

Training Working Memory for Two Years—No Evidence of Transfer to Intelligence

Luc Watrin¹, Gizem Hülür², and Oliver Wilhelm¹

¹ Institute of Psychology and Education, Department of Individual Differences and Psychological Assessment, Ulm University

² School of Aging Studies, University of South Florida

Working memory (WM) training has been proposed as a promising intervention to enhance cognitive abilities, but convincing evidence for transfer to untrained abilities is lacking. Prevalent limitations of WM training studies include the narrow assessment of both WM and cognitive abilities, the analysis of manifest variables subject to measurement error, and training dosages too low to likely cause changes in the cognitive system. To address these limitations, we conducted a 2-year longitudinal study to investigate the effects of working memory training on latent factors of working memory capacity, fluid intelligence and crystallized intelligence. One hundred twelve students initially attending 9th grade practiced a heterogeneous set of validated WM tasks on a biweekly basis. A control group of 113 students initially attending 9th grade participated in the pretest and posttest. Broad and prototypical measures of fluid and crystallized intelligence served as measures of nearer and farther transfer. We found substantial and reliable training effects on the practiced WM tasks, as well as on a latent WM factor constituted by them. However, no transfer of training effects to fluid or crystallized intelligence were observed. These results add to the literature questioning the utility and validity of WM training as means of improving cognitive abilities.


Keywords: cognitive training, intelligence, latent change score models, transfer effects, working memory


Given the paramount importance of cognitive ability in everyday life, even small reliable improvements in the construct would be worth substantial effort and lead to improvements in related abilities, higher intellectual performance in everyday life, and meaningful delay of cognitive decline. Accordingly, researchers have extensively


investigated the utility of cognitive interventions and commercial providers confidently advertise various types of *brain trainings*, *neuro enhancements* and the like. However, early cognitive trainings did not live up to their expectations (see Carroll, 1993, p. 669ff for a review), and the relative hopelessness of this endeavor has caused significant scientific and political controversy (e.g., Cronbach, 1969; Jensen, 1969). Renewed enthusiasm arose from studies reporting positive effects of working memory (WM) training on cognitive abilities (Au et al., 2015; Jaeggi et al., 2008; Klingberg et al., 2002), and it has quickly become the most prevalent type of cognitive intervention. Unfortunately, the enthusiasm created by these studies has largely again led to disappointment, after several systematic reviews and meta-analyses of available evidence failed to provide unequivocal evidence for the generalizability of training effects to untrained abilities (Redick, 2019; Sala, Aksayli, Tatlidil, Tatsumi, et al., 2019; Shipstead et al., 2012). The lack of effects has been attributed to a series of shortcomings commonly encountered in cognitive intervention studies, leaving the key question of whether cognitive abilities are malleable through tailored interventions still unanswered. In the following, we will first discuss the relation of WM with established cognitive abilities. We will then elaborate on a number of key issues in WM training research and specify how the present article contributes to the ongoing debate on the effectiveness of WM training.

Working Memory and Cognitive Abilities

Competing models of WM (Engle, 2002; Oberauer, 2009; Unsworth & Engle, 2007) agree that WM is a system for storage and

Luc Watrin  <https://orcid.org/0000-0003-4343-3781>

Gizem Hülür  <https://orcid.org/0000-0002-9225-3878>

Oliver Wilhelm  <https://orcid.org/0000-0001-7980-1166>

Gizem Hülür is now at Department of Psychology, University of Bonn.

Parts of the data reported in this article have also been used in other articles (Gasimova, Robitzsch, Wilhelm, & Hülür, 2014; Gasimova, Robitzsch, Wilhelm, Boker, et al., 2014; Hülür et al., 2017, 2018; Hülür, Wilhelm, & Robitzsch, 2011; Hülür, Wilhelm, & Schipolowski, 2011). We have no conflicts of interests to disclose. This research was supported by the Institute for Educational Progress, Humboldt University, Berlin, Germany, and by a Grant from the German Research Foundation (DFG WI 2667/7-1) to Oliver Wilhelm and Alexander Robitzsch. At the time the data were collected (2005–2008), neither the funding organization (German Research Foundation DFG) nor the university at which the research was conducted (Humboldt University, Berlin, Germany) required ethical approval for this kind of purely behavioral research. The data that support the findings of this study are available at <https://osf.io/x8znf/>.

Correspondence concerning this article should be addressed to Luc Watrin, Institute of Psychology and Education, Department of Individual Differences and Psychological Assessment, Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany. Email: luc.watrin@uni-ulm.de

processing of a limited amount of information. In our view, it is a system that allows to flexibly build, maintain, and update bindings (Oberauer et al., 2008; Wilhelm et al., 2013). WM allows to consciously keep chunks of information needed for ongoing cognitive processes available for direct access by placing and maintaining this information in a cognitive coordinate system. Chunks in the coordinate system can be accessed and subjected to cognitive operations, such as building new relations between them. Critically, people differ in the amount of information they can simultaneously use in their WM. This individual differences construct is termed working memory capacity (WMC) and is positively related to a large number of real-world outcomes, like the acquisition of mathematics (Peng et al., 2016), second language acquisition (Linck et al., 2014), or reading skills in children (Peng et al., 2018). Impairments in WMC, in turn, have been observed in a variety of disorders, such as attention deficit hyperactivity disorder (ADHD; Alderson et al., 2013), autism spectrum disorder (Wang et al., 2017), and learning disabilities (Alloway, 2009).

Importantly, WMC is highly correlated with psychometric intelligence (Oberauer et al., 2005) and has been discussed as the explanatory construct for intellectual abilities for almost 40 years (Fogarty & Stankov, 1982; Johnson-Laird, 1983). This claim is supported both theoretically and empirically. Reasoning items are not good measures of fluid intelligence by virtue of some surface characteristic, by imposing unbearable loads of some logic calculus on participants, or by asking subjects to solve somewhat novel tasks—the core of effortful cognitive processing as required in reasoning tasks is to detect regularities and to draw valid, useful, non-obvious inferences (Wilhelm, 2005). It is the core cognitive processes of building, maintaining, and updating relations between new chunks of information that are the foundation for engaging in and solving tasks that are good indicators of fluid intelligence (Oberauer et al., 2008). Fluid intelligence, in turn, has been reported to correlate perfectly with general cognitive ability (e.g., Gustafsson, 1984) and is widely accepted as the key ingredient in human cognitive abilities as it is central to all cognitive abilities and an essential determinant of learning of knowledge and skills (Carroll, 1993; Flanagan & Harrison, 2012).

Empirically, individual differences research has repeatedly and successfully shown that individual differences in WMC are the key limiting factor of reasoning and are the best predictor of fluid intelligence (Kyllonen & Christal, 1990; Oberauer et al., 2005, 2008). Disattenuated correlations of broad measurements of WMC and cognitive ability are estimated at around $\rho = .85$ (Kyllonen & Christal, 1990; Oberauer et al., 2005; Süß et al., 2002), indicating that WMC shares over 70% common variance with psychometric intelligence. Although this correlation does not and cannot equal unity owing to differences in conception, measurement, and factor structure (see Oberauer et al., 2005, for a discussion), models of WM are the best theoretical cognitive basis to date to explain individual differences in fluid intelligence and relevant lower-order cognitive processes. It is therefore instrumental to consider established theories of WM to better understand what limits, or conversely, what might improve, cognitive ability.

From a pragmatic view, focusing on WMC has some key advantages over fluid intelligence when it comes to cognitive interventions. Training material is easier and more efficient to construct. Item difficulty is easier to predict, variation of item difficulty is easier to achieve, and the effects of item attributes can be

explained more reliably based on the vast corpus of research from cognitive psychology that is as of yet not paralleled in individual differences research about fluid intelligence (Oberauer et al., 2018). Importantly, from an epistemic point of view, training WM supposedly addresses the underlying mechanism whereas training fluid intelligence is indirectly focusing on its application.

Working Memory Training

The aim of cognitive interventions, as opposed to simple test practice or coaching (e.g., SAT practice; The Princeton Review, 2020), is not to ameliorate performance on the practiced task(s) alone but to improve the overarching ability that is assumed to be causal for observed performance. Training-induced changes must therefore not be limited to improved performance on the trained tasks but transcend to improved performance on sufficiently dissimilar untrained measures of the targeted construct. Applied to WM, improvements must not be restricted to the practiced WM task but transfer to untrained WM tasks or further, given the postulated causal relation with fluid intelligence, improvements should positively transfer to measures of reasoning ability and eventually real-life variables such as educational achievement or job success.

Evidence regarding the effects of WM training, however, is disappointing. A plethora of primary studies, several meta-analyses (e.g., Melby-Lervåg et al., 2016; Melby-Lervåg & Hulme, 2013; Sala, Aksayli, Tatlidil, Gondo, et al., 2019; Schwaighofer et al., 2015), a second-order meta-analysis (Sala, Aksayli, Tatlidil, Tatsumi, et al., 2019), and several literature reviews (e.g., Redick, 2019; Shipstead et al., 2010, 2012) have investigated and discussed the effects of WM trainings of different kinds (see Table 1 for an incomplete overview of meta-analytical results of the last 5 years).

In summary, the observed effects of WM training seem to be a function of the distance between the trained and the targeted task: Meta-analyses report medium to large performance improvements in the trained tasks, small to medium improvements in different tasks tapping the same underlying ability (i.e., WM-tasks), and little to no improvements in untrained but still proximal cognitive tasks tapping distinct but closely related abilities (e.g., fluid intelligence; Aksayli et al., 2019; Melby-Lervåg et al., 2016; Nguyen et al., 2019; Sala, Aksayli, Tatlidil, Gondo, et al., 2019; Sala, Aksayli, Tatlidil, Tatsumi, et al., 2019; Teixeira-Santos et al., 2019; see Table 1). The small transfer effects to cognitive abilities reported in some meta-analyses (e.g., Au et al., 2015; Karbach & Verhaeghen, 2014) have been attributed to different levels of motivation in experimental designs with passive control groups, inadequate control for baseline differences between experimental groups, or the selection and coding of studies (e.g., Melby-Lervåg et al., 2016; see Shipstead et al., 2010, 2012, for a general discussion of prevalent shortcomings in WM training studies). Studies that do report transfer from single task WM training to a single far transfer task (e.g., matrices; Jaeggi et al., 2008) should be interpreted with caution and skepticism. Although it is possible that training effects occurred at the latent ability level, it is more plausible and parsimonious to attribute the observed effects to one or more alternative explanations (such as shared task characteristics between trained and targeted tasks) rather than a true and persisting change in underlying abilities.

