**This manuscript is published in:**

von Bastian, C. C. & Eschen, A. (2016). Does working memory training have to be adaptive? *Psychological Research, 80*(2), 181-194. doi: 10.1007/s00426-015-0655-z

**The final publication is available at www.springerlink.com via http://dx.doi.org/10.1007/s00426-015-0655-z**

# Does Working Memory Training Have to Be Adaptive?

Claudia C. von Bastian
Department of Psychology
University Research Priority
Program ‘‘Dynamics of Healthy Aging’’
University of Zurich, Switzerland

Anne Eschen
Department of Psychology
International Normal Aging and
Plasticity Imaging Center (INAPIC),
University Research Priority
Program ‘‘Dynamics of Healthy Aging’’
University of Zurich, Switzerland

Correspondence concerning this article should be addressed to Claudia C. von Bastian, now at the Department of Psychology, Bournemouth University.

E-mail: cvonbastian@bournemouth.ac.uk

Abstract

This study tested the common assumption that, to be most effective, working memory (WM) training should be adaptive (i.e., task difficulty is adjusted to individual performance). Indirect evidence for this assumption stems from studies comparing adaptive training to a condition in which tasks are practiced on the easiest level of difficulty only [cf. Klingberg (Trends Cogn Sci 14:317–324, 2010)], thereby, however, confounding adaptivity and exposure to varying task difficulty. For a more direct test of this hypothesis, we randomly assigned 130 young adults to one of the three WM training procedures (adaptive, randomized, or self-selected change in training task difficulty) or to an active control group. Despite large performance increases in the trained WM tasks, we observed neither transfer to untrained structurally dissimilar WM tasks nor far transfer to reasoning. Surprisingly, neither training nor transfer effects were modulated by training procedure, indicating that exposure to varying levels of task difficulty is sufficient for inducing training gains.

*Keywords:* cognitive training, adaptive training, transfer, working memory capacity

# Does Working Memory Training Have to Be Adaptive?

Can fluid cognitive abilities such as working memory (WM) and reasoning be improved through computer-based WM training? This is a highly controversial question, with prior empirical studies (for reviews, see Morrison & Chein, 2011; von Bastian & Oberauer, 2014) and meta-analyses (Au et al., in press; Karbach & Verhaeghen, 2014; Lampit, Hallock, & Valenzuela, 2014; Melby-Lervåg & Hulme, 2013) providing contradictory findings. Although multiple previous studies revealed promising effects (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Jaeggi et al., 2010; Klingberg et al., 2005; Schmiedek, Lövden, & Lindenberger, 2010; Schweizer, Hampshire, & Dalgleish, 2011; Stepankova et al., 2014; von Bastian & Oberauer, 2013), a growing number of other WM training interventions failed to induce such broad transfer (e.g., Chein & Morrison, 2010; Chooi & Thompson, 2012; Colom et al., 2013; Harrison et al., 2013; Redick et al., 2013; Salminen, Strobach, & Schubert, 2012; Sprenger et al., 2013; Thompson et al., 2013; von Bastian, Langer, Jäncke, & Oberauer, 2013). The factors contributing to the success of WM training interventions in terms of improving WM and reasoning are still unclear (see von Bastian & Oberauer, 2014), and large variations (and, in some occasions, serious flaws) in the methodologies and training regimens used complicate comparisons across studies (cf. Shipstead, Redick, & Engle, 2012), and thus the identification of such factors. Therefore, before we can conclude whether and under which circumstances WM training can induce transfer, carefully controlled, systematic investigations of factors potentially contributing to training effectiveness are needed.

In theory, cognitive plasticity occurs if there is a "prolonged mismatch between functional organismic supplies and environmental demands" (Lövden, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010, p. 659). According to Lövden and colleagues (2010), this mismatch occurs if the environmental demands exceed the routine demands the cognitive system usually faces. If those environmental demands are too high, however, individuals might simply give up on the task or develop task-specific strategies to solve this seemingly otherwise unsolvable task. Hence, in order to trigger cognitive plasticity, the authors argue that the demands should still be manageable with the current range of functional supplies. In other words, improvement in cognitive abilities such as WM can be induced by constantly challenging individuals slightly above their current routine performance level. Hence, the authors suggest that cognitive training programs should be adaptive, that is during training, task difficulty should be continuously adjusted automatically to the individual's current level of performance to maximize and prolong the supply-demand mismatch. WM training studies showing larger performance gains after adaptive than low-level training seem to support this theoretical assumption (Brehmer, Westerberg, & Bäckman, 2012; Karbach, Strobach, & Schubert, in press; see Klingberg, 2010 for an overview).

However, participants in the adaptive training condition do not only experience adjustment of task difficulty to individual performance, but are also exposed to various levels of task difficulty, whereas participants in the low-level training condition practice constantly on the easiest level of task difficulty only. Thus, adapting task difficulty to individual performance and exposure to varying levels of task difficulty are confounded in those studies. Such varying levels of task difficulty, however, pose constantly changing environmental demands forcing the cognitive system continuously out of its routines and hence could be sufficient to trigger cognitive plasticity. In line with this assumption, Schmidt and Bjork (1992) gave an overview of motor and verbal concept training studies demonstrating that training with variability in task demands leads to greater transfer effects than training with constant task demands.

