuis et al., 2007), which can lead to false decisions based on test results in all kinds of contexts (Randall & Villado, 2017).

With regard to moderator analyses, determinants of retest effects should be considered in applied settings, as retest effects can be expected to differ between test-retest intervals, cognitive ability tests, and equivalence of test forms. Several approaches have been suggested to reduce effect sizes of retest effects that are mostly based on determinants of retest effects (e.g., Arendasy & Sommer, 2017). In order to control for retest effects, the recommendation of using parallel instead of identical test forms was supported by our results (Arendasy & Sommer, 2017), as long as only a few retests are administered. However, the finding that the administration of alternate test versions also leads to retest effects suggests that any earlier experience with a similar cognitive ability test can account for higher scores, which should be taken seriously as well. Especially in personnel selection settings, test experience is an important factor accounting for test fairness and equal treatment of all applicants (e.g., Hausknecht et al., 2007). As the internet offers numerous possibilities to practice cognitive ability tests, an extensive preparation of tests can be done easily by anyone from their home computers. Applicants that choose to prepare themselves extensively might therefore have advantages in the selection procedure, which the hiring institution is unable to control for. Thus, it is recommended to always inquire test experience from a test subject (Freund & Holling, 2011), regardless of whether it is a research or practice context. A longer test-retest interval between administrations and the use of numerical contents in tests are further strategies to prevent retest effects, for which this meta-analysis found support.

In contrast to the goal of preventing retest effects, it might be beneficial for applicants in selection settings to take advantage of retest effects and improve their test performance on an aptitude test (e.g., Lievens et al., 2005). On the one hand, there are tasks for which it can be worthwhile training for before taking an application test. On the other hand, there are tests, for a preparation might not be very efficient. It was found that processing speed tasks show smaller retest effects over several repetitions. In contrast, for reasoning and memory tasks, practicing a few times can be more efficient. Also, verbal contents seem to be promising to train, and numerical contents should probably be practiced more than twice, as score gains seem to be achieved later for these kinds of tasks. However, repeating a test more than three times does not seem to be beneficial as score gains appear to remain at a certain level from this point.

Adverse impact of retesting against older people, as suggested to be problematic by Van Iddekinge et al. (2011), was not found to be an issue. It is, in fact, an encouraging result that age and general intelligence were not significant moderators for the first repetition. Based on the results, smaller retest effects in older people or those with an intelligence below average do not necessarily have to be expected. However, when retesting multiple times though, younger people might profit more.

Further, our findings stress the importance of control groups in intervention evaluations. Especially regarding growing effects with further repetitions, control groups are equally important for pre to post comparisons as for follow-up evaluations. The use of alternate test versions for both experimental and control groups can possibly decrease the influence of retest effects.

4.3.2. Future research

In this meta-analysis, the existence of retest effects in cognitive ability tests was shown to be reasonable and its size and a few moderators could be identified, despite that high variance has to be admitted. Although it was not the goal of the meta-analysis to explore causes of retest effects, this is a very important question. Distorting

factors like test anxiety and motivation, and test-specific skills and strategies seem to play a crucial role, as related hypotheses found approval. We outlined the causes of retest effects above (Freund & Holling, 2011; Lievens et al., 2007), but, to date, only few studies yield at explaining mechanisms to cause retest effects directly (Randall & Villado, 2017), and at how their influence changes over the course of multiple test repetitions. In this meta-analysis, hypotheses that were based on the assumption that the causes' influence decreases with the number of repetitions mostly found approval. This approach, however, needs more solid direct investigation in order to be confirmed.

Our results suggest that there might be differences in the mechanisms causing retest effects between cognitive ability operations, test contents, test forms, and different levels of participant intelligence and age, especially when retested multiple times. Future research should address determinants of these mechanisms in detail, as only very few studies examine mechanisms causing retest effects in multiple retests. For example, results suggest that for processing speed tasks, test-specific skills and strategies increase faster and that distorting factors decrease faster than in other cognitive ability tasks, when retested multiple times. This assumption needs to be investigated directly by measuring the development of these mechanisms during retesting. In this regard, it should be considered that it might, at the same time, be more difficult to deter complex and highly *g*-loaded tasks (te Nijenhuis et al., 2007; Olenick et al., 2016). Also, results regarding the influence of the test-retest interval imply that these mechanisms should be investigated with a high level of differentiation, as different factors might by triggered with different emphasis by different moderators. One explanation for the results is that some mechanisms, such as item-specific memory (Nevo, 1976), rely on the test-retest-interval more heavily than others when retested multiple times. With regard to participant intelligence and age, results could be explained by the reduction of distorting and construct-irrelevant factors (e.g., anxiety, familiarity) being independent from these sample variables. It needs to be investigated whether this assumption holds. These mechanisms might rely on other differential factors, such as personality variables (Randall & Villado, 2017). Also, the assumption that younger participants might develop strategies faster than older ones when retested multiple times might give further insights into how age moderates retest effects (Van Iddekinge et al., 2011; Villado et al., 2016).