Table 1
Overview of Meta-Analyses Published Between 2016 and 2021 Which Estimated Far Transfer Effects of WM Training and Reported Training Duration in Hours

Meta-Analysis	Target	Near	Far	Far _{passive}	Far _{active}	Dosage (h)
Aksayli et al. (2019)	—	.44 [.38, .55]	.05 [−.02, .11]	—	—	<i>M</i> = 15.08, <i>SD</i> = 5.62, Range = 6.25–35.00
Nguyen et al. (2019)	1.17 [.92, 1.43]	.37 [.20, .55]	.12 [−.01, .25]	—	—	<i>M</i> = 8.30, <i>SD</i> = 4.25, Range = 2.33–20.00
Sala, Aksayli, Tatlidil, Gondo, et al. (2019)	.88 [.69, 1.06]	.27 [.19, .36]	.12 [.03, .21]	.26 [.12, .40]	−.01 [−.10, .09]	<i>M</i> = 10.93, <i>SD</i> = 7.49, Range = 1.75–24.00
Teixeira-Santos et al. (2019)	—	.23 [.07, .39] ^a	.10 [−.03, .23]	.08 [−.11, .27]	.14 [−.04, .33]	<i>M</i> = 7.17, <i>SD</i> = 4.20, Range = 1.50–20.00
Mewborn et al. (2017)	.44 [.36, .52]	.23 [.03, .43] ^b	.15 [.09, .20]	—	—	<i>M</i> = 17.66, <i>SD</i> = 30.85, Range = 1.00–270.00
Sala and Gobet (2017)	—	.46 [.35, .57]	.12 [.06, .18]	—	—	<i>M</i> = 6.60, <i>SD</i> = 3.16, Range = 2.00–14.58
Soveri et al. (2017)	.63 [.44, .82]	.24 [.16, .33]	.16 [.08, .24]	.22 [NA]	.11 [NA] n.s.	<i>M</i> = 6.23, <i>SD</i> = 2.76, Range = 1.00–15.00
Melby-Lervåg et al. (2016)	.80 [.62, .97] ^c	.28 [.16, .40] ^c	—	.20 [.11, .28] ^e	.05 [−.02, .13] ^e	<i>M</i> = 10.71, <i>SD</i> = 5.93, Range = 1.00–30.00
Weicker et al. (2016)	1.88 [1.33, 2.42] ^d	.51 [.34, .69] ^d	.22 [.09, .35]	.03 [−.09, .24] ^f	.05 [−.07, .17] ^f	<i>M</i> = 12.78; <i>SD</i> = 14.92; Range = 2.50–100.00

Note. Effect sizes reported in the columns target, near, far, far_{passive} and far_{active} are net training effect in Hedge's *g*. Numbers in brackets indicate 95% confidence intervals. Meta-analyses contain redundancies in primary studies. Categorizations of target, near transfer and far transfer were taken from the original studies. NA = not available; n.s. = not significant. Supplementary information on the sources of the computed estimates is provided in the online supplement (<https://osf.io/tpmr3/>).

^a Verbal WM. ^b Visuospatial WM. ^c Active control group. ^d Passive control group. ^e Nonverbal ability. ^f Verbal ability.

The gradual decrease of training effects with increasing dissimilarity between trained and transfer tasks indicates that interventions do not seem to affect fundamental cognitive processes and functions that WM and other cognitive tasks share. Instead, task-specific skills and strategies are improved, which is in line with the distinction of “elements of skill” or “behavioral flexibility” versus “abilities” (Lövdén et al., 2010; Thorndike, 1906). Put simply, participants get better at specific WM tasks, but this has no impact on their general cognitive functioning whatsoever.

These results are sobering but there is reason to believe that they can be considered inconclusive. Theoretical and methodological critique has been put forward that most WM training studies suffer from a number of shortcomings that hinder them from producing substantial transfer effects and from subsequently detecting or for testing them, should they occur (Noack et al., 2014; Schmiedek et al., 2019; Shipstead et al., 2012). In the following, we review the ones we deem the most essential at the current point in research, which are dosage of intervention, multivariate assessment of constructs, and latent variable analysis.

Requirements for WM Training Studies

To cause profound and lasting changes, a cognitive intervention must be sufficiently long and intense (Lövdén et al., 2010; Shipstead et al., 2012). Day in and day out, we rely on and practice our WM and our fluid intelligence. Our societies devote enormous resources to practicing and rewarding cognitive achievements. Given the ubiquitous importance of both constructs in everyday life, it is unrealistic to expect substantial changes therein after a short and low-dosage intervention. Training dosage has been investigated in primary studies and meta-analyses as a potential moderator of the training outcome and has received mixed results (Jaeggi et al., 2008; Redick et al., 2013; Sala, Aksayli, Tatlidil, Gondo, et al., 2019; Teixeira-Santos et al., 2019). However, in their review of WM training studies Noack et al. (2014) reported a median training duration of only 8 hr which conforms with our review of training dosages reported in meta-analyses of the last five years (see Table 1). Training durations go as low as a single hour of intervention and rarely exceed 20 hours. We therefore argue that most training studies are not even close to providing a sufficient dosage to elicit substantial and lasting changes in a cognitive system that is and has already been under constant strain for years (e.g., students) or decades (e.g., older people).

WM and transfer constructs are latent ability constructs and should be measured as such. A single task, as implemented in many studies, cannot be equated with its underlying ability. WM cannot directly be measured but must be inferred from performance on tasks that allegedly measure WM in different kinds and extents. Conceptually, single measurements always contain true variance related to the ability of interest, task-specific variance, and measurement error. For example, Soveri et al. (2017) showed that large parts of observed n-back training effects are task specific, that is, they only transfer to other n-back tasks but barely to other WM tasks or intelligence. In turn, if training and transfer task share specific task characteristics, transfer effects are overestimated. To overcome task-specificity and obtain a valid measurement of the latent construct of WM it is thus necessary to administer multiple validated tasks, which in turn should vary in content and paradigm, and use structural equation modeling to

decompose the different sources of variance (Bollen, 1989). If a number of WM tasks that vary in content and paradigm are used, latent variables allow to abstract away from specificities of tasks, and training induced change can be studied at the level of latent factors (Könen & Karbach, 2021; McArdle & Nesselroade, 1994). Modeling latent variables instead of manifest tasks further allows to test for measurement invariance, which addresses the question if the meaning of the latent factors stayed the same before and after the intervention. Measurement invariance is implicitly assumed when analyzing composite scores, but it is rarely tested in the training literature. Testing measurement invariance is important as training might alter not only the means of indicators and latent factors but also the measurement properties of targeted tasks. A lack of invariance can lead to misleading interpretations of the observed change (Noack et al., 2014).

Opting for a latent variable approach further comes with the advantage that one can evaluate transfer effects on hierarchical levels of consensual models of cognitive abilities (e.g., Carroll, 1993) instead of arbitrary categories of near and far distance between specific tasks (Noack et al., 2009). Rather than defining the transfer distance based on superficial task characteristics that might be difficult to generalize and agree on, the nature of transfer can be described by the level of the hierarchy at which it occurs. Transfer that occurs at the level of observed variables is likely task-specific, but transfer occurring at higher levels in the hierarchy is increasingly likely to indicate veritable change in abilities.

Several other points concerning the design and analysis of cognitive training studies have been put forward to strengthen the validity of observed effects, such as the implementation of active control groups (Au et al., 2020; Shipstead et al., 2010), tests for the specificity of transfer effects (Noack et al., 2014), or the formulation of a priori hypotheses about the size of transfer effects (McArdle & Prindle, 2008). Given the current status of WM training research, where strong evidence for far transfer at the level of latent abilities is scarce (see Schmiedek et al., 2010 for an exception), we argue that the training dosage (i.e., high), the measurement (i.e., multivariate) and the method of analysis (i.e., latent) are the pivotal features to produce and detect the intended training effects.

The Present Research

Following the above line of reasoning, we use latent variable modeling to investigate whether two years of biweekly WM training leads to reliable positive effects on WM and cognitive abilities. In a large sample of students, we implement a pretest–posttest control group design with a diverse set of training and transfer tasks and ample opportunity for training. Thus, the current study surpasses the broad majority of WM training studies with respect to sample size, training dosage, breadth of measurement, and statistical analysis (Noack et al., 2014).

In line with established results, we expect medium to large performance gains on trained WM tasks in the training group. These gains will be larger than gains in the untrained control group. We expect all WM tasks to load on a common latent WM factor. Building on the expected manifest training effects, we test for a significant gain of nontrivial magnitude in a latent WM factor in the training group that are higher than the gains in the control group.

Concerning transfer effects to intelligence, we test for significant gains of nontrivial magnitude that are expected to be smaller both at the manifest and latent ability level. Consistent with theories on the relation of WMC and fluid intelligence (Oberauer et al., 2008), as well as investment theories of cognitive abilities (Cattell, 1987), we expect larger transfer effects of WM training to fluid intelligence than to crystallized intelligence.

Method

The training study was conducted as part of a large multivariate longitudinal study on the development of student achievement. The present data have partly been used in a number of earlier articles to investigate different research questions. Hülür, Wilhelm, and Robitzsch (2011) investigated the longitudinal relation between student achievement and school grades. Hülür, Wilhelm, and Schipolowski (2011) investigated overclaiming in the nomological net of cognitive abilities using data of the fluid and crystallized intelligence tests. Gasimova, Robitzsch, Wilhelm, Boker, et al. (2014) investigated fluctuations in memory updating using dynamical systems analysis. Gasimova, Robitzsch, Wilhelm, and Hülür (2014) presented an overview of methods for modeling interindividual and intraindividual variability in longitudinal data based on data of the memory updating task. Hülür et al. (2017) reported longitudinal trajectories of German language and mathematics achievement. Finally, Hülür et al. (2018) investigated the role of intellectual engagement in the change of fluid intelligence, crystallized intelligence, and student achievement.