In the present study, we therefore tested the hypothesis that adaptive WM training is superior to other training procedures because task difficulty is continuously adapted to individual performance instead of being varied

performance-independently, thus differentiating between adaptivity and variability of task difficulty. Hence, adaptive training was compared to another WM training procedure in which task difficulty varied randomly. In addition, a third WM training procedure was included in which participants themselves could modify training task difficulty. The purpose of this training procedure was to explore whether change in training task difficulty across the training period in the adaptive training condition approximately matches what the average individual would choose as the optimal modification of task difficulty across training. Finally, to evaluate whether we could replicate our earlier findings showing benefits after adaptive WM training on untrained, structurally dissimilar WM and reasoning tasks(von Bastian & Oberauer, 2013), we added an adaptive active control group solving trivia quizzes with low WM demand.

In our pretest-posttest study design, we aimed at avoiding the methodological issues occasionally observed in previous training research. First, training tasks were selected both theory-driven and based on empirical findings. We chose the complex span paradigm, which is a well-established measure of WM capacity (cf. Conway et al., 2005), as well as an excellent predictor for reasoning (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). Moreover, in our recent study mentioned above (von Bastian & Oberauer, 2013), we found that training with complex span tasks was more effective than training with other tasks of WM capacity in terms of transfer to untrained WM and reasoning tasks. Second, the training regimen was intensive (20 sessions within four weeks, each lasting approximately 30 min) and followed recommendations for facilitating transfer effects such as providing variability and feedback (Schmidt & Bjork, 1992). To enhance variability, each group practiced three different tasks (each for approximately 10 min per session). Feedback was provided after each trial, after each task, and across sessions at the beginning of each session. Third, we assessed each transfer range (intermediate transfer to structurally dissimilar WM tasks and far transfer to reasoning tasks) with multiple indicators to avoid task-specific features being responsible for the detection of transfer effects (cf. Noack, Lövden, & Schmiedek, 2014; Shipstead, et al., 2012). Fourth, the study included a relatively large sample of $N$ = 130 participants.

## METHOD

Over the course of four weeks, participants completed 20 sessions of intensive cognitive training (approximately 30-45 min per session). They were randomly assigned to one of the three WM training procedures (adaptive, randomized, or self-selected task difficulty) or an adaptive active control group practicing tasks with low WM demand (trivia questions on general knowledge). The study was double-blind, hence neither the participants nor the experimenters collecting the outcome measures were aware of which group the participants were assigned to. To assess training and transfer effects, we administered a test battery immediately before and after training. For facilitating between-groups baseline comparisons, which are essential for establishing the comparability across groups and occasions, we used an identical test battery at both assessments.

### PARTICIPANTS

Participants were recruited from the participant pools of the Department of Psychology and the International Normal Aging and Plasticity Center of the University of Zurich, and through advertisements at the campuses of several universities in Zurich. Participants were informed that they would take part in a cognitive training study, but not about the different training conditions. All participants were German native speakers or highly proficient in German, and gave written consent to participate. Of the overall 145 recruited individuals, 8 dropped out during the training phase due to lack of time (2), loss of interest (1), or technical issues (1). Four participants withdrew consent

without comment. We excluded four additional participants as they lacked compliance in proceeding with the training sessions, and three participants as they reported medical conditions potentially impacting cognitive functioning (traumatic brain injury, epilepsy, or medication with possible cognitive side-effects). Basic demographics of the remaining 130 individuals (93 female, $M_{age}$ = 23, *SD* = 3, age range 18-34 years) who completed the study are listed in Table 1. There were no significant group differences in these variables. At study completion, participants received CHF 80 (about USD 88) or course credits.

**Table 1. Participant Demographics**

| | Group | | | |
|---|---|---|---|---|
| Demographics | Adaptive | Randomized | Self-Selected | Active Control |
| Sample size (*n*) | 34 | 30 | 34 | 32 |
| Gender (f/m) | 25/9 | 21/10 | 24/10 | 23/9 |
| Age (*M* ± *SD*) | 23.00 ± 3.01 | 22.50 ± 3.33 | 23.12 ± 3.80 | 23.00 ± 3.05 |

## DESIGN AND MATERIALS

### TRAINING

Training was self-administered at home using Tatool (von Bastian, Locher, & Ruflin, 2013), a Java-based open-source training and testing tool (www.tatool.ch). After each training session, data were automatically uploaded to a web server running Tatool Online, which permits to constantly control participants' compliance. Several measures were taken to maximize compliance and experimental control, such as automated online analyses of training data for detecting irregularities (e.g., accuracy below chance level). Another experimenter than those collecting the outcome measures monitored the participants' training compliance and served as their contact during training. To increase individual commitment, participants signed a participant agreement and were informed that their training data would be monitored. To stay in regular contact with the participants, they received e-mails at multiple events (e.g., when half of the training sessions were completed, or when the time since the last data upload exceeded two days). In addition, participants could always contact the experimenters in case of technical difficulties. For each group, the training intervention comprised three tasks (each approximately 10 min per session), the order of which was randomized for each session.