Several moderators that might be determinants of the size of the retest effect in cognitive ability tests were outlined in the limitations subsection above and were reviewed by Randall and Villado (2017). These moderators should be investigated in future research, also with regard to multiple retests. Factors like personality, motivation or prior experience with cognitive ability tests might give further insight. Another way to find possible alternative moderators, is to analyze related fields of research. In the area of the testing effect for memory tasks (Roediger & Butler, 2011), several further moderators were suggested that apply to their paradigm. Unfortunately, the paradigm did not fit the criteria of this analysis, which is why none of the studies investigating the testing effect were included in this meta-analysis. Nonetheless, moderators like feedback (which could not be tested as a control variable in this meta-analysis because too little information was available), number of retrieval errors and participants' confidence, as suggested by Roediger and Butler (2011), might determine the size of retest effects in other cognitive ability tests as well. In general, it is strongly recommended to give as much information as possible on the study design, sample and tests in order to be able to include such information in future meta-analyses on this or related topics.

Due to the restriction of including samples between 12 and 70 years of age, we can exclusively transfer our findings to this age group. Examining mere retest effects in the elderly or in children would contribute to the understanding of age differences. As illustrated in Fig. 2, a lot of studies were excluded because of too low or too high age of the samples. A meta-analytic study of retest effects in children or elderly would be helpful to address age differences in further detail.

As outlined above, only very few primary studies investigate retest effects over several administrations. In fact, 31 studies were found that applied our criteria and administered more than two tests. Undoubtedly, more primary research is needed to investigate retest effects for a high number of test repetitions and to evaluate determinants and causes of retest effects in this regard, because in many settings, tests are administered more than twice.

A relevant question for which evidence is still mixed (e.g., Allalouf & Ben-Shakar, 1998; Freund & Holling, 2011; Lievens, Buyse & Sackett, 2005; te Nijenhuis et al., 2007), is how and why retesting changes the validity of a test. Each of the three groups of causes of retest effects has their implications on validity changes due to retesting (Lievens et al., 2007). If causes and mechanisms leading to retest effects are understood in more detail, an improved comprehension of validity changes due to retesting can be achieved as well. Future studies should therefore focus on this and the aspects discussed above in the context of numerous test repetitions in order to fully comprehend the nature of retest effects and to prevent false decisions based on cognitive ability tests.

Acknowledgements

We thank Wolfgang Viechtbauer for providing valuable support concerning the data analysis during the revision process.

This work was supported by the Deutsche Forschungsgemeinschaft [HO 1286/6-4].

Appendix A

Table A1
Coding Scheme.

Level	Variable	Explanation
Study	Authors	
	Year	
	Country	
	Published	Yes, no
	Type of publication	Peer-reviewed paper, thesis, test manual, other
	Post hoc analysis	Yes, no
	Field study	Yes, no
	Number of administrations	
	N	Sample size

Sample	Age	Years
	Gender	% Male
	Intelligence	Average sample intelligence, z-scaled
	Intelligence test	Name of intelligence test
	Control group	Yes, no
	Dropouts	% Of participants dropped out
Test	Test form	Alternate, identical
	Test-retest interval	Time between administrations in hours
	Test	Name of test
	Subtest	Name of subtest
	Cognitive operation category	Processing speed, divergent thinking, memory, reasoning, general intelligence
	Spatial component	Yes, no (for reasoning tasks only)
	Test content category	Figural, numerical, verbal, several
	Explanation	Rule explanation before test: yes, no
	Practice	Practice items before test: yes, no
	Feedback	Feedback about correctness of the given answer: yes, no
	Total tests	Total number of subtests administered
	Scholastic aptitude test	Yes, no
	Format	Outcome format: score, accuracy, errors, interference score
	Comp	Comparison of administrations: first to second test = 1.2, first to third test = 1.3, first to fourth test = 1.4
	t2	Contrast variable 1: first to second test = 1, first to third test = 0, first to fourth test = 0
	t3	Contrast variable 1: first to second test = 0, first to third test = 1, first to fourth test = 0
	t4	Contrast variable 1: first to second test = 0, first to third test = 0, first to fourth test = 1
	M	Mean
	SD	Standard deviation
	$r_{1.t}$	Pearson's correlation between administrations (1.2, 1.3, 1.4)
	ES	Given effect size
	type *ES*	Type of effect size (e.g., *t*-value, *g*)

Table A2
Cognitive ability tests, their paradigms and categorization into the BIS model.

Cognitive ability operation	Test content	Name of test	Paradigm
PS	F	CogState (detection task, identification task, continuous monitoring task, reaction time tasks), CogSport/Axon Test Battery/Concussion Sentinel (detection, identification and matching tasks)	Quickly react to cards
		W-JTCA (pair cancellation task)	Quickly identify equal figures
		pattern comparison test (Salthouse & Babcock, 1991)	Quickly judge if patterns are equal or different
		attentive matrices test (Iuliano et al., 2015), WAIS (symbol search task)	Quickly identify targets
		TEA (map search task)	Quickly identify specific locations on map
		Simon Test	Quickly react to colors
	N	Connections Test (number task), TMT A	Quickly connect numbers
		W-JTCA (visual matching task)	Quickly circle equal numbers
		number comparison test (Salthouse & Babcock, 1991)	Quickly judge if two number are equal or different
	V	letter comparison test (Salthouse & Babcock, 1991)	Quickly judge if two letters are equal or different
		CPT, Go/NoGo test (Enge et al., 2014; Langenecker, Zubieta, Young, Akil & Nielson, 2007), Flanker test (Eriksen & Eriksen, 1974)	Quickly react to specific letter(s)
		stop signal test (Enge et al., 2014)	Quickly judge if letter presented is vowel or consonant
		letter cancellation test (Sharma et al., 2014), d2, finding A's test (French, Ekstrom & Price, 1963)	Quickly cancel specific letters in a list
	S	ANAM (code substitution task), RBANS (coding task), WAIS/WISC (coding task), digit symbol test (Nguyen et al., 2015), NAI (digit symbol test), SDMT	Quickly code digits and symbols according to coding scheme
		Connections Task (L-N and N-L tasks), TMT B	Quickly connect numbers and letters
		ImPACT (processing speed subscale)	Quickly react to letters, digit symbol coding