Participants

One hundred ninety-six students were recruited for the training group, and 137 students were recruited for the control group. Given the extent and duration of the intervention, a random allocation of participants to experimental groups was not feasible. As is to be expected in intensive longitudinal studies, a number of participants dropped out along the study period and attrition was related to experimental group membership (experimental vs. control) and sociodemographic variables. The final analysis sample consisted of $N = 112$ participants from the training group (57.1% of the pretest sample) and $N = 113$ participants in the control group for whom complete pre- and posttest data was available (82.5% of the pretest sample). Participants in the training group completed $M = 39.2$ ($SD = 1.05$) of 40 training sessions. All participants gave informed consent prior to inclusion in the study.

Undergraduate and graduate students conducted the testing/training sessions and ensured that participants felt comfortable at all times and made them feel that they were making an important contribution to research through their participation. Participants in the training group received 420€ (distributed over time), participants in the control group received 50€.

The mean age of the training group was 14.7 years ($SD = .72$) and 72 students (64%) were female. Students attended different German school tracks, with 76 students (68%) coming from *Gymnasium* (usually preparing for university), 23 students (21%) from *Realschule* (usually preparing for vocational education), and 13 students (12%) from *Gesamtschule* (comprehensive school). The no-training control group was investigated to differentiate training effects from test-retest effects and effects of normal cognitive

development across the 2-year span. It only participated in the pre- and posttest. The mean age was 14.2 ($SD = .71$) and 65 students (58%) were female. Ninety-two students (81%) attended *Gymnasium* and 21 students (19%) attended *Gesamtschule*.

Procedure

Students in the training group participated in testing sessions once every 2 weeks over a period of 2 school years to engage with the tasks on a regular basis without imposing too much temporal constraints on them. Importantly, we chose to implement such a long and regular training period to provide enough time for effects to unfold and to provide enough dosage to allow transfer of WM training to cognitive ability. At each measurement time point, a group of up to 12 students completed two measurements of WM comprising three parallel tests each (Alpha Span, Memory Updating, N-Back), achievement tests in German and in mathematics, a questionnaire on school related behavior and varying personality questionnaires (see <https://osf.io/2mpwx/> for a complete list of measures). Parallel versions of the WM tasks were generated based on task parameters and randomly assigned to measurement time points. The order of WM tasks, as well as the order of trials, was fixed within time points and subjects. Each task lasted approximately 10 minutes, resulting in a total of 1 hour of training per measurement time point or 40 hr of training over the course of the entire training phase. Thus, the training dosage surpassed the vast majority of WM training studies considered in Table 1.

The pre- and posttest were conducted in two separate sessions each and comprised a sociodemographic questionnaire, measures of fluid and crystallized intelligence, three parallel test versions of the WM tasks used in the training phase, and achievement tests in German and mathematics. Two parallel versions of each measure in the pre- and posttest were constructed. Participants were randomly assigned to one version in the pretest and worked on the other version in the posttest. Participants in the control group did not undergo any testing between the pretest and the posttest. In the following, we describe the tasks investigated in the current study.

Working Memory Tasks

The WM tasks were similar to those employed in the COGITO study (Schmiedek et al., 2010). In all tasks, items were presented in ascending difficulty and in each training session, each subject was presented with items of all difficulty levels. Although adaptive task difficulty is often considered a superior approach in WM training, the empirical evidence for this claim is questionable (e.g., Karbach & Verhaeghen, 2014; Weicker et al., 2016). Instead, experimental studies suggest that different training procedures (e.g., adaptive, self-selected, random) are equally effective as long as subjects are exposed to varying levels of task difficulty (von Bastian & Eschen, 2016).

Alpha Span

In the Alpha Span task, participants were presented with a series of single letters on a screen for a short period of time. Each letter was presented together with a number. For each letter, participants had to determine at which position of the alphabetical sequence of already presented letters it stood. They should then decide whether or not this position matched the number presented with the letter.

Each measurement consisted of eight alpha span items with eight letters (trials) each. Presentation times for the letters were 2,000 ms in half of the trials and 1,500 ms for the other half. The interstimulus interval was always 500 ms.

Memory Updating

In the memory updating task, participants saw two lines of X horizontally arranged squares. First, positive single-digit numbers ranging from 1–9 were presented in the above squares. Next, a succession of Z arithmetic operations was presented in horizontally arranged squares below. These operations, either an addition or a subtraction, had to be applied to the numbers of the squares above. The result always remained in the range between 1 and 9. After the operations, a retrieval request was presented in the squares and participants had to enter the final numbers for each square. Each measurement consisted of 8 items. Two items consisted of three numbers and six operations (six trials), four items consisted of four numbers and eight operations (16 trials), and two items consisted of five numbers and 10 operations (10 trials). The initial presentation time for the numbers was 4,000 ms, followed by an interstimulus interval of 500 ms. Afterward, the presentation time for the arithmetic operators was 2,000 ms for one half of the trials and 1,500 ms for the other half of the trials. Participants had no time limit to enter the results.

N-Back

In the spatial N-Back task, participants were presented with a grid of 4×4 cells. A stimulus was successively presented at different positions and participants had to decide whether the current position i is identical to the position of the stimulus $i - n$. Each measurement consisted of six items: Two items with 38 2-back trials, four items with 39 3-back trials, and one item with 40 4-back trials. The presentation time of the stimulus was 500 ms, and the interstimulus interval 1,000 ms in one half of the trials and 1,500 ms in the other half of the trials.

Transfer Tasks

Fluid and crystallized intelligence were measured with the Berlin test of Fluid and Crystallized Intelligence (BEFKI; Wilhelm et al., 2014). Fluid intelligence was measured with three subtests of verbal (relational reasoning), numerical (algebra word problems) and figural-spatial (figural sequences) content with 16 items each. Crystallized intelligence was measured with a declarative knowledge test comprising 64 items from the content domains of science, humanities and social sciences. Students were randomly assigned to one of two parallel versions of both tests at pre- and posttest to avoid retest effects. Parallel test versions were equated in difficulty using a linear equating procedure (Kolen & Brennan, 1995).

Statistical Analysis

Total scores were computed as percent correct score for all tasks. Total scores below guessing probability were set to missing in the intelligence tasks (1.9% of all cases). Effect sizes for individual tasks were computed as the difference between mean pre- and posttest scores divided by pretest SD . Net effect sizes were computed by subtracting the effect size in the control group from

the effect sizes in the training group. Mixed models testing the interaction of time point (pre vs. post) and group (training vs. control) were used to investigate if the manifest net effects were statistically significant.

Parcels were used as indicators in the latent factor models. Latent training and transfer effects were estimated with latent change score models (McArdle & Nesselroade, 1994; McArdle & Prindle, 2008). Three parcels for each WM task were computed as percent correct score of a sequential series of items. For fluid intelligence, three parcels based on the subtests were computed as percent correct. For crystallized intelligence, three parcels were computed as percent correct based on the three knowledge domains. All models were estimated with the maximum likelihood (ML) estimator. Latent factors were identified and scaled with the effects-coding method (Little et al., 2006). Missing data was handled with full information maximum likelihood (FIML; Schafer & Graham, 2002).

Equality constraints for strict measurement invariance across groups (training/control) and time points (pre/post) were imposed on model parameters to make the factor metric interpretable. The tenability of these constraints was tested with stepwise model tests (Little et al., 2007; Meredith, 1993). Following Hu and Bentler (1999), model fit was considered good with a comparative fit index (CFI) > .95 and root mean square error of approximation (RMSEA) > .06. A stronger emphasis was put on the CFI for the evaluation of model fit, as the RMSEA has been shown to be too conservative in models with few degrees of freedom, such as in the change models estimated in this study (Kenny et al., 2015). Deterioration in model fit caused by invariance constraints were investigated via differences in CFI. A Δ CFI > .01 was considered a substantial deterioration in fit (Cheung & Rensvold, 2002).

All statistical analyses were performed in R 4.0 (R Core Team, 2020). Latent models were estimated using the package *lavaan* (Version .6-6; Rosseel, 2012), and mixed models were estimated using the package *rstatix* (Version .6.0; Kassambara, 2020). We provide annotated syntax for the main analyses in a repository of the Open Science Framework: <https://osf.io/x8znf/>.

Results

Descriptive statistics of pre- and posttest performance across all tasks are reported in Table 2 (see Appendix A for descriptive statistics of the performance in WM task across training sessions). A comprehensive correlation matrix is provided in Appendix B.

Working Memory Capacity

Task-Wise Analysis

A substantial manifest net training effect of $d = .83$ ($p < .01$) was observed for the Alpha Span task. To investigate the training effect at the latent level, we estimated a LCSM with three indicators per time point and constraints for strict measurement invariance across time points and groups (see Appendix C and D for tests of measurement invariance). A comprehensive output of all estimated models and results is provided in the online supplement. The LCSM for the Alpha Span task had an acceptable fit with $\chi^2(37) = 55.1$, $p = .03$; CFI = .95; RMSEA = .07. Because the effects-coding method was used for scaling, the latent means and variances can be interpreted on the observed metric of the indicators. The control group improved by .03, whereas the training group improved by .13, resulting in a latent training effect of $d = 1.07$. A comparison of the estimated model with a model where

Table 2

Descriptive Statistics of the Working Memory, Fluid Intelligence and Crystallized Intelligence Subtests by Experimental Group and Time Point, as Well as Effect Sizes Within Time Points