### WM TRAINING TASKS

Modeled after the storage and processing training intervention in an earlier study (von Bastian & Oberauer, 2013), WM training consisted of three complex span tasks (Conway, et al., 2005; Daneman & Carpenter, 1980) with varying material (numerical, verbal, and figural-spatial). In these tasks, the presentation of memoranda (each for 1 s) alternates with a secondary distractor task, in which participants have to make a decision as quickly and as accurately as possible. After a certain number of memory/decision-sequences (i.e., the set size), participants have to recall the memoranda in correct serial order, for which they have unlimited time. In the numerical version, participants had to memorize two-digit numbers and judge the correctness of equations. In the verbal complex span, letters served as memoranda and a lexical decision (word vs. non-word) had to be made on strings of characters. In the figural-spatial version of the task, memoranda were positions (i.e., colored squares) in a 5 x 5 grid. In-between the display of memoranda, participants had to decide whether the long side of an L-formed shape composed of colored squares displayed in the grid was oriented horizontally or vertically. In each session, participants completed up to 12 trials in each task. As the level of difficulty was varied by adjusting the set size, trial

length increased with difficulty (see also Chein & Morrison, 2010; von Bastian & Oberauer, 2013). To keep the average duration of training sessions between 30-45 min, each task ended when task duration exceeded 15 minutes.

## ACTIVE CONTROL TRAINING TASKS

Participants had to solve trivia questions on general knowledge with four alternative answers, one of them being correct. To hold variability of the training tasks constant across groups, the active control training also comprised three task versions, which differed in respect to the subject (geography, history, and natural science). Participants completed 50 trials per task and session. The level of difficulty was raised by presenting increasingly difficult questions. We ran a pilot-study to determine the questions' difficulties (i.e., the percentage of correct answers for each question). We then rank-ordered the questions by their difficulty and assigned 50 questions to each training level (e.g., the 50 easiest questions were assigned to level 1). Thus, questions were repeated in case participants remained on the same training level across multiple sessions.

## TRAINING ALGORITHMS

Depending on WM training condition, task difficulty was adjusted adaptively, varied randomly, or was self-selected. Apart from this manipulation, we aimed at maximizing the between-groups comparability regarding the overall task difficulty across the training phase. All WM training groups started all training tasks at the same level of difficulty with three memoranda, and the active control group started with the 50 easiest questions. In the *adaptive* WM training condition and in the active control training, task difficulty was adjusted to individual performance using the default adaptive score and level handler included in Tatool (see von Bastian, Locher, et al., 2013), and corresponded to the presentation of one additional memorandum or one less (WM training) or more challenging or easier quiz questions (active control), respectively. Task difficulty was increased if participants scored at least 80% correct in the preceding session or decreased in case performance dropped below 60%. In WM training, participants had to additionally score 80% correct in the processing component of the complex span task in order to move up a level.

In the *randomized* WM training condition, task difficulty varied randomly and independently of individual performance between 3 and 9 memoranda. We chose this range because it approximates the range which most participants practiced on in a previous study that implemented a similar adaptive WM training regimen as the adaptive one used in the present study (von Bastian & Oberauer, 2013). In the *self-selected* WM training condition, participants were instructed to modify the task demands themselves by setting the level of task difficulty for the next training session at the end of each task. Task difficulty could be set to remain on the same level, to increase, or to decrease one level (i.e., one additional memorandum or one less). This mirrors the range of possible change in task difficulty in the adaptive WM training condition from session to session as well as across sessions (i.e., 3 to 22 memoranda due to the total number of 20 sessions).

## TRAINING FEEDBACK

Participants in all conditions received performance-based feedback across sessions, after each trial, and after each task. Feedback across sessions was presented at the beginning of each session, visualized in form of a graph plotting level against session for each task. Trial-by-trial feedback was presented as a green check mark for a correct response, and a red cross for a wrong answer. In addition, after each task, participants received feedback

visualized as 1 to 5 stars[1] reflecting their overall performance in this task in the current session. After receiving feedback about their task performance in the current session, participants in the self-selected WM training condition were asked to choose the level of task difficulty for that task in the next session. At the same time, participants in the other conditions were informed about the level of task difficulty they would practice on in the next session. Thus, participants in the self-selected condition could make informed decisions without sacrificing comparability between conditions regarding the quantity of instructions and information about the upcoming level of difficulty.