		Stroop Test, D-KEFS (color word interference test)	Quickly name color of colored color words
		Memoro (processing speed task)	Quickly judge if letters or digits are equal
		WAIS (processing speed subscale)	(Digit symbol coding and symbol search tasks)
M	F	ANAM (match to sample task), CANTAB (delayed match to sample task), Memoro (pattern separation task), CogState (matching task), CogSport/Axon Test Battery/Concussion Sentinel (one card learning task)	Recognize patterns/cards
		GMLT	Recall path in maze
		CogState (continuous paired associate learning task), AWMA (dot matrix task), WMS (spatial span (forward) task), block tapping test (Schellig, 1997), Memoro (objects in grid task), visual memory test (Lyall et al., 2016)	Recall positions
		spatial working memory test (Erickson et al., 2011), CANTAB (spatial recognition memory task)	Recognize positions
		RCFT (copy, immediate and delayed recall tasks), RBANS (figure recall task), BVMT (immediate and delayed recall tasks), adapted version of RVDLT (Robinson et al., 2013)	Recall figures
		RCFT (recognition task), visual-short term memory test (Luck & Vogel, 1997)	Recognize figures
		WMS (family pictures task)	Recall details about family picture
		WMS (faces task)	Recognize faces
		WMS (visual memory subscale: family pictures and faces tasks)	(Several figural memory tasks)
	N	RBANS (digit span task), WAIS/WISC (digit span (forward) task), AWMA (digit recall task), forward digit span test (Sharma et al., 2014), digit span forward test (Nguyen et al., 2015), running memory span test (Cowan et al., 2005)	Recall digits (forward)
		reading span test (Ölhafen et al., 2013), whole report test (Schubert et al., 2015)	Recall letters
	V	RAVLT (recall tasks), RBANS (list recall task, list learning task, HVLT (immediate and delayed recall tasks), short-term memory test (Flanagan et al., 1962; word and sentence tasks), recall of concrete nouns test (Dahlin, Nyberg, Baeckman & Stigsdotter Neely, 2008), WAIS (word recall test), VMPT (recall tasks), NAI (immediate and delayed recall tasks), recall of concrete nouns test (Sandberg, Ronnlund, Nyberg & Stigsdotter Neely, 2014), GMCT (immediate, delayed and free recall tasks), Memoro (verbal memory task), VCLT (immediate and delayed recall tasks)	Recall words
		RAVLT (recognition tasks), RBANS (list recognition tasks), WMS (word recall task), VMPT (recognition tasks)	Recognize words
		CVLT	Recall and recognize words
		short-term memory test (Flanagan et al., 1962)	Recall sentences
		RBANS (story memory task, story recall task), WMS (logical memory task)	Recall story
		paired associations test (Dahlin et al., 2008), paired associates test (Salthouse et al., 2004), WMS (verbal paired associates task), GMCT (paired associates task)	Recall word pairs
		WMS (auditory memory subscale: logical memory and verbal paired associates tasks)	(Several verbal memory tasks)
	S	updating test (Salminen, Strobach & Schubert, 2012)	Recall last items of a list
		attentional blink test (Salminen et al., 2012)	Recall first and last auditory or verbal items of a list
		ImPACT (verbal memory subscale)	Recognize words, match symbols and numbers, recall letters
		ImPACT (visual memory subscale)	Recognize figures, identify letters
DT	F	RBMT	(Several memory tasks)
	V	RWT, D-KEFS (word and letter fluency task), COWAT (category and letter fluency tasks), phonemic verbal fluency test (Gil-Gouveia, Oliveira & Martins, 2015), verbal categorical fluency test (Özer Celik et al., 2015), word fluency test (Mehlsen, Pedersen, Jensen & Zachariae, 2009)	generate as many words as possible for a given category or from a first letter
		verbal fluency test (Benton & Hamsher, 1976), COWAT (letter fluency task)	Generate as many words as possible from a given first letter
		COWAT (letter fluency task)	Generate as many words as possible for a given category
		PST	Find solutions for problematic situations
		alternate uses test (Sun et al., 2016)	Generate diverse uses of an object
		Torrance Test, divergent thinking test battery (Sun et al., 2016)	(Several verbal divergent thinking tasks)