Task	Time Point	Training						Control							
		<i>n</i>	<i>M</i>	<i>SD</i>	Skew	Kurt.	$d_{pre/post}$	<i>n</i>	<i>M</i>	<i>SD</i>	Skew	Kurt.	$d_{pre/post}$	$d_{train/control}$	
Training															
WMC Alpha Span	pre	112	.47	.12	-.77	.09		113	.49	.10	-.33	.38		-.24	
	post	110	.60	.13	-.19	-.16	1.10	113	.52	.11	-.20	.12	.26	.65	
WMC Memory Updating	pre	111	.25	.16	1.49	2.67		113	.27	.11	.37	-.56		-.12	
	post	111	.59	.23	-.09	-.88	2.07	113	.39	.15	.28	.63	1.13	1.02	
WMC N-Back	pre	112	.50	.17	-.45	-.51		113	.56	.13	-.51	-.14		-.36	
	post	111	.77	.18	-.76	-.39	1.57	113	.62	.13	-.82	.60	.46	.99	
Transfer															
Gf verbal	pre	107	.56	.16	.28	-.68		113	.61	.14	.29	-.32		-.30	
	post	112	.64	.18	-.19	-.53	.50	113	.70	.15	-.54	.14	.62	-.33	
Gf numerical	pre	110	.53	.16	.08	-.82		111	.58	.15	.22	-.60		-.30	
	post	109	.61	.17	-.35	-.35	.46	112	.63	.16	-.21	-.44	.35	-.15	
Gf figural	pre	108	.56	.17	-.23	-.58		112	.60	.16	-.13	-.35		-.23	
	post	112	.62	.20	-.13	-.61	.34	112	.65	.19	-.24	-.42	.33	-.18	
Gc science	pre	111	.56	.15	-.14	-.77		112	.59	.14	-.17	-.41		-.21	
	post	112	.62	.15	-.45	-.48	.42	112	.67	.15	-.88	.56	.61	-.33	
Gc humanities	pre	108	.48	.13	-.05	-.56		111	.51	.14	.10	-.17		-.25	
	post	109	.51	.14	.15	-.56	.25	111	.59	.16	.08	-.71	.58	-.52	
Gc social sciences	pre	102	.48	.13	-.01	-.67		111	.53	.13	.15	-.62		-.37	
	post	108	.57	.16	-.36	-.68	.73	111	.60	.18	-.18	-.69	.52	-.16	

Note. WMC = working memory capacity; Gf = fluid intelligence; Gc = crystallized intelligence; *n* = sample size; *M* = mean; *SD* = standard deviation; Kurt. = Kurtosis; *d* = Cohen's *d*. Unequal *ns* result from case-wise deletions during the data cleaning procedure reported in the methods section.

the means of the latent change factor were constrained to equality across groups indicated that the latent mean difference was significant, $\chi^2(1) = 27.6, p < .01$.

In the Memory Updating task we observed a manifest training effect of $d = .93 (p < .01)$. The LCSM that was set up just like for Alpha Span fit the data well, with $\chi^2(37) = 35.1, p = .65$; CFI = 1.00; RMSEA = .00. From pretest to posttest, participants in the control group improved by .125 while participants in the training group improved by .33. The resulting large latent training effect of $d = 1.80$ was significant, $\chi^2(1) = 53.4, p < .01$.

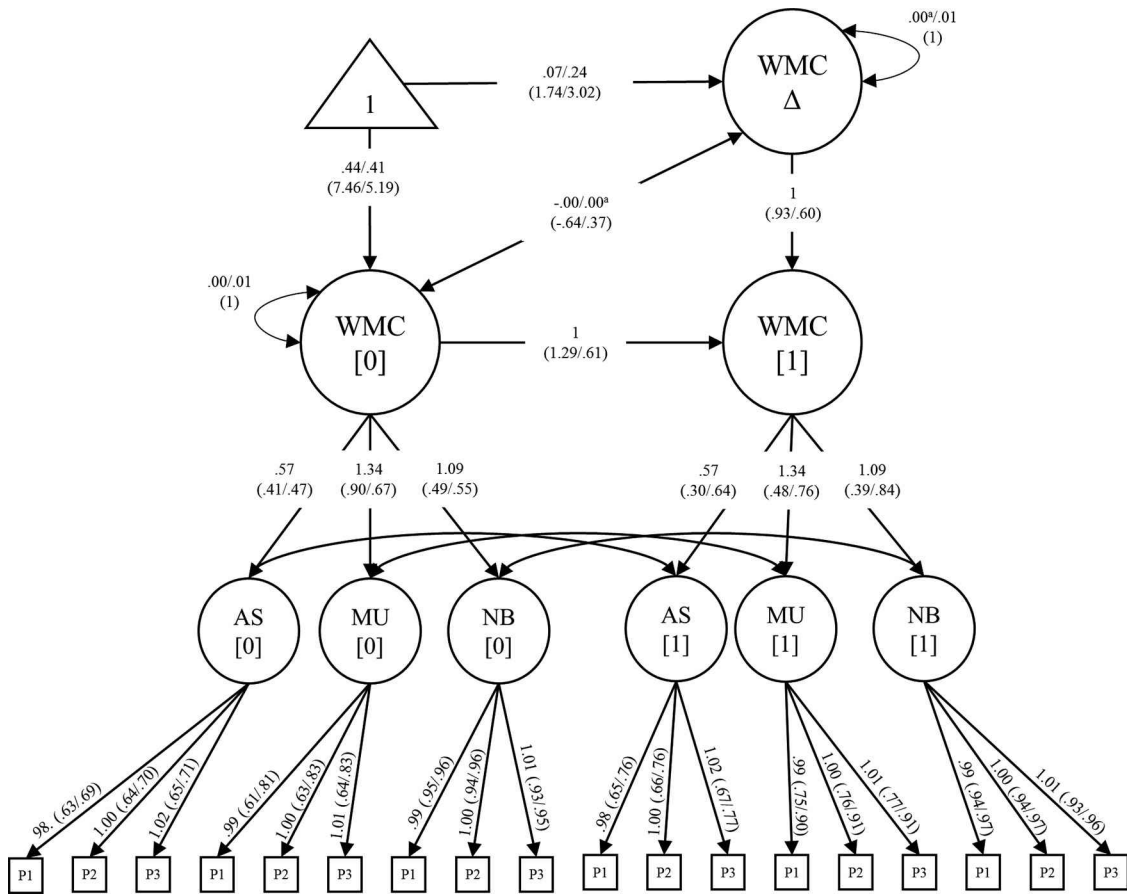
As for the previous tasks, a large manifest training effect was observed for the N-Back task, with $d = 1.12 (p < .01)$. The LCSM that was set up just like the models for the two other WM tasks had an acceptable fit of $\chi^2(37) = 111.8, p < .01$; CFI = .96; RMSEA = .13. At the latent level, participants in the control group improved by .06 and participants in the training group improved by .27. As for the previous tasks, the resulting latent training effect of $d = 1.38$ was significantly different to zero, $\chi^2(1) = 78.6, p < .01$.

Higher-Order Model

So far, manifest and latent analyses consistently indicated substantial training effects at the individual task level. However, task-level improvements are necessary but not sufficient evidence for improvements at the level of the common latent factor underlying the performance in all tasks. We therefore fitted a higher-order model of WMC to decompose task-specific variance and common-factor variance. Thereto, we estimated a LCSM where the previously reported task-level models were integrated under a common WMC factor, as illustrated in Figure 1. Residual correlations between first-order factors were allowed to accommodate variance not captured by the common WMC factor.

The higher-order LCSM of WMC had a reasonable fit, with $\chi^2(325) = 505.6, p < .01$; CFI = .94; RMSEA = .07. Substantial differences in the intercepts of the latent change factor between the experimental groups confirmed the effects observed at the task-level. While participants in the control group only improved by .07 in the percent correct metric over the two-year period, participants in the training group improved by .24, that is, about three

Figure 1
Higher-Order Latent Change Score Model for Modeling Training-Inducted Changes in the Latent WMC Factor



Note. Loadings, intercepts, and residual variances are constrained to equality across time points and groups. Parameters omitted for clarity are available in the online supplement (OS1). Estimates of the control group are reported first. Standardized estimates are reported in parentheses. “ = not significant. [0] = Pretest, [1] = Posttest, AS = Alpha Span, MU = Memory Updating, NB = N-Back, WMC = working memory capacity. $r_{AS[0], AS[1]} = .00/-.00 (.45/-.06)$, $r_{MU[0], MU[1]} = .00/.00 (.16/.19)$, $r_{NB[0], NB[1]} = .01/.00 (.46/.05)$.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

times more. This resulted in a large net latent training effect of $d = 2.37$, which was significant, $\chi^2(1) = 97.9, p < .01$. However, intercept differences from pre- to posttest in both groups (see Appendix C) indicate that the observed change in the mean structure cannot solely be attributed to changes in the latent WMC factor but might be attributed to task specific improvements that are not mirrored by improvements in the latent factor. Further, the standardized auto-regressive path between the pretest and posttest factor of WMC was lower in the training group ($\beta = .61$) than in the control group ($\beta = 1.29^1$), indicating larger changes in the rank order of participants following training. Larger changes in the training group are plausible to assume as not all participants benefit from the training equally.

Cognitive Abilities

Fluid Intelligence

The manifest training effects on individual fluid intelligence tasks were $d = -.12$ ($p = .83$; verbal), $d = .11$ ($p = .38$; numerical), and $d = .00$ ($p = .83$; figural). A LCSM of fluid intelligence fit the data well, $\chi^2(31) = 30.9, p = .47$; CFI = 1.00; RMSEA = .00 (Figure 2A) and confirmed the lack of transfer effects at the latent level. The latent training effect was $d = .08$ and not significantly different from zero, $\chi^2(1) = .40, p = .52$. Bivariate LCSMs, where change models of WMC and Gf are estimated simultaneously, allow to directly test the covariance between latent change factors. Given the lack of significant variance in the latent change factors of gf, and the absence of mean effects between both groups, such computations would be futile. Inspection of the latent change factor means of gf puts the observed effect into perspective: with an improvement of .073 in the percent correct metric, the training group was 8% (or one eighth of an item per subtest) better than the control group which improved by .065. Therefore, no generalization of training effects to fluid intelligence were present.

Crystallized Intelligence

For crystallized intelligence tasks, manifest net training effects were $d = -.18$ ($p = .31$; Science), $d = -.33$ ($p = .01$; Humanities), and $d = .21$ ($p = .14$; Social Sciences), respectively. Whereas the larger improvement in the domain of humanities for the control group was significant at the task level, no significant training effect was observed at the latent level. The LCSM of crystallized intelligence fit the data reasonably, $\chi^2(31) = 61.8, p = .01$; CFI = .95; RMSEA = .09 (Figure 2B) and indicated a latent training effect of $d = -.10$, which was not significant, $\chi^2(1) = .76, p = .38$.