## TRAINING QUESTIONNAIRES

At the end of each session, participants were asked to complete a short questionnaire comprising two questions adapted from the Intrinsic Motivation Inventory (Deci & Ryan, n.d.) on their enjoyment and effort concerning the training tasks ("Today's training session was fun to do" and "I tried to do well in today's training session", respectively), and one question on the perceived fit between difficulty and ability ("The difficulty of today's training session was just right"). They had to indicate their agreement or disagreement with these statements on a 7-point scale (1 = *does not apply at all*, 7 = *does apply very well*). In addition, participants were asked to indicate their arousal and valence on a 9-point scale using self-assessment manikins (Bradley & Lang, 1994). These data will be reported elsewhere. As a further measure of motivation, participants completed the Questionnaire on Current Motivation (QCM, Rheinberg, Vollmeyer, & Bruns, 2001) after the first training session (in which all participants practiced on the same level of difficulty) and after the tenth training session (i.e., after half of the training intervention was completed). The QCM comprises 18 items that assess four factors of achievement motivation (anxiety, probability of success, interest, and challenge) in current learning situations.

## PRE AND POSTASSESSMENTS

Before and after the training intervention, we administered a test battery comprising three tests assessing practice effects in the WM training tasks, three tests measuring intermediate transfer to untrained and structurally dissimilar WM tasks, and five tests determining far transfer to reasoning. In addition, participants completed a control test (trivia quiz) to which we did not expect any transfer. Participants were tested in groups of up to four individuals in one lab session that took about 3 h including two 10 min breaks. Half of the participants in each group completed the test battery in reverse order (relative to the other half of participants) to control for linear effects of fatigue and practice. For each task, participants completed several practice trials preceding test blocks of pseudo-randomized trials. The tasks were programmed with Java in Tatool (von Bastian, Locher, et al., 2013).

In addition to cognitive assessment, participants were asked to complete several questionnaires preceding the pretest and at the posttest assessment (Need for Cognition, Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994;

[1] The number of stars corresponded to the proportion of correct responses: 5 stars for at least 80 % correct, 4 stars for more than 70% correct, 3 stars for more than 60 % correct, and 2 stars for less than 60 % correct. In WM training, 1 star was given if recall performance was less than 60 % or performance in the processing task was below 80 % (having at least 80 % correct in the processing task was a prerequisite to receive any higher number of stars). In the active control condition, participants received 1 star if performance was below 50 %.

NEO-FFI, Borkenau & Ostendorf, 2008; Cacioppo & Petty, 1982; Costa & McCrae, 1992; Intrinsic Motivation Inventory, Deci & Ryan, n.d.; Theories of Intelligence Scale, Dweck, 1999; Cognitive Failures Questionnaire, Klumb, 2001; Prospective and Retrospective Memory Questionnaire, Smith, Del Sala, Logie, & Maylor, 2000), the results of which will be reported elsewhere.

### TRAINING TASKS.

To compare practice effects between the training conditions, we administered three complex span tasks. The design and type of material was the same as for the training tasks. Each of the tasks consisted of 16 trials with varying set sizes (4 to 7 memoranda). The proportion of items recalled at the correct position served as dependent variable (partial-credit unit score; for details, see Conway, et al., 2005).

### INTERMEDIATE TRANSFER TASKS

To measures intermediate transfer, participants completed three tasks which are assumed to capture WM but are structurally dissimilar to the complex span training tasks (von Bastian & Oberauer, 2013; Wilhelm, Hildebrandt, & Oberauer, 2013).

#### WORD-POSITION BINDING TASK

Participants had to memorize the positions of 3 to 5 words presented sequentially on the screen (cf. the local recognition task in Oberauer, 2005). Each word was displayed for 2 s. Probe words in a different color were shown immediately afterward. Participants had to decide for each of the probes whether it matched the word previously shown at this exact position. Probes not matching the original stimulus at that position could be new probes (distractors not presented anywhere in the list) or intrusion probes (words presented in the list but at a different position). Whereas new probes can be correctly rejected based solely on item recognition, correct rejection of intrusion probes requires recollection of the word-position binding. Participants completed two blocks of 8 trials per set size. Across all 48 trials, 50% of the probes were positive, 25% were negative new probes, and 25% were negative intrusion probes. The positive probes were distributed equally across temporal and spatial positions. Scores were derived by computing the discrimination parameter d' from signal-detection theory, taking hits and false alarms to intrusion probes into account: $d' = z(\mathrm{H}) - z(\mathrm{FA})$, where $H$ is the hit rate, $FA$ the false alarm rate, and $z$ refers to the $z$ value corresponding to the probability of the given argument. Using only false alarms to intrusion probes, $d'$ serves a pure measure of binding memory (see also von Bastian & Oberauer, 2013).

#### BROWN-PETERSON TASK

We adapted the classical Brown-Peterson paradigm (Brown, 1958) to serve as a dual task combining a simple span and a distractor decision task. Participants first had to memorize sequentially presented words, and then to decide for a series of letter pairs whether they rhyme (e.g., "A" and "K") or not (e.g., "A" and "E"). After four such decisions, participants had to recall the words memorized before in correct serial order. The task consisted of 16 trials with set sizes varying between 3 and 6. As for the complex span tasks, scores were derived from the proportion of items recalled at the correct position.