R	F	RAPM, RSPM, RPM, RPM adaptation (Crone et al., 2009), figural matrices test (Freund & Holling, 2011), BOMAT, EAS (symbolic reasoning task), DAT (abstract reasoning task), WAIS/WASI (matrix reasoning task), BETAIII (matrix reasoning task)	Solve figural matrices
		CRTB (abstract reasoning task)	Complete objects
		GATB (form perception task)	Compare forms in detail
		WAIS/WASI/WISC (block design task)	Arrange blocks to complete image
		WCST, D-KEFS (sorting task), CANTAB (intra/extra dimensional shift test), categorization test (Soveri, Waris & Laine, 2013)	Sort cards according to rules
		Brixton Test	Predict where an object will appear
		D-KEFS (tower task), Tower of London Test	Build tower
		WAIS/WISC (object assembly)	Solve puzzle
		WAIS/WISC (picture completion), form boards test (French et al., 1963; Ekstrom et al., 1976)	Identify missing part of image
		WAIS/WISC (picture arrangement)	Arrange pictures in a logical order
		CANTAB (stockings of Cambridge task)	Arrange balls to copy Figure
		map planning, choosing a path and maze tracing speed tests (Ekstrom et al., 1976)	Find shortest way on a map
		GATB (spatial aptitude task), EAS (space visualization task), mental rotation test (Vandenberg & Kuse, 1978), GZSOT, GZSVT, mental rotation test (Shepard & Metzler, 1971; Peters et al., 1995), flags test of mental rotation (Thurstone & Jeffrey, 1959)	Rotate objects mentally (2D spatial component)
		figures test (Thurstone, 1958), cube comparisons and card rotation tests (French et al., 1963; Ekstrom et al., 1976), Elliot-Prive Board Block (Perspectives) Test, RBANS (line orientation task), primary mental abilities test (Thurstone & Thurstone, 1963)	Rotate objects mentally (3D spatial component)
		paper folding test (French et al., 1963; Ekstrom et al., 1976)	Match unfolded to folded object (spatial component)
		spatial relations test (Bennett et al., 1990)	Match 3D object to corresponding 2D object (spatial component)
		relative position test (Albers & Höft, 2009), Eliot-Donelly Test, object perspective test (Kozhevnikov & Hegary, 2001), environment pointing test (Meneghetti, Borella & Pazzaglia, 2016)	Take a perspective from a defined point to an object (spatial component)
		MGMP (spatial visualization task)	(Several figural reasoning tasks, spatial component)
	N	ANAM (mathematical processing task), TTB (numerical reasoning task), GATB (numerical aptitude task), SAT (regular mathematics task), WAIS/WISC (arithmetic task), EAS (numerical ability task), DAT (numerical reasoning task), AAET (arithmetical problems task)	Solve algebraic problems
		NAT (number completion task), VMNC, EAS (numerical reasoning task), number series test (Thurstone, 1938), numerical reasoning test (Do, 2011), number sequencing test (Özer Celik et al., 2015), AAET (number series completion task)	Complete series of numbers
		SAT (quantitative comparisons task)	Solve quantitative comparisons
		(P)SAT (math subscale), PET (quantitative reasoning task), PISA (math subscale)	(Several mathematical operations, e.g., algebra, geometry, combinatorics)
	V	VAT (verbal analogies task), GIT (verbal analogies task), MAT, DAT (verbal reasoning task), analogies test (Berger et al., 1990), verbal reasoning test (Do, 2011), AAET (analogies task)	Solve verbal analogies
		WAIS/WASI/WISC (similarities task)	Describe similarities of two concepts
		SAT (data sufficiency task)	Judge if given info is sufficient to solve problem
		EAT (verbal reasoning task), grammatical reasoning test (Baddeley, 1968)	Decide if conclusion based on given facts is true
		PMAB (inductive reasoning task)	Complete series of letters
		letter sets test (Ekstrom et al., 1976)	Identify letter set that violates rule
		inferences test (Ekstrom et al., 1976)	identify logical result of A statement
		AAET (disarranged sentences task)	Fix disarranged sentences
	S	WAIS (perceptual organization subscale)	(Several nonverbal reasoning tasks)
		CFT, medical college admission exam (Lievens et al., 2005; Lievens et al., 2007), Shipley abstract test (Zachary & Shipley, 1986), verbal-numerical reasoning test (Lyall et al., 2016)	(Several reasoning tasks)

G	S	WAIS, WPT, OGIT, LSAT, cognitive ability test (Arthur et al., 2010), ACT, SweSAT, AH4, TCEE	(Several cognitive ability tasks)