Discussion

The purpose of this study was to contribute to the ongoing debate on the validity of WM training by examining whether WM training leads to improvements in latent factors of WM and cognitive ability. The training group underwent two years of training, with the goal to induce lasting changes in the cognitive system. A multivariate measurement of both WM and transfer tasks ensured that the abilities of interest were measured in adequate breadth and allowed to investigate if training effects generalized to latent abilities.

Training Effects on Working Memory Capacity and Transfer to Intelligence

Substantial and reliable training effects on all three practiced WM tasks ensured that a basic presupposition of training interventions was met: Performance on practiced tasks improved and the training group significantly outperformed the control group at posttest. An overarching latent change score model of WM confirmed that improvements were not restricted to individual tasks but were present at the level of a common latent WM factor and the magnitude of the net latent training effect was very large. Importantly, the presence of training effects at the latent level indicates a substantial degree of general improvement across the practiced tasks, that is, improvements in task-specific strategies and familiarity with the testing materials cannot fully account for the observed improvements. To our knowledge, a training effect at the factor level has only been reported once (Schmiedek et al., 2010).

The observed improvements at the task level were in line with meta-analytical effect sizes (e.g., Melby-Lervåg et al., 2016; see also Table 1) and comparable with the ones reported in a similarly powered study by Schmiedek et al. (2010). This suggests that our results are credible, but it also reveals that the length and dosage of the current intervention did not result in substantially larger training effects than are reported elsewhere. To the contrary, some short low-dosage interventions have been reported where net training effects exceeded $g = 3$ (e.g., see Figure 2 in Weicker et al., 2016). Although such effects are theoretically possible, we would caution to interpret effect sizes that are so far of the ordinary in the educational sciences (e.g., Hülür, Wilhelm, & Robitzsch, 2011) without convincing evidence for the measurement invariance of the pre- and postintervention measurement.

Despite the striking improvements in WM, we did not observe transfer to intelligence. Participants did improve in fluid and crystallized intelligence over the two-year period of this study, as is expected at this age and given some effect of familiarization with the testing material, but the group that underwent WM training did not improve more than the control group. Thus, the training-induced improvements in WM were not accompanied by significant improvements in either of two prominent factors of intelligence.

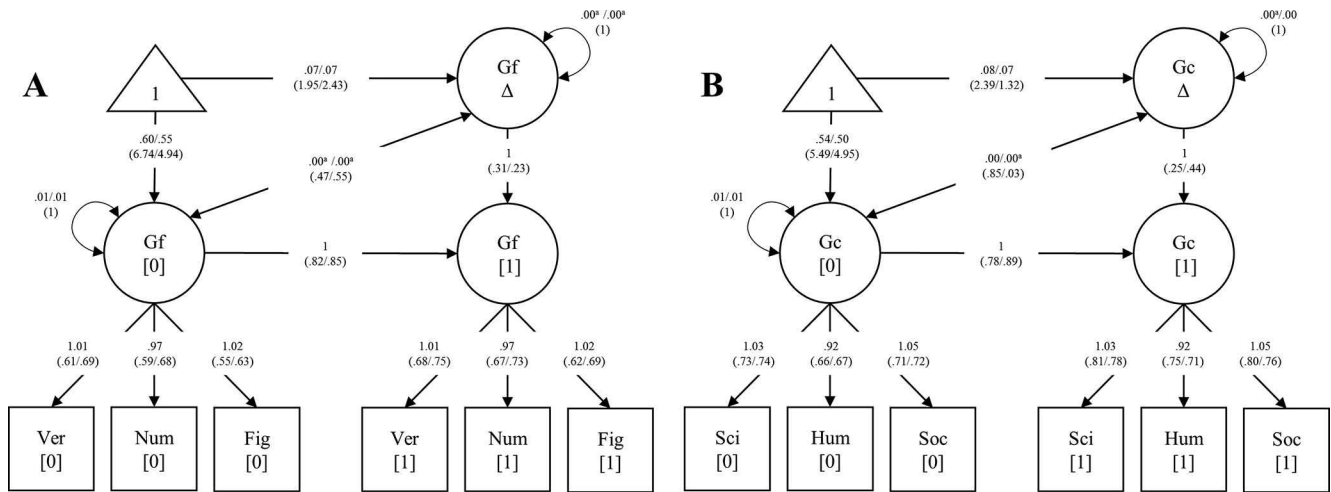
Can Transfer Effects Be Expected?

In the current study, we implemented a substantial training dosage to increase the probability of an overarching change in WM (Lövdén et al., 2010) and fluid intelligence as a closely related ability that is deemed a key ingredient of human cognitive abilities (Carroll, 1993). We chose a substantial dosage of WM training because both WM and fluid intelligence are essential in real life cognitive functioning and under constant strain throughout our lives. To provide reliable estimates for the training effect at the latent ability level, we investigated the performance of comparatively large experimental groups (Bogg & Lasecki, 2014) on broad

¹ Obviously, standardized regression coefficients greater than 1 are not plausible. We attribute this estimation problem to the constraints imposed on the autoregression and the very high pre-post correlation. In a simple, jointly estimated, correlated factor model the latter was $\rho = .83$ in the control group and $\rho = .69$ in the training group.

Figure 2

Latent Change Score Models for Modeling Training-Induced Changes in the Latent Factors of Fluid (A) and Crystallized (B) Intelligence Factors



Note. Loadings, intercepts, and residual variances are constrained to equality across time points and groups. Variable intercepts, residual variances, and covariances between repeated measurements are omitted for clarity. Estimates of the control group are reported first. Standardized estimates are reported in parentheses. * = not significant. [0] = Pretest, [1] = Posttest, Gf = Fluid intelligence, Ver = Verbal, Num = Numerical, Fig = Figural, Gc = crystallized intelligence, Sci = science, Hum = humanities, Soc = social sciences.

measurements of the investigated constructs with adequate statistical models (Noack et al., 2014). Thus, this setup addressed common theoretical and methodological shortcomings observed in the training literature. In sum, however, this approach did not lead to the effects proposed in parts of the training literature. To the contrary, our findings showed reliable evidence for the lack of transfer from WM training to intelligence, if substantial but experimentally still feasible interventions are implemented.

The lack of transfer is far from surprising and clearly in line with results from several meta-analyses. Transfer effects are often small or nil and question the utility of WM training for improving intelligence (e.g., Melby-Lervåg & Hulme, 2013, Melby-Lervåg et al., 2016; Sala, Aksayli, Tatlidil, Tatsumi, et al., 2019). Meta-analyses that did report small significant transfer effects (e.g., Au et al., 2015; Karbach & Verhaeghen, 2014) have been criticized on a methodological basis and reanalyses have shown that significant transfer effects of WM training to intelligence are biased by study selection, type of control group, and methodological flaws (Melby-Lervåg et al., 2016; see also Table 1). On an individual-study level, several attempts to replicate seminal studies reporting far transfer have repeatedly failed (Chooi & Thompson, 2012; Redick et al., 2013; and see Redick, 2015, for a discussion of issues in studies which supposedly found that WM training improves cognitive abilities). In addition, transfer effects are only a necessary, not sufficient, condition for inferring broad improvements in the trained ability. Even if transfer were found, further challenges for training effects would target the scope (i.e., process specificity of improvements) and persistence (i.e., presence of effects at catamnestic time points).

We therefore concur with Redick et al. (2013) observation that “WM transfer effects to intelligence are actually not commonly observed” (p. 373). Establishing transfer to intelligence would be an important milestone in training research. Arguably, however,

the ultimate goal of all WM training efforts is to induce changes that have a meaningful and measurable influence on real-life outcomes, for example, educational achievement or job success. Thus, if transfer to intelligence is ever established, the question still remains how much change is needed to achieve meaningful changes in these variables.

Improvement in WMC Without Improvement in Gf

Given the meta-analytical evidence on the strong correlation of WMC and fluid intelligence (Oberauer et al., 2005), as well as theories stressing the causal relationship between WMC and fluid intelligence (Oberauer et al., 2008), the lack of transfer to fluid intelligence given the substantial latent training effect on WMC is surprising. In our sample, WMC and gf were highly correlated in both the control and training group at both pre- and posttest. Although the correlations were substantial, they were not perfect and training apparently has led to changes in task performance that cannot be attributed to improvements in cognitive processes shared by WM and fluid intelligence. Longitudinal invariance tests indicated that the meaning of the WMC factor itself did not change through the intervention (as indicated by invariant factor loadings). Differences in intercepts between the pre- and posttest, however, showed that parts of the observed changes in the tasks could not be explained with changes in the latent factor alone but led to a combination of factor-level and task-level improvements.

Von Bastian and Oberauer (2014) described two general mechanisms, from which improvements in performance following WM training can result: improved WM capacity and improved WM efficiency. Improvements in capacity are arguably the aim of all WM trainings—individuals with higher capacity can, for instance, hold more chunks of information simultaneously in WM, which

benefits performance in a broad variety of cognitive tasks. It is theorized that change in capacity results from a prolonged mismatch between the available and necessary requirements of the cognitive system which leads to changes of neural structures (Lövdén et al., 2010). As a result, cognitive abilities relying on the same neural structures (e.g., fluid intelligence) should exhibit performance increases, too.

Improvements in WM efficiency, in turn, relate to a better use of the available (unaltered) WM capacity and can result from the acquisition of task-specific skills and strategies. Contrary to improvements in WM capacity, improvements in WM efficiency are generally expected to remain specific to comparable materials or tasks. Given the lack of transfer to fluid intelligence, it is more parsimonious to interpret the current results as improvements in WM efficiency rather than capacity. Interestingly, however, the improvements in the distinct WM tasks that represent substantially different paradigms to the measurement of WM (Wilhelm et al., 2013) were sufficiently correlated to manifest at the level of the common latent WM factor. Just like status, change was therefore correlated across different paradigms. The improvement at the latent factor level might therefore reflect improvements in specific processes which benefit performance in more than one task. As was illustrated repeatedly in the literature (Schmiedek et al., 2009; Wilhelm et al., 2013), tasks of very different paradigms can be equally good indicators of WM as long as they share the cognitive mechanisms of building, maintaining and updating bindings. This correlated change is a necessary, yet given our results obviously not sufficient, condition for transfer effects of WM training to fluid other constructs. Future studies will have to investigate the nature of these correlated improvements more thoroughly to rule out alternative explanations. For example, correlated change might also appear from the causally independent co-occurrence of improvements in independent skills and strategies.