#### MEMORY UPDATING TASK

In this task, participants have to constantly manipulate and update information (cf. Oberauer, 2006; see also von Bastian & Oberauer, 2013). Each trial started with the simultaneous presentation of 1 to 3 digits, which were shown in different colors (blue, orange, and purple). Afterward, participants had to complete a series of 20 arithmetic operations (additions or subtractions indicated by signs in the digit colors) that had to be applied to the digit in the same color. The previously memorized digit had to be replaced by the result of the operation and the result had to be entered via the keyboard. All digits (i.e., memoranda, summands or subtrahends, and the results) ranged from 1 to 9. Participants had to complete 24 trials presented in three blocks. The proportion of correct responses to the arithmetic operations served as score.

## FAR TRANSFER TASKS

Transfer to reasoning was assessed with five tasks adapted from standard test instruments.

### RAVEN'S ADVANCED PROGRESSIVE MATRICES (RAPM, RAVEN, 1990)

Participants had to complete a pattern presented by choosing one of eight alternatives. We used the 12-item short version developed by Arthur and Day (1994; see also Arthur, Tubre, Paul, & Sanchez-Ku, 1999).Time for completion was restricted to 15 min.

### LETTER SETS TEST (EKSTROM, FRENCH, HARMAN, & DERMEN, 1976)

Five sets of four letters were presented. Except for one set, all sets followed a certain logical pattern. The task was to choose which of the letter sets deviated from the others. Participants had 14 min to complete 30 problems.

### LOCATIONS TEST (EKSTROM, ET AL., 1976)

In this task, five rows of dashes separated by blank spaces are given. In the first four rows, one dash is replaced by an “x”, following a certain pattern across rows. Participants have to discover the rule and to choose which position of the “x” out of five is the correct one in the fifth row. Participants had 12 min to complete 28 problems.

### NONSENSE SYLLOGISMS TEST (EKSTROM, ET AL., 1976)

The task was to judge whether the conclusion drawn from two premises was logically valid (e.g., following the premises “all trees are fish” and “all fish are horses”, it would be logically correct to conclude that “therefore all trees are horses”). Nonsensical content was used to avoid the scores being influenced by past learning. Participants had 8 min to complete 30 problems.

### DIAGRAMMING RELATIONSHIPS

Sets of three nouns (e.g., animals, cats, and dogs) were presented. Participants had to choose which one out of five diagrams represents the relationship between the nouns (in this example, one circle representing animals containing two separate circles representing cats and dogs, respectively). Participants had 8 min to complete 30 problems.

### CONTROL TASK

We included a trivia quiz as a control test to which we did not expect any transfer of WM training (cf. Noack, et al., 2014). In addition, the test served to increase the believability of the control condition because all participants experienced a task in pre and post assessment that was similar to their training tasks. The test included 30 questions which were drawn from the same subjects (i.e., geography, history, and natural science) but had not been presented during the active control training. Therefore, the knowledge required to solve these questions could not have been acquired in the active control training. Hence, the control group was not expected to perform better than the WM training groups in this task as a result of training. In addition, another response format than in the training version was chosen (open text instead of multiple choice questions). Time for responses was not restricted.

## RESULTS

### MISSING DATA

Due to technical issues at pretest, data for the memory updating task were lost for one participant. Data of three participants are missing for the QCM assessment after the tenth training session. Participants with missing data were excluded from analyses including the respective measure. Some participants had difficulties scheduling their training sessions and hence did not complete the required 20 sessions, but only 17 (1 participant), 18 (1 participant) or 19 sessions (9 participants), while one participant completed 21 training sessions. For the analyses of training progress, we included only participants with complete training data sets. For the analyses of training and transfer gains, the results were qualitatively similar, independent of whether the participants with less or more than 20 sessions were included or excluded in the analyses. Therefore, we included also participants with irregular numbers of training sessions to maximize power.

### GROUP COMPARABILITY AT BASELINE

To determine whether baseline cognitive performance was comparable across groups, we first conducted a multivariate analysis of variance (MANOVA) with all pretest measures as dependent variables. The main effect of group was not significant, $F(36, 348) = 1.09$, $p = .345$, $\eta_p^2 = .10$. In addition, none of the Bonferroni corrected post-hoc between-groups comparisons for single tasks was significant, with one exception. The adaptive training group showed worse baseline performance in the figural complex span training task than the active control group ($M_{diff} = .13$, $p = .033$) with a medium effect size ($d = 0.67$). Table 2 lists the means and standard deviations for each group in each cognitive task.