Note. PS = processing speed; M = memory; DT = divergent thinking; R = reasoning; G = general intelligence; F = figural; N = numerical; V = verbal; S = several; ACT = American College Test; AAET = Army Alpha Examinations Test; ANAM = Automated Neuropsychological Assessment Metrics; AWMA = Automated Working Memory Assessment; BOMAT = Bochumer Matrizentest; BVMT = Brief Visual Memory Test; CANTAB = Cambridge Neuropsychological Automated Test Battery; CPT = Continuous Performance Test; COWAT = Controlled Oral Word Associations Test; CRTB = Critical Reasoning Test Battery; CFT = Culture Fair Test; D-KEFS = Delis-Kaplan Executive Function System; DAT = Differential Aptitude Test; EAS = Employee Aptitude Survey; GATB = General Aptitude Test Battery; GMCT = Green's Memory and Concentration Test; GIT = Groningen Intelligence Test; GMLT = Groton Maze Learning Test; GZSOT = Guildford-Zimmerman Spatial Orientation Test; GZSVT = Guildford-Zimmerman Spatial Visualization Test; HVLT = Hopkins Verbal Learning Test; ImPACT = Immediate Post-Concussion Assessment and Cognitive Testing; LSAT = Law School Admission Test; MGMP = Middle Grade Mathematics Project; MAT = Miller Analogies Test; NAT = Numerical Aptitude Test; NAI = Nürnberger Altersinventar; OGIS = Otis Group Intelligence Scale; PSAT = Preliminary Scholastic Aptitude Test; PMAB = Primary Mental Abilities Test Battery; PET = Psychometric Entrance Test; PST = Problems Situations Test; RAPM = Raven's Advanced Progressive Matrices Test; RSPM = Raven's Standard Progressive Matrices Test; RPM = Raven's Progressive Matrices; RWT = Regensburger Wortflüssigkeitstest; RBANS = Repeatable Battery for the Assessment of Neuropsychological Status; RAVLT = Rey Auditory Verbal Learning Test; RCFT = Rey Complex Figure Test; RVDLT = Rey Visual Design Learning Test; RBMT = Rivermead Behavioral Memory Test; RFFT = Ruff Figural Fluency Test; S-ILS = Shipley Institute of Living Scale; SweSAT = Swedish Scholastic Aptitude Test; SDMT = Symbol Digit Modalities Test; TTB = Technical Test Battery; TEA = Test of Everyday Attention; TCEE = Thorndike College Entrance Examination; TMT (A + B) = Trail Making Test (Parts A + B); VMNC = Van der Maesen Number Completion; VAT = Verbal Aptitude Test; VLMT = Verbal Learning and Memory Test; VMPT = Verbal Memory Process Test; WAIS = Wechsler Adult Intelligence Scale; WISC = Wechsler Intelligence Scale for Children; WMS = Wechsler Memory Scale; WPT = Wonderlic Personnel Test; W-JTCA = Woodcock-Johnson III test of cognitive abilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2018.01.003.

References*

*Albers, F., & Hoeft, S. (2009). Do it again and again. And again - Übungseffekte bei einem computergestützten Test zum räumlichen Vorstellungsvermögen [Do it again and again. And again – Practice effects in a computer-based test of spatial ability]. *Diagnostica, 55*(2), 71–83.

*Allalouf, A., & Ben-Shakar, G. (1998). The effect of coaching on the predictive validity of Scholastic Aptitude Tests. *Journal of Educational Management, 35*(1), 31–47.

*Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects in unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*(1), 1–16.

*Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *Neuroscience, 11*, 118–129.

*Broglio, S. P., Ferrara, M. S., Macciocchi, S. N., Baumgartner, T. A., & Elliott, R. (2007). Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training, 42*(4), 509–514.

*Buschkühl, M. (2007). *Arbeitsgedächtnistraining: Untersuchungen mit jungen und älteren Erwachsenen [Working Memory Training: Studies with Young and Older Adults].* Doctoral dissertationUniversität Bern.

*Colom, R., Quiroga, M.Á., Solana, A. B., Burgaleta, M., Román, F. J., Privado, J., ... Karama, S. (2012). Structural changes after videogame practice related to a brain network associated with intelligence. *Intelligence, 40*, 479–489.

*Dunlop, P. D., Morrison, D. L., & Cordery, J. L. (2011). Investigating retesting effects in a personnel selection context. *International Journal of Selection and Assessment, 19*(2), 217–221.

*Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No evidence for true training and transfer effects after inhibitory control training in young healthy adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 987–1001.

*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.

*Lo, A. Y., Humphreys, M., Byrne, G. J., & Pachana, N. A. (2012). Test-Retest reliability and practice effects of the Wechsler Memory Scale-III. *Journal of Neuropsychology, 6*(2), 212–231.

*Randall, J. G., Villado, A. J., & Zimmer, C. U. (2016). Is retest bias biased? Examining race and sex differences in retest performance. *Journal of Personnel Psychology, 15*(2), 45–54.

*Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence, 33*, 535–549.

*Robinson, S. J., Leach, J., Owen-Lych, P. J., & Sünram-Lea, S. I. (2013). Stress reactivity and cognitive performance in a simulated firefighting emergency. *Aviation, Space, and Environmental Medicine, 84*(6), 592–599.

*Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristic on retest score gains and criterion-related validity. *Journal of Business and Psychology, 31*, 233–248. http://dx.doi.org/10.1007/s10869-015-9408-7.

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*(1), 3–27.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*(10), 1086–1093.

Arendasy, M. E., & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*, 181–192.

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence, 62*, 89–98.

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.

Baddeley, A. D. (1968). A three-minute reasoning test based on grammatical transformation. *Psychonomic Science, 10*, 341–342.

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

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*(2), 257–278.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*(3), 373–417.

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administration of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 339–352.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1990). *Differential Aptitude Tests, Levels 1 and 2, Forms C and D.* San Antonio, USA: The Psychological Corporation.

Benton, A. L., & Hamsher, K. (1976). *Multilingual aphasia examination.* Iowa City: AJA.