Future Directions

Clearly, having participants do tests repeatedly is only one—although prevalent—form of cognitive intervention. Evidently, it did not deliver the transfer effects still suggested in many publications. Even if one was to take the optimistic position that WM training can produce meaningful transfer effects, one would need to acknowledge that they are small, that they vary interindividually, that they can only be identified reliably with latent variable modeling and that they must persist after training to be meaningful. From a purely statistical perspective, none of these points can be addressed without substantial sample sizes, comprehensive multivariate measurements, and seriously longitudinal designs over long time frames. Because small individual studies have produced inconsistent and underpowered results in the past years, one way forward might be to join forces in joint research collaborations (e.g., The Psychological Science Accelerator; Moshontz et al., 2018). A major task will be to identify determinants of cognitive malleability as, right now, meta-analytical effect sizes not only center around zero but the amount of true heterogeneity between studies is small or null (e.g., Sala & Gobet, 2020).

On the other hand, nobody does WM training solely to improve scores on a working memory test, but for the potential implications increases in intelligence have on real-life outcomes. If WM training in its current form does not provide generalizable evidence for

beneficial effects on such outcomes, individuals might be better off investing their time and energy to influence these outcomes directly (e.g., through the choice of environments which fit their abilities or interests). This is an important aspect to consider as even free, well-intentioned cognitive trainings come at an opportunity cost if they lack proof of effectiveness; even more so if we acknowledge that we need to get away from the idea that we can achieve meaningful and lasting effects with brief interventions. From this perspective, any intervention that affects the outcome of interest somehow competes with WM training. We would also like to stress a point that seems to be neglected in the WM literature. Given WM is a critical resource for all mental activities that reflect cognitive effort and given that we all use our WM intensely everyday throughout our lives, is it really reasonable that even high dosage studies such as the present intervention will cause a lasting change in WM (presupposing WM can in principle be trained)? We argue that the present results cast more doubt on this perspective. Still, the sheer number of variables which are positively correlated with WM or intelligence will always make cognitive interventions targeting these constructs attractive and worth pursuing.

If WM training in its current form delivers disappointing results, are there other interventions that do the trick of improving intellectual abilities? Basically, all other brief cognitive interventions have been criticized and called into question on similar grounds than WM training (see Moreau, 2021, for a discussion). Education, on the other hand, has been described as the “most consistent, robust, and durable method . . . for raising intelligence” (Ritchie & Tucker-Drob, 2018, p. 1358). In many ways, education fulfills requirements that have been raised for cognitive interventions, way beyond anything that can realistically be achieved in conventional intervention studies: A dosage of several hours per day over the period of many years (at least in most industrialized countries) coupled with heterogeneous “tasks” in the form of different subjects. For normally developing children, it might therefore be more instrumental to focus on early assistance and continued tutoring to ensure they can fully benefit from the school environment, than to conduct specific cognitive interventions with questionable utility.

Limitations

Some limitations of the current study need to be acknowledged. For temporal and financial reasons, our study did not include a random allocation of participants to experimental groups, as well as an active control group. Although the overall pattern of results did not indicate problems attributable to pretest differences, the lack of blinding might have led to different levels of motivation and thereby task performance across the experimental groups. Yet, given the lack of group differences in the intelligence posttests (i.e., transfer effects), we do not consider this an acute threat to the validity of our results.

Even though the participants were well remunerated, many participants from the training group dropped out during the study, presumably due to the substantial time demand of the training. Also, the average performance in and intercorrelation of the working memory tasks was comparatively low. Thus, our data comprise problems commonly encountered in demanding longitudinal studies and low-stakes ability testing of adolescents. Still, we did not

observe either floor or ceiling effects and psychometric properties of all tasks were adequate. Regarding selective attrition, we argue that this would likely inflate the probability of false-positive training effects, which we did not observe on the key outcome (i.e., transfer).

Although we did observe improvement in a latent factor of WM, it is important to bear in mind that this factor was constituted of parallel versions of the practiced tasks. Thus, this effect is better interpreted as correlated improvement in different WM tasks. To test for near transfer to WM, a set of distinct tasks would have been necessary. Also, the observed difference in indicator intercepts partly limits the interpretability of the mean structure of the latent change score models. This underlines the necessity to implement latent factor analyses in future WM training studies to substantiate commonly performed mean comparisons.

Finally, although a consensus or scientific justification for what constitutes a “typical cognitive training study” is lacking, we acknowledge that the training regimen implemented in this study differs from the majority of prior work in some regards, for example, tasks were not adaptive and while the overall dosage (total hours trained and length of the training period) was larger, the frequency of training sessions (every two weeks) was lower. Given that most cognitive interventions fail to produce an effect, we believe that to move forward it is essential to vary such key parameters of the intervention in a theoretically sound manner and to examine their effects. Our study contributes to this goal, albeit at the price of a reduced comparability with other studies.

Conclusion

In conclusion, our study lines up with and adds to the comprehensive research questioning the validity of WM training as a method to improve intelligence. In a large sample of students, forty training sessions across 2 years on a set of heterogeneous WM tasks led to substantial improvements both at the level of individual tasks as well as a common latent ability factor. However, these improvements did not transfer to fluid or crystallized intelligence. The pattern of results suggests that the observed manifest improvements cannot be attributed to improvements in the latent construct of WMC. Thus, given our results and the available meta-analytic evidence, we do not think that WM training in its current form allows to improve cognitive abilities. Because the consequences of successful interventions would be far-reaching, research into (alternative) cognitive interventions will persist. However, we are convinced that real progress will only be made if the comprehensive theoretical and methodological requirements outlined in the extant literature are considered.

References

- Aksayli, N. D., Sala, G., & Gobet, F. (2019). The cognitive and academic benefits of Cogmed: A meta-analysis. *Educational Research Review*, 27, 229–243. <https://doi.org/10.1016/j.edurev.2019.04.003>
- Alderson, R. M., Kasper, L. J., Hudec, K. L., & Patros, C. H. G. (2013). Attention-deficit/hyperactivity disorder (ADHD) and working memory in adults: A meta-analytic review. *Neuropsychology*, 27(3), 287–302. <https://doi.org/10.1037/a0032371>
- Allaway, T. P. (2009). Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment*, 25(2), 92–98. <https://doi.org/10.1027/1015-5759.25.2.92>
- Au, J., Gibson, B. C., Bunarjo, K., Buschkuhl, M., & Jaeggi, S. M. (2020). Quantifying the difference between active and passive control groups in cognitive interventions using two meta-analytical approaches. *Journal of Cognitive Enhancement: Towards the Integration of Theory and Practice*, 4(2), 192–210. <https://doi.org/10.1007/s41465-020-00164-6>
- Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuhl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, 22(2), 366–377. <https://doi.org/10.3758/s13423-014-0699-x>
- Bogg, T., & Lasecki, L. (2014). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in Psychology*, 5, 1589. <https://doi.org/10.3389/fpsyg.2014.01589>
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley; <https://doi.org/10.1002/9781118619179>
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511571312>
- Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. North-Holland.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, 40(6), 531–542. <https://doi.org/10.1016/j.intell.2012.07.004>
- Cronbach, L. (1969). Heredity, environment, and educational policy. *Harvard Educational Review*, 39(2), 338–347. <https://doi.org/10.17763/haer.39.2.nvr226676j010551>
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1), 19–23. <https://doi.org/10.1111/1467-8721.00160>
- Flanagan, J. C. & Harrison, P. L. (Eds.). (2012). *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed.). Guilford Press.
- Fogarty, G., & Stankov, L. (1982). Competing tasks as an index of intelligence. *Personality and Individual Differences*, 3(4), 407–422. [https://doi.org/10.1016/0191-8869\(82\)90006-X](https://doi.org/10.1016/0191-8869(82)90006-X)
- Gasimova, F., Robitzsch, A., Wilhelm, O., & Hülür, G. (2014). A hierarchical Bayesian model with correlated residuals for investigating stability and change in intensive longitudinal data settings. *Methodology*, 10(4), 126–137. <https://doi.org/10.1027/1614-2241/a000083>
- Gasimova, F., Robitzsch, A., Wilhelm, O., Boker, S. M., Hu, Y., & Hülür, G. (2014). Dynamical systems analysis applied to working memory data. *Frontiers in Psychology*, 5, 687. <https://doi.org/10.3389/fpsyg.2014.00687>
- Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8(3), 179–203. [https://doi.org/10.1016/0160-2896\(84\)90008-4](https://doi.org/10.1016/0160-2896(84)90008-4)
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Hülür, G., Gasimova, F., Robitzsch, A., & Wilhelm, O. (2017). An intensive longitudinal study of the development of student achievement over two years (LUISE). In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 333–354). Springer International Publishing; https://doi.org/10.1007/978-3-319-50030-0_20
- Hülür, G., Gasimova, F., Robitzsch, A., & Wilhelm, O. (2018). Change in fluid and crystallized intelligence and student achievement: The role of intellectual engagement. *Child Development*, 89(4), 1074–1087. <https://doi.org/10.1111/cdev.12791>