**Table 2. Mean Performance for the Test Battery Tasks as a Function of Training Group and Time of Assessment**

| | Group | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Adaptive | | Randomized | | Self-Selected | | Active Control | |
| Task | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest |
| *Training tasks (complex span)* | | | | | | | | |
| Numerical | 0.40 (0.14) | 0.59 (0.2) | 0.41 (0.18) | 0.62 (0.20) | 0.38 (0.16) | 0.58 (0.19) | 0.44 (0.17) | 0.48 (0.18) |
| Verbal | 0.81 (0.12) | 0.94 (0.08) | 0.79 (0.12) | 0.92 (0.10) | 0.82 (0.13) | 0.94 (0.08) | 0.81 (0.12) | 0.86 (0.12) |
| Figural | 0.51 (0.21) | 0.80 (0.18) | 0.53 (0.18) | 0.78 (0.17) | 0.60 (0.17) | 0.82 (0.13) | 0.63 (0.18) | 0.69 (0.15) |
| *Intermediate transfer* | | | | | | | | |
| Word-position binding | 2.45 (0.96) | 2.94 (1.00) | 2.45 (0.79) | 2.56 (1) | 2.30 (0.96) | 2.86 (1.03) | 2.37 (0.95) | 2.80 (0.84) |
| Brown-Peterson | 0.70 (0.16) | 0.80 (0.14) | 0.69 (0.17) | 0.75 (0.17) | 0.72 (0.16) | 0.77 (0.14) | 0.73 (0.15) | 0.77 (0.13) |
| Memory updating | 0.85 (0.13) | 0.90 (0.11) | 0.83 (0.12) | 0.89 (0.11) | 0.86 (0.08) | 0.91 (0.07) | 0.87 (0.11) | 0.90 (0.10) |
| *Far transfer* | | | | | | | | |
| RAPM | 7.44 (2.63) | 8.00 (2.74) | 8.00 (2.68) | 8.10 (2.70) | 8.38 (2.26) | 8.18 (2.29) | 8.31 (2.72) | 8.81 (2.21) |
| Letter sets | 20.71 (5.05) | 22.59 (4.45) | 21.77 (4.49) | 22.5 (3.69) | 20.53 (4.95) | 21.38 (4.99) | 22.31 (5.29) | 23.28 (4.16) |
| Locations test | 15.18 (5.37) | 16.91 (5.41) | 15.2 (4.46) | 17.8 (3.74) | 14.00 (5.09) | 15.71 (4.58) | 14.69 (4.37) | 17.94 (6.12) |
| Diagramming relationships | 22.91 (4.50) | 23.06 (4.02) | 21.8 (4.58) | 23.83 (3.72) | 22.44 (3.54) | 24.03 (2.70) | 23.34 (4.01) | 24.78 (4.01) |
| Nonsense syllogisms | 17.18 (4.41) | 19 (4.04) | 17.03 (4.54) | 18.33 (4.51) | 17.06 (5.03) | 17.59 (4.45) | 18.25 (4.38) | 19.88 (4.72) |
| *Control task* | | | | | | | | |
| Trivia Quiz | 0.59 (0.07) | 0.73 (0.08) | 0.60 (0.06) | 0.72 (0.09) | 0.61 (0.07) | 0.74 (0.06) | 0.59 (0.07) | 0.69 (0.07) |

***Note.*** **Standard deviations are given in parentheses. All values are given in proportional accuracy, except binding (d') and far transfer measures (number of correctly solved items).**

# TRAINING PROGRESS

## TRAINING PERFORMANCE

For each training task, we ran mixed ANOVAs using the level of difficulty achieved as dependent variable, and training session and group as independent variables. We coded training session as linear contrast to evaluate monotonic trends instead of potentially erratic fluctuations across sessions. As summarized in Table 3 and reflected by Figure 1, all groups but the one completing the randomized condition showed large effects of session for all training tasks (all *p*s ≤ .001). Furthermore, there was no difference in level of difficulty achieved between adaptive and self-selected training (linear contrasts of session x group interaction ($F(1, 58) = 0.38$, $p = .541$, $\eta_p^2 = .01$; $F(1, 58) = 1.01$, $p = .318$, $\eta_p^2 = .02$; $F(1, 58) = 0.87$, $p = .356$, $\eta_p^2 = .02$ for the numerical, verbal, and figural complex span, respectively). As a consequence of the study design, the average level of difficulty did not follow a monotonic trend across sessions (all *p*s ≥ .256) in the randomized WM training condition because the level of difficulty varied randomly across sessions and participants. The active control group also showed large linear training effects in all three versions of the trivia quiz (geography: $F(1, 30) = 1131.77$, $p < .001$, $\eta_p^2 = .97$; history: $F(1, 30) = 1142.45$, $p < .001$, $\eta_p^2 = .97$; natural science: $F(1, 30) = 1554.20$, $p < .001$, $\eta_p^2 = .98$).