Bors, D. A. (1993). The factor-analytic approach to intelligence is alive and well: A review of Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. *Canadian Journal of Experimental Psychology, 47*(4), 763–766.

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Behavioral Reviews, 26*(7), 809–817.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three level hierarchichal linear model. *American Journal of Education, 97*(1), 65–108.

Buehner, M., Krumm, S., & Pick, M. (2005). Reasoning = working memory ≠ attention. *Intelligence, 33*(3), 251–272.

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.

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

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

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

Childers, T. L., & Viswanathan, M. (2000). Representation of numerical and verbal product information in consumer memory. *Journal of Business Research, 47*, 109–120.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*(2nd ed.). New York: Russell Sage Foundation.

Cooper, H. M. (2017). *Research Synthesis and meta-analysis: A Step-by-Step Approach* (5th ed.). Sage.

Cowan, N., Elliot, E. M., Saults, J. S., Morey, C. C., Mattox, S., ... Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42–100.

* References marked with an asterisk are studies included in the meta-analysis discussed in the text. For a complete list of included studies, see Online Supplement 2.

Coyle, T. R. (2005). Test-retest changes on scholastic aptitude tests are not related to g. *Intelligence, 34*, 15–27.

Crone, E. A., Wendelken, C., Van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science, 12*, 55–66. http://dx.doi.org/10.1111/j.1467-7687.2008.00743.x.

*Dahlin, E., Nyberg, L., Baeckman, L., & Stigsdotter Neely, A. (2008). Plasticity of executive functioning in young and older adults: Immediate training gains, transfer, and long-term maintenance. *Psychology and Aging, 23*(4), 720–730.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*(1), 1–15.

*Do, B.-R. (2011). *Test coaching on assessments of cognitive constructs.* University of Illinois: Unpublished doctoral dissertation.

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review, 22*, 1308–1319.

*van Eersel, M. E. A., Joosten, H., Koerts, J., Gansevoort, R. T., Slaets, J. P. J., & Izaks, G. J. (2015). A longitudinal study of performance on the Ruff Figural Fluency Test in persons aged 35 years or older. *PLoS One, 10*(3), 1–14.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., ... Kramer, A. F. (2011). Exercise training increases size of hippocampus and improvesmemory. *Proceedings of the National Academy of Sciences, 108*, 3017–3022. http://dx.doi.org/10.1073/pnas.1015950108.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study of American youth.* Boston: Houghton Mifflin.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Kif of reference tests for cognitive factors.* Princeton, NJ: Educational Testing Service.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiment involving paired comparisons. *Journal of Educational Statistics, 18*(3), 271–279.

*Gil-Gouveia, R., Oliveira, A. G., & Martins, I. P. (2015). Sequential brief neuropsychological evaluation of migraineurs is identical to controls. *Acta Neurologica Scandivica.* http://dx.doi.org/10.1111/ane.12530.

Grissom, N., & Bhatnagar, S. (2009). Habituation to repeated stress: Get used to it. *Neurobiology of Learning and Memory, 92*(2), 215–224.

Guthke, J., & Beckmann, J. (2001). Intelligenz als "Lernfähigkeit" – Lerntests als Alternative zum herkömmlichen Intelligenztest [Intelligence as "learning ability" - Study Tests as an Alternative for Common Intelligence Tests]. In E. Stern, & J. Guthke (Eds.). *Perspektiven der Intelligenzforschung.* Lengerich: Pabst Publisher.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognition and Science, 11*(6), 236–242.

Hasegawa, H. (1997). On the relationship between recognition memory of items and test-retest effect. *The Japanese Journal of Psychology, 68*(5), 417–422.

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.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87*(2), 243–254.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1–14.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65.

Holland, B. S., & DiPonzio Copenhaver, M. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin, 104*(1), 145–149.

Holling, H., Preckel, F., & Vock, M. (2004). *Intelligenzdiagnostik [Diagnosing Intelligence].* Göttingen: Hogrefe.

Holm (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Huber, O. (1980). The influence of some task variables on cognitive operations in an information-processing decision model. *Acta Psychologica, 45*, 187–196.

Hülsheger, U. R., Maier, G. W., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15*(1), 3–18.

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

Ishak, K. J., Platt, R. W., Joseph, L., Hanley, J. A., & Caro, J. J. (2007). Meta-analysis of longitudinal studies. *Clinical Trials, 4*, 525–539.

*Iuliano, E., di Cagno, A., Aquino, G., Fiorilli, G., Mignogna, P., Calcagno, G., & Di Costanzo, A. (2015). Effects of different types of physical activity on the cognitive functions and attention in older people: A randomized controlled study. *Experimental Gerontology, 70*, 105–110.

Jaber, M. Y., & Glock, C. H. (2013). A learning curve for tasks with cognitive and motor elements. *Computers & Industrial Engineering, 64*(3), 866–871.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. [Translation of multimodal classification of intelligence test performance: Experimentally controlled development of a descriptive intelligence structure model.]. *Diagnostica, 28*(3), 195–225.

Karch, D., Albers, L., Renner, G., Lichtenauer, N., & von Kries, R. (2013). The efficacy of cognitive training programs in children and adolescence: A meta-analysis. *Deutsches Aerzteblatt International, 110*(39), 643–652.

Kelly, M., 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.