- Hülür, G., Wilhelm, O., & Robitzsch, A. (2011). Multivariate veränderungsmodelle für schulnoten und schülerleistungen in deutsch und mathematik [Multivariate change models for school grades and test performance in German and mathematics]. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, 43(4), 173–185. <https://doi.org/10.1026/0049-8637/a000051>
- Hülür, G., Wilhelm, O., & Schipolowski, S. (2011). Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. *Learning and Individual Differences*, 21(6), 742–746. <https://doi.org/10.1016/j.lindif.2011.09.006>
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19), 6829–6833. <https://doi.org/10.1073/pnas.0801268105>
- Jensen, A. (1969). How much can we boost IQ and scholastic achievement. *Harvard Educational Review*, 39(1), 1–123. <https://doi.org/10.17763/haer.39.1.13u15956627424k7>
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
- Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science*, 25(11), 2027–2037. <https://doi.org/10.1177/0956797614548725>
- Kassambara, A. (2020). *rstatix: Pipe-friendly framework for basic statistical tests*. <https://cran.r-project.org/web/packages/rstatix/rstatix.pdf>
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507. <https://doi.org/10.1177/0049124114543236>
- Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, 24(6), 781–791. <https://doi.org/10.1076/jcen.24.6.781.8395>
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating*. Springer. <https://doi.org/10.1007/978-1-4757-2412-7>
- Könen, T., & Karbach, J. (2021). Analyzing individual differences in intervention-related changes. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592097917–251524592097919. <https://doi.org/10.1177/2515245920979172>
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, 14(4), 389–433. [https://doi.org/10.1016/S0160-2896\(05\)80012-1](https://doi.org/10.1016/S0160-2896(05)80012-1)
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, 21(4), 861–883. <https://doi.org/10.3758/s13423-013-0565-2>
- Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development*, 31(4), 357–365. <https://doi.org/10.1177/0165025407077757>
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, 13(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3
- Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin*, 136(4), 659–676. <https://doi.org/10.1037/a0020080>
- McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 223–267). Erlbaum.
- McArdle, J. J., & Prindle, J. J. (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging*, 23(4), 702–719. <https://doi.org/10.1037/a0014349>
- Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, 49(2), 270–291. <https://doi.org/10.1037/a0028228>
- Melby-Lervåg, M., & Hulme, C. (2016). There is no convincing evidence that working memory training is effective: A reply to Au et al. (2014) and Karbach and Verhaeghen (2014). *Psychonomic Bulletin & Review*, 23(1), 324–330. <https://doi.org/10.3758/s13423-015-0862-z>
- Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer”: Evidence from a meta-analytic review. *Perspectives on Psychological Science*, 11(4), 512–534. <https://doi.org/10.1177/17456916166635612>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Mewborn, C. M., Lindbergh, C. A., & Stephen Miller, L. (2017). Cognitive interventions for cognitively healthy, mildly impaired, and mixed samples of older adults: A systematic review and meta-analysis of randomized-controlled trials. *Neuropsychology Review*, 27(4), 403–439. <https://doi.org/10.1007/s11065-017-9350-8>
- Moreau, D. (2021). How malleable are cognitive abilities? A critical perspective on popular brief interventions. *American Psychologist*. Advance online publication. <https://doi.org/10.1037/amp0000872>
- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., . . . Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, 1(4), 501–515. <https://doi.org/10.1177/2515245918797607>
- Nguyen, L., Murphy, K., & Andrews, G. (2019). Immediate and long-term efficacy of executive functions cognitive training in older adults: A systematic review and meta-analysis. *Psychological Bulletin*, 145(7), 698–733. <https://doi.org/10.1037/bul0000196>
- Noack, H., Lövdén, M., & Schmiedek, F. (2014). On the validity and generality of transfer effects in cognitive training research. *Psychological Research*, 78(6), 773–789. <https://doi.org/10.1007/s00426-014-0564-6>
- Noack, H., Lövdén, M., Schmiedek, F., & Lindenberger, U. (2009). Cognitive plasticity in adulthood and old age: Gauging the generality of cognitive intervention effects. *Restorative Neurology and Neuroscience*, 27(5), 435–453. <https://doi.org/10.3233/RNN-2009-0496>
- Oberauer, K. (2009). Design for a working memory. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 51, pp. 45–100). Elsevier. [https://doi.org/10.1016/S0079-7421\(09\)51002-X](https://doi.org/10.1016/S0079-7421(09)51002-X)
- Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schwaninger, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144(9), 885–958. <https://doi.org/10.1037/bul0000153>
- Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence—their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 61–65. <https://doi.org/10.1037/0033-2909.131.1.61>
- Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2008). Individual differences in working memory capacity and reasoning ability. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (pp. 49–75). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195168648.003.0003>
- Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between

- reading and working memory. *Psychological Bulletin*, 144(1), 48–76. <https://doi.org/10.1037/bul0000124>
- Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, 108(4), 455–473. <https://doi.org/10.1037/edu0000079>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Redick, T. S. (2015). Working memory training and interpreting interactions in intelligence interventions. *Intelligence*, 50, 14–20. <https://doi.org/10.1016/j.intell.2015.01.014>
- Redick, T. S. (2019). The hype cycle of working memory training. *Current Directions in Psychological Science*, 28(5), 423–429. <https://doi.org/10.1177/0963721419848668>
- Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, 142(2), 359–379. <https://doi.org/10.1037/a0029082>
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, 29(8), 1358–1369. <https://doi.org/10.1177/0956797618774253>
- Rossee, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology*, 53(4), 671–685. <https://doi.org/10.1037/dev0000265>
- Sala, G., & Gobet, F. (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin & Review*, 27(3), 423–434. <https://doi.org/10.3758/s13423-019-01681-y>
- Sala, G., Aksayli, N. D., Tatlidil, K. S., Gondo, Y., & Gobet, F. (2019). Working memory training does not enhance older adults' cognitive skills: A comprehensive meta-analysis. *Intelligence*, 77, 1–13. <https://doi.org/10.1016/j.intell.2019.101386>
- Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra Psychology*, 5(1), 1–22. <https://doi.org/10.1525/collabra.203>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 1089–1096. <https://doi.org/10.1037/a0015730>
- Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, 2, 1–10. <https://doi.org/10.3389/fnagi.2010.00027>
- Schmiedek, F., Lövdén, M., & Lindenberger, U. (2019). Training working memory for 100 days: The COGITO study. In J. M. Novick, M. F. Bunting, M. R. Dougherty, & R. W. Engle (Eds.), *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development*. Oxford University Press. <https://doi.org/10.1093/oso/9780199974467.003.0003>
- Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*, 50(2), 138–166. <https://doi.org/10.1080/00461520.2015.1036274>
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychologica Belgica*, 50(3-4), 245–276. <https://doi.org/10.5334/pb-50-3-4-245>
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, 138(4), 628–654. <https://doi.org/10.1037/a0027473>
- Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*, 24(4), 1077–1096. <https://doi.org/10.3758/s13423-016-1217-0>
- 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), 261–288. [https://doi.org/10.1016/S0160-2896\(01\)00100-3](https://doi.org/10.1016/S0160-2896(01)00100-3)
- Teixeira-Santos, A. C., Moreira, C. S., Magalhães, R., Magalhães, C., Pereira, D. R., Leite, J., Carvalho, S., & Sampaio, A. (2019). Reviewing working memory training gains in healthy older adults: A meta-analytic review of transfer for cognitive outcomes. *Neuroscience and Biobehavioral Reviews*, 103, 163–177. <https://doi.org/10.1016/j.neubiorev.2019.05.009>
- The Princeton Review. (2020). *10 practice tests for the SAT, 2021: Extra prep to help achieve an excellent score (College Test Preparation)*. The Princeton Review.
- Thorndike, E. L. (1906). *Principles of teaching*. Seiler.
- Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114(1), 104–132. <https://doi.org/10.1037/0033-295X.114.1.104>
- von Bastian, C. C., & Eschen, A. (2016). Does working memory training have to be adaptive? *Psychological Research*, 80(2), 181–194. <https://doi.org/10.1007/s00426-015-0655-z>
- von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: A review. *Psychological Research*, 78(6), 803–820. <https://doi.org/10.1007/s00426-013-0524-6>
- Wang, Y., Zhang, Y.-B., Liu, L.-L., Cui, J.-F., Wang, J., Shum, D. H. K., van Amelsvoort, T., & Chan, R. C. K. (2017). A Meta-analysis of working memory impairments in autism spectrum disorders. *Neuropsychology Review*, 27(1), 46–61. <https://doi.org/10.1007/s11065-016-9336-y>
- Weicker, J., Villringer, A., & Thöne-Otto, A. (2016). Can impaired working memory functioning be improved by training? A meta-analysis with a special focus on brain injured patients. *Neuropsychology*, 30(2), 190–212. <https://doi.org/10.1037/neu0000227>
- Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373–392). SAGE Publications. <https://doi.org/10.4135/9781452233529.n21>
- Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4, 433. <https://doi.org/10.3389/fpsyg.2013.00433>
- Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. Bis 10. Jahrgangsstufe* [Berlin test of fluid and crystallized intelligence for grades 8–10]. Hogrefe.

(Appendices follow)