Even though we defined the range of possible levels of task difficulty in the randomized training condition based on observations from a previous study (von Bastian & Oberauer, 2013), an ANOVA using task difficulty averaged across sessions and tasks as dependent variable and WM training condition as independent variable revealed a significant main effect of group, $F(2, 75) = 10.13$, $p < .001$, $\eta_p^2 = .18$. Bonferroni corrected post-hoc comparisons showed that on average, the randomized training group practiced on lower levels of task difficulty than both the adaptive ($M_{diff} = 1.81$, $p = .001$, $d = 1.16$) and the self-selected training group ($M_{diff} = 1.39$, $p = .010$, $d = 1.04$). As described in the Method section, each WM task ended when task duration exceeded 15 min. Therefore, because participants in the randomized training condition practiced with trials of shorter list lengths, they completed slightly more trials than the other two WM training groups (adaptive: $M = 11.37$, randomized: $M = 11.90$, self-selected: $M = 11.22$). An ANOVA using the number of trials averaged across sessions and tasks as dependent variable and WM training condition as independent variable yielded a significant effect of group, $F(2, 84) = 3.14$, $p = .049$, $\eta_p^2 = .07$. Bonferroni corrected post-hoc between-groups comparisons showed a trend for more completed trials in the randomized than in the self-selected training group ($p = .056$). None of the other comparisons was significant (*p*s > .193).

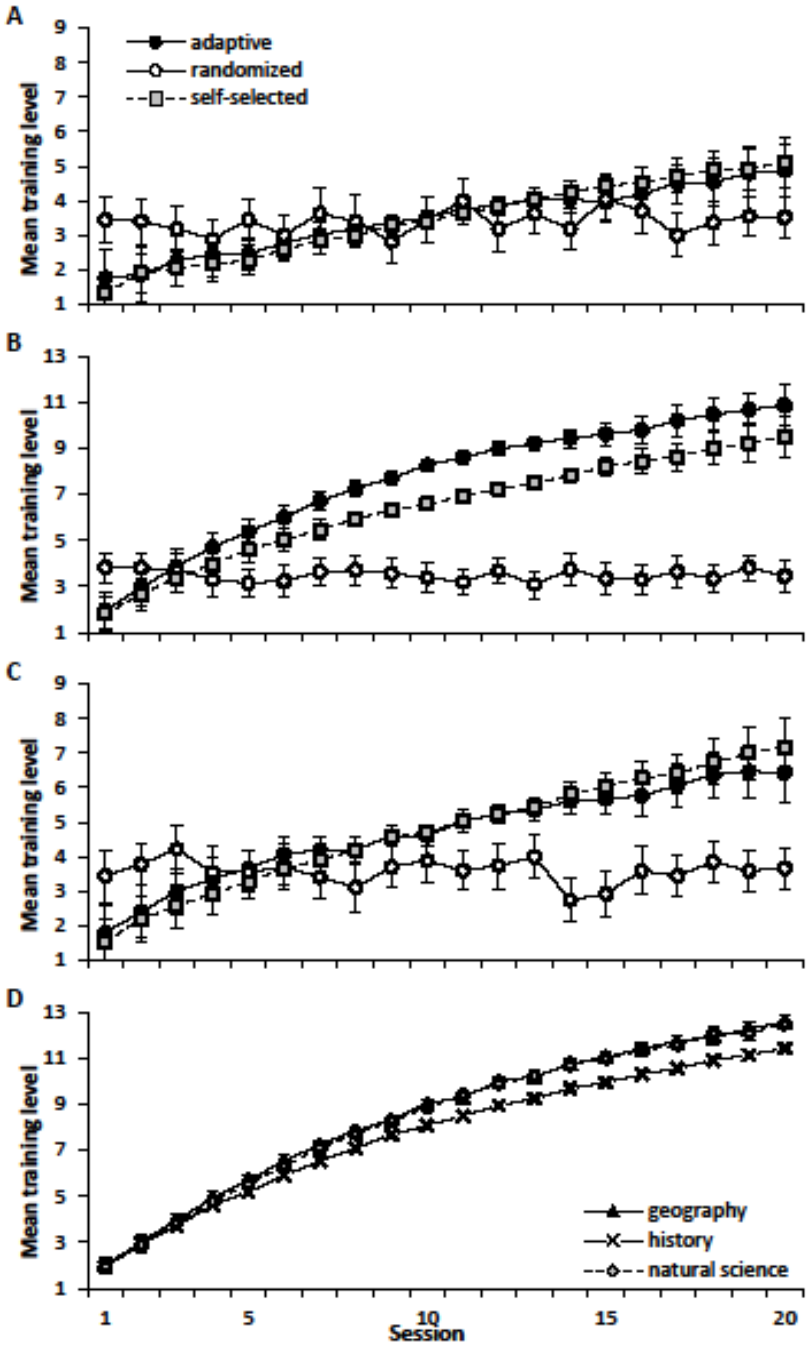


*Figure 1.* **Change in performance during the training phase: WM training progress in (A) numerical complex span, (B) verbal complex span, (C) figural complex span, and in (D) active control training (trivia questions). Note the varying scaling of the dependent variable. Error bars represent confidence intervals (95%) for within-subjects comparisons, calculated according to Cousineau (2005) and Morey (2008).**