Klauer, K. J. (2014). Training des induktiven Denkens – Fortschreibung der Metaanalyse von 2008 [Training inductive reasoning – update on the meta-analysis from 2008]. *Zeitschrift für Pädagogische Psychologie, 28*(1–2), 5–19.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research, 78*(1), 85–123.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 17*, 279–301.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition, 29*, 745–756. http://dx.doi.org/10.3758/BF03200477.

Kubinger, K. D., & Jäger, R. S. (2003). *Schlüsselbegriffe der Psychologischen Diagnostik [Key Concepts of Psychological Diagnosis].* Beltz: Weinheim.

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.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (littler more than) working memory capacity?!. *Intelligence, 14*(4), 389–433.

Lader, M. H., & Wing, L. (1964). Habituation of the psycho-galvanic reflex in patients with anxiety states and in normal subjects. *Journal of Neurology, Neurosurgery & Psychiatry, 27*(3), 210–218.

Lampit, A., Hallock, H., & Valenzuela, M. (2014). Computerized cognitive training in cognitive healthy older adults: A systematic review and meta-analysis of effect modifiers. *PLoS Medicine, 11*(11), 1–18.

Lane, R. G., Penn, N. E., & Fischer, R. F. (1966). Miller Analogies Test: A note on permissive retesting. *Journal of Applied Psychology, 50*(5), 409–411.

*Langenecker, S. A., Zubieta, J.-K., Young, E. A., Akil, H., & Nielson, K. A. (2007). A task to manipulate attentional load, set-shifting, and inhibitory control: Convergent validity and test-retest reliability of the Parametric Go/No-Go Test. *Journal of Clinical and Experimental Neuropsychology, 29*(8), 842–853.

*Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–1007.

*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.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281. http://dx.doi.org/10.1038/36846.

*Lyall, S. M., Cullen, B., Allerhand, M., Smith, D. J., Mackay, D., Evans, J., ... Pell, J. P. (2016). Cognitive test scores in UK Biobank: Data reduction in 480,416 participants and longitudinal stability in 20,346 participants. *Plos One, 11*(4), e0154222. http://dx.doi.org/10.1371/journal.pone.0154222.

Maerlender, A. C., Masterson, C. J., James, T. D., Beckwith, J., & Brolinson, P. G. (2016). Test-retest, retest, and retest: Growth curve models of repeat testing with Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT). *Journal of Clinical and Experimental Neuropsychology*. http://dx.doi.org/10.1080/13803395.2016.1168781.

Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence, 37*, 412–421.

*Mehlsen, M., Pedersen, A. D., Jensen, A. B., & Zachariae, R. (2009). No indications of cognitive side-effects in a prospective study of breast cancer patients receiving adjuvant chemotherapy. *Psycho-Oncology, 18*, 248–257.

*Meneghetti, C., Borella, E., & Pazzaglia, F. (2016). Mental rotation training: Transfer and maintenance effects on spatial abilities. *Psychological Research, 80*, 113–127.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*(2), 191–216.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational Psychological Measurement, 15*(3), 707–726.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Groupi, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62*(10), 1006–1012.

Morris, S. B., & DeShon, R. P. (2002). Combining effect sizes in meta-analysis with repeated measures and independent groups designs. *Psychological Methods, 7*(1), 105–125.

Musekiwa, A., Manda, S. O. M., Mwambi, H. G., & Chen, D.-G. (2016). Meta-analysis of effect sizes reported at multiple time points using general linear mixed model. *PLoS One, 11*(10), http://dx.doi.org/10.1371/journal.pone.0164898.

Nevo, B. (1976). The effects of general practice, specific practice, and item familiarization on change in aptitude test scores. *Measurement and Evaluation in Guidance, 9*, 16–20.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.). *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Ng, E., & Lee, K. (2015). Effects of trait test anxiety and state anxiety on children's working memory task performance. *Learning and Individual Differences, 40*, 141–148.

*Nguyen, H. T., Quandt, S. A., Summers, P., Morgan, T. M., Chen, H., Walker, F. O., ... Arcury, T. A. (2015). Learning ability as a function of practice: Does it apply to farmworkers? *Journal of Occupational Environmental Medicine, 57*(6), 676–680.

te Nijenhuis, J., van Vianen, A. E., & van der Vlier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence, 35*(3), 238–300.

Olenick, J., Bhatia, S., & Ryan, A. M. (2016). Effects of g-loading and time lag on retesting in job selection. *International Journal of Selection and Assessment, 24*(4), 324–336.