Appendix A

Descriptive Statistics of the Performance in WM Task Across Training Sessions

Time point	Task														
	Alpha Span					Memory Updating					N-Back				
	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max
1	111	.52	.11	.05	.71	112	.33	.16	.05	.67	112	.59	.16	.08	.93
2	107	.54	.11	.30	.75	112	.40	.17	.08	.80	109	.60	.17	.19	.95
3	100	.52	.13	.10	.76	111	.40	.19	.03	.81	100	.62	.17	.27	.96
4	90	.53	.14	.10	.77	108	.44	.18	.05	.83	90	.60	.19	.12	.98
5	105	.53	.13	.17	.80	112	.46	.20	.00	.86	105	.62	.17	.21	.97
6	109	.54	.11	.20	.75	109	.43	.20	.08	.86	109	.63	.17	.05	.97
7	105	.54	.13	.12	.77	112	.44	.21	.05	.84	105	.62	.18	.18	.97
8	105	.55	.13	.20	.79	111	.45	.21	.00	.83	107	.64	.20	.12	.98
9	112	.55	.13	.12	.78	112	.49	.21	.06	.88	112	.68	.18	.19	.97
10	110	.57	.11	.23	.85	112	.53	.19	.03	.92	110	.71	.17	.21	.98
11	109	.56	.12	.27	.81	112	.53	.19	.02	.89	108	.71	.18	.25	.99
12	108	.57	.11	.30	.80	111	.53	.19	.11	.91	111	.71	.17	.26	.98
13	110	.57	.12	.20	.88	112	.53	.21	.05	.88	112	.71	.19	.10	.97
14	110	.58	.12	.20	.81	112	.54	.21	.05	.88	112	.73	.18	.05	.99
15	111	.58	.11	.27	.83	111	.53	.21	.06	.94	111	.72	.18	.15	.99
16	108	.58	.12	.27	.90	110	.54	.20	.06	.89	109	.72	.18	.21	.98
17	110	.57	.13	.12	.84	111	.53	.20	.06	.89	109	.71	.18	.15	.98
18	110	.58	.12	.22	.83	112	.56	.19	.05	.89	112	.73	.18	.07	.98
19	111	.58	.11	.24	.84	112	.55	.20	.05	.95	112	.73	.17	.25	.97
20	112	.58	.12	.21	.84	111	.56	.20	.08	.91	112	.73	.18	.23	.99
21	111	.58	.11	.21	.80	110	.54	.20	.09	.93	111	.73	.18	.08	.98
22	111	.59	.12	.24	.84	110	.55	.21	.05	.91	112	.72	.18	.24	.99
23	111	.58	.13	.21	.84	111	.55	.21	.05	.89	110	.73	.18	.22	.98
24	109	.59	.11	.36	.83	111	.56	.21	.06	.94	110	.75	.16	.24	.99
25	112	.58	.11	.24	.81	111	.54	.21	.03	.91	112	.74	.17	.21	.98
26	110	.57	.13	.14	.91	111	.56	.21	.02	.92	111	.73	.19	.18	.98
27	109	.59	.13	.24	.84	111	.56	.22	.05	.89	109	.73	.18	.17	.97
28	104	.58	.13	.26	.86	107	.53	.21	.03	.89	107	.72	.19	.18	.98
29	112	.58	.11	.33	.82	112	.54	.22	.06	.94	112	.73	.19	.19	.98
30	110	.57	.14	.10	.84	112	.54	.23	.03	.91	110	.73	.19	.22	.98
31	111	.58	.12	.21	.88	112	.53	.22	.02	.89	112	.73	.19	.20	.98
32	110	.58	.13	.27	.92	110	.54	.21	.07	.92	109	.72	.20	.20	.99
33	111	.59	.12	.23	.82	112	.54	.23	.06	.94	110	.73	.19	.22	.98
34	112	.59	.12	.22	.84	111	.54	.22	.07	.89	111	.74	.18	.18	.98
35	111	.57	.13	.16	.90	112	.55	.23	.06	.97	112	.73	.19	.18	.98
36	111	.59	.13	.23	.89	111	.56	.24	.06	.95	111	.73	.19	.17	.99
37	110	.59	.13	.00	.94	112	.56	.23	.05	.92	112	.74	.19	.18	.98
38	111	.59	.12	.29	.84	111	.56	.22	.03	.94	111	.74	.17	.32	.99
39	108	.60	.12	.32	.90	110	.55	.24	.06	.98	111	.73	.17	.27	.98
40	110	.59	.11	.33	.93	110	.54	.25	.03	.95	111	.74	.17	.27	.99

(Appendices continue)

Appendix B
Zero-Order Correlation Matrix of Tasks Within Experimental Groups

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. as_pre	—	.26**	.57**	.21*	.25**	.35**	.31**	.31**	.23*	.40**	.30**	.26**	.23*	.20*	.23*	.33**	.20*	.19*
2. mu_pre	.28**	—	.21*	.22*	.40**	.39**	.27**	.24*	.15	.19*	.37**	.25**	.10	.05	.13	.13	.23*	.20*
3. nb_pre	.33**	.34**	—	.38**	.42**	.48**	.42**	.38**	.40**	.53**	.43**	.49**	.41**	.23*	.13	.39**	.31**	.36**
4. as_post	.36**	.16	.29**	—	.41**	.52**	.33**	.36**	.22*	.25**	.31**	.30**	.26**	.15	.10	.27**	.25**	.25**
5. mu_post	.09	.31**	.28**	.03	—	.59**	.36**	.29**	.24*	.40**	.41**	.32**	.25**	.12	.16	.30**	.29**	.25**
6. nb_post	.23*	.06	.47**	.42**	.16	—	.33**	.42**	.31**	.49**	.51**	.41**	.31**	.24*	.12	.39**	.26**	.24*
7. gfv_pre	.22*	.04	.06	.26**	.07	.14	—	.52**	.43**	.57**	.51**	.44**	.53**	.37**	.23*	.40**	.35**	.40**
8. gfn_pre	.35**	.12	.17	.17	.04	.11	.39**	—	.42**	.52**	.63**	.44**	.36**	.32**	.25*	.38**	.31**	.38**
9. gff_pre	.25**	.19*	.19*	.12	.25**	.18	.37**	.43**	—	.44**	.42**	.60**	.38**	.24*	.32**	.47**	.30**	.36**
10. gfv_post	.25**	.18	.24*	.39**	.21*	.29**	.41**	.18	.36**	—	.50**	.46**	.49**	.43**	.31**	.54**	.41**	.50**
11. gfn_post	.46**	.29**	.33**	.18	.26**	.14	.44**	.47**	.39**	.39**	—	.50**	.37**	.20*	.36**	.45**	.36**	.46**
12. gff_post	.34**	.18	.34**	.30**	.17	.29**	.41**	.48**	.62**	.38**	.45**	—	.40**	.27**	.20*	.54**	.39**	.34**
13. sci_pre	.20*	-.00	.29**	.11	.01	.20*	.30**	.33**	.26**	.33**	.35**	.35**	—	.56**	.49**	.65**	.53**	.49**
14. hum_pre	.32**	.06	.15	.21*	.07	.16	.37**	.28**	.31**	.22*	.31**	.37**	.52**	—	.40**	.42**	.48**	.44**
15. soc_pre	.31**	.07	.18	.09	.05	.17	.24*	.35**	.20*	.20*	.42**	.29**	.63**	.43**	—	.42**	.40**	.50**
16. sci_post	.29**	.15	.21*	.25**	.10	.21*	.30**	.28**	.25**	.43**	.45**	.32**	.58**	.45**	.57**	—	.49**	.57**
17. hum_post	.25**	.01	.11	.12	.05	.20*	.28**	.17	.24*	.20*	.29**	.26**	.50**	.60**	.62**	.66**	—	.51**
18. soc_post	.32**	.06	.26**	.18	-.01	.22*	.26**	.44**	.26**	.23*	.50**	.35**	.59**	.49**	.66**	.57**	.52**	—

Note. Correlations of the control group are below the diagonal, correlations of the training group are above the diagonal.
* $p < .05$. ** $p < .01$.

(Appendices continue)

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix C
Measurement Invariance Across Groups at Pretest

Task	χ^2	<i>df</i>	<i>p</i>	CFI	RMSEA
Alpha Span					
Configural	0	0		1	0
Weak	2.5	2	.29	.997	.045
Strong	4.3	4	.36	.998	.027
Strict	6.8	7	.45	1	0
Memory Updating					
Configural	0	0		1	0
Weak	2.0	2	.37	1	0
Strong	3.6	4	.46	1	0
Strict	7.5	7	.38	.997	.025
N-Back					
Configural	0	0		1	0
Weak	.08	2	.96	1	0
Strong	.94	4	.92	1	0
Strict	3.6	7	.82	1	0
WMC					
Configural	65.6	48	.05	.985	.057
Weak	77.0	56	.03	.982	.058
Strong	84.4	64	.04	.983	.053
Strict	111.0	76	.01	.971	.064
Gf					
Configural	0	0		1	0
Weak	1.4	2	.49	1	0
Strong	1.6	4	.81	1	0
Strict	1.7	7	.98	1	0
Gc					
Configural	0	0		1	0
Weak	1.7	2	.42	1	0
Strong	6.0	4	.20	.988	.066
Strict	9.6	7	.21	.984	.057

Note. *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation. Bolded values indicate a deterioration of CFI > .01 compared with the previous model.

(Appendices continue)

Appendix D

Longitudinal Measurement Invariance

Task	Control					Training				
	χ^2	<i>df</i>	<i>p</i>	CFI	RMSEA	χ^2	<i>df</i>	<i>p</i>	CFI	RMSEA
Alpha Span										
Configural	14.3	8	.07	.952	.084	12.1	8	.14	.980	.075
Weak	16.7	10	.08	.950	.077	16.7	10	.08	.967	.085
Strong	19.1	12	.09	.946	.072	25.4	12	.01	.941	.104
Strict	20.1	15	.17	.961	.055	28.7	15	.02	.939	.095
Memory Updating										
Configural	8.2	8	.42	.999	.013	4.1	8	.85	1	0
Weak	10.3	10	.41	.997	.017	4.7	10	.91	1	0
Strong	18.4	12	.10	.948	.069	5.4	12	.95	1	0
Strict	19.9	15	.18	.960	.054	7.5	15	.94	1	0
N-Back										
Configural	9.7	8	.29	.998	.043	19.0	8	.01	.990	.111
Weak	11.1	10	.35	.998	.032	19.7	10	.03	.991	.093
Strong	18.0	12	.12	.992	.067	33.4	12	.00	.980	.126
Strict	21.9	15	.11	.990	.064	76.1	15	.00	.944	.191
WMC										
Configural	156	125	.03	.970	.047	160	125	.02	.981	.050
Weak	165	133	.03	.970	.046	181	133	.00	.974	.057
Strong	198	141	.00	.945	.060	218	141	.00	.959	.070
Strict	219	153	.00	.936	.062	273	153	.00	.936	.084
Gf										
Configural	9.3	5	.10	.977	.088	1.5	5	.91	1	0
Weak	11.8	7	.11	.974	.078	2.4	7	.94	1	0
Strong	18.1	9	.03	.951	.094	3.2	9	.95	1	0
Strict	20.5	12	.06	.955	.079	6.6	12	.88	1	0
Gc										
Configural	20.5	5	.00	.954	.166	2.4	5	.80	1	0
Weak	23.2	7	.00	.952	.143	8.8	7	.27	.993	.048
Strong	24.5	9	.00	.954	.124	18.4	9	.03	.961	.096
Strict	33.5	12	.00	.936	.126	20.1	12	.06	.966	.078

Note. *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation. Bolded values indicate a deterioration of CFI > .01 compared with the previous model.

Received March 3, 2021
Revision received February 1, 2022
Accepted February 4, 2022 ■