*Ölhafen, S., Nikolaidis, A., Padovani, T., Blaser, D., Koenig, T., & Perrig, W. J. (2013). Increased parietal activity after training of interference control. *Neuropsychologica, 51*(13), 2781–2790.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in personnel selection

decisions. In A. Evers, O. Voskuijl, & N. Anderson (Eds.). *Handbook of selection* (pp. 143–183). Oxford, UK: Blackwell.
*Pereira, D., Costa, P., & Cerqueira, J. J. (2015). Repeated assessment and practice effects of the written symbol digit modalities test using a shorter inter-test interval. *Archives of Clinical Neuropsychology, 30*, 424–434.
Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition, 28*(1), 39–58.
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*, 1055–1079.
Puddey, I. B., Mercer, A., Andrich, D., & Styles, I. (2014). Practice effects in medical school entrance testing with the undergraduate medicine and health sciences admission test (UMAT). *Medical Education, 14*, 48–62.
R Core Team (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. URL: https://wwwRpoject.org.
Randall, J. G., & Villado, A. J. (2017). Take two: Sources and deterrents of score change in employment testing. *Human Resource Management, 27*, 536–553.
Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 310–319.
Raven, J. C. (1936). *Standard progressive matrices, Sets A, B, C, D, E.* London: Lewis.
Raven, J. C. (1962a). *Advanced progressive matrices, Set II.* London: Lewis.
Raven, J. C. (1962b). *Coloured progressive matrices, Sets A, AB, B.* London: Lewis.
Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence, 37*, 31–34.
Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.
Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*(3), 249–255.
Salanti, G., Higgins, J. P. T., Ades, A. E., & Ioannidis, J. P. A. (2008). Evaluation of networks of randomized trials. *Statistical Methods in Medical Research, 17*, 279–301.
*Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Frontiers in Human Neuroscience, 6*, 166. http://dx.doi.org/10.3389/fnhum.2012.00166.
Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psycholology, 27*, 763. http://dx.doi.org/10.1037/0012-1649.27.5.763.
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.
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.
Sanches de Oliveira, R., Trezza, B. M., Busse, A. L., & Jocob-Filho, W. (2014). Learning effect of computerized cognitive tests in older adults. *Einstein, 12*(2), 149–153.
*Sandberg, P., Ronnlund, M., Nyberg, L., & Stigsdotter Neely, A. (2014). Executive process training in young and old adults. *Aging, Neuropsychology, and Cognition, 21*, 577–605.
Sarason, I. G., & Palola, E. G. (1960). The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology, 59*(3), 185–191.
Schellig, D. (1997). *Block-tapping test.* Frankfurt am Main: Swets Tests Services.
Schleicher, D. J., Van Iddekinge, C. H., 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), 603–627.
Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45*(2), 294–301.
Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. *Creativity Research Journal, 16*(4), 361–388.
Shaffer, D. R., & Kipp, K. (2010). *Developmental Psychology: Childhood and adolescence* (8th ed.). Belmont, CA, US: Thomson Brooks/Cole Publishing Co.
*Sharma, V. K., Rajajeyakumar, M. R., Velkumary, S., Subramanian, S. K., Bhavanani, A. B., Madanmohan, A. S., ... Thangavel, D. (2014). Effect of fast and slow pranayama practice on cognitive functions in healthy volunteers. *Journal of Clinical & Diagnostic Research, 8*(1), 10–13.
Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 191*, 952–954.
*Soveri, A., Waris, O., & Laine, M. (2013). Set shifting training with categorization tasks. *Plos One, 8*(12), http://dx.doi.org/10.1371/Journal.pone.0081693.
Spielberger, C. D. (1959). Evidence of a practice effect on the miller analogies test. *Journal of Applied Psychology, 43*(4), 259–263.
Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working memory capacity explains reasoning ability – And a little bit more. *Intelligence, 30*(3), 228–261.
*Sun, J., Chen, Q., Zhang, Q., Li, Y., Li, H., Wie, D., ... Qiu, J. (2016). Training your brain to be more creative: Brain functional and structural changes induces by divergent thinking training. *Human Brain Mapping.* http://dx.doi.org/10.1002/hbm.23246.
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.
Trikalinos, T. A., & Olkin, I. (2012). Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. *Clinical Trials, 9*, 610–620.
Thurstone, L. L. (1938). *Primary mental abilities.* Chicago, IL: University of Chicago Press.
Thurstone, L. L., & Jeffrey, T. E. (1959). *Test administration manual for the space thinking (flags) test.* Chicago, IL: Education-Industry Service.
Thurstone, T. G. (1958). *Manual for the SRA primary mental abilities.* Chicago: Science Research Associates.
Thurstone, T. G., & Thurstone, L. L. (1963). *Primary mental ability.* Chicago: Science Research Associates.
Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352–402.
Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of 3-dimensional spatial visualization. *Perceptual and Motor Skills, 47*, 599–604. http://dx.doi.org/10.2466/pms.1978.47.2.599.
Van Iddekinge, C. H., 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.
Venkatesan, M., Lancaster, W., & Kendall, K. W. (1986). Empirical study of alternate formats for nutritional information disclosure in advertising. *Journal of Public Policy and Marketing, 5*, 29–43.
Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. URL: http://www.jstatsoft.org/v36/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*, 907. http://dx.doi.org/10.3389/psyg.2016.00907.
Wechsler, D. (1958). *Measurement and appraisal of adult intelligence* (4th ed.). Baltimore, USA: Williams & Wilkins Co.
Wechsler, D. (1987). *Wechsler memory scale-revised.* New York: Psychological Corporation.
Woehlke, A. B., & Wilder, D. H. (1963). Differences in difficulty of forms A and B of the Otis self-administering test of mental ability. *Personnel Psychology,* 198–395.
Zachary, R. A., & Shipley, W. C. (1986). *Shipley institute of living scale: Revised manual.* Los Angeles: Western Psychological Serviceshttp://dx.doi.org/10.1002/acp.1202.
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.