**Table 3. Linear Contrasts of Training Effects on Performance in the Trained Tasks during Working Memory Training**

| | Group | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Adaptive | | | | Randomized | | | | Self-Selected | | | |
| Training task (complex span) | *M* (*SD*) | *F* (1, 29) | *p* | $\eta_p^2$ | *M* (*SD*) | *F* (1, 26) | *p* | $\eta_p^2$ | *M* (*SD*) | *F* (1, 29) | *p* | $\eta_p^2$ |
| Numerical | 4.87 (4.42) | 14.04 | **.001** | .33 | 3.52 (1.70) | 1.35 | .256 | .05 | 5.10 (3.55) | 31.71 | < **.001** | .52 |
| Verbal | 10.87 (4.34) | 95.39 | < **.001** | .77 | 3.44 (1.85) | 0.23 | .638 | .01 | 9.50 (4.48) | 74.40 | < **.001** | .72 |
| Figural | 6.43 (4.47) | 27.03 | **< .001** | .48 | 3.67 (1.62) | 0.52 | .477 | .02 | 7.17 (4.31) | 45.77 | **< .001** | .61 |

***Note.*** **Bold *p*-values indicate significant effects. Only participants with complete training data sets were included in the analyses. The dependent variable was the level of difficulty achieved in each training session. Means and standard deviations are given for the last training session.**

## MOTIVATION DURING TRAINING

To determine whether the three training algorithms had differential effects on motivation during training, we ran a set of mixed ANOVAs using the three one-item training motivation measures (enjoyment, effort, and perceived fit between task difficulty and ability) that participants completed after each session as dependent variables, and group (3) and session (20) as independent variables. There was no main effect of group on two of the three motivation measures, showing that the three experimental training groups did not differ in their overall enjoyment experienced during training ($F(2, 84) = 0.44$, $p = .643$, $\eta_p^2 = .01$) or overall effort spent on training, $F(2, 84) = 0.21$, $p = .810$, $\eta_p^2 = .01$. However, there was a marginal effect of group on the overall perceived fit between task difficulty and ability, $F(2, 84) = 2.93$, $p = .059$, $\eta_p^2 = .07$. Randomized training yielded a smaller perceived fit than the other two training conditions, which reached significance for the comparison to self-selected training ($M_{Diff} = .53$, $p = .019$), but not adaptive training, $M_{Diff} = .34$, $p = .129$. There was no difference between adaptive and self-selected training, $M_{Diff} = -.19$, $p = .374$.

For enjoyment, neither the linear ($F(1, 84) = 0.96$, $p = .335$, $\eta_p^2 = .01$) nor the quadratic trend for session ($F(1, 84) = 1.20$, $p = .277$, $\eta_p^2 = .01$) were significant. However, the interaction between group and the quadratic trend of session ($F(2, 84) = 3.96$, $p = .023$, $\eta_p^2 = .09$) was significant, indicating that participants' enjoyment in the randomized condition decreased after the first session and increased again in the last sessions, whereas enjoyment ratings in the other groups did not vary much across sessions. For effort, the data followed a quadratic trend for session ($F(1, 84) = 22.27$, $p < .001$, $\eta_p^2 = .21$) with higher effort ratings in the beginning and the end of the training phase than in-between. This trend was not modulated by group, $F(2, 84) = 1.31$, $p = .276$, $\eta_p^2 = .03$. For the rating of perceived fit between task difficulty and ability, neither the linear ($F(1, 84) = 2.59$, $p = .111$, $\eta_p^2 = .03$) nor the quadratic trend ($F(1, 84) < 0.01$, $p = .995$, $\eta_p^2 < .01$) were significant. Furthermore, we observed no significant group x session interactions ($F(2, 84) = 1.35$, $p = .265$, $\eta_p^2 = .03$ and $F(2, 84) = 0.44$, $p = .645$, $\eta_p^2 = .01$ for the linear and the quadratic trend, respectively).

In addition to the one-item motivation measures, we administered the QCM after the first and the tenth session. For the QCM, there was no main effect of group ($F(2, 92) = 0.39$, $p = .676$ , $\eta_p^2 = .01$), but a large main effect of session ($F(1, 92) = 24.49$, $p < .001$, $\eta_p^2 = .21$), with motivation decreasing from session 1 to session 10. However, the effect was not modulated by group, $F(2, 92) = 0.52$, $p = .596$, $\eta_p^2 = .01$.

## TRAINING AND TRANSFER GAINS

To evaluate gain from pre to post assessment, we computed standardized gain scores (i.e., difference between posttest and pretest score divided by the pretest standard deviation) for each individual and each task (cf. von Bastian & Oberauer, 2013). We then ran linear mixed-effects (LME) models to estimate these gain scores on the level of generalization range (i.e., training, intermediate transfer and far transfer effects) rather than on the level of single tasks (for a more detailed discussion of the advantages of using LME models over analyses of variance, see Baayen, Davidson, & Bates, 2008; Bates, 2010; see also von Bastian & Oberauer, 2013). We ran a separate LME models on the gain scores for each range of generalization. LME models can simultaneously account for multiple sources of variances, which can be either fixed effects or random effects. The fixed-effects predictor was group (adaptive, randomized, self-selected, and active control). The four levels of group were coded as three contrasts according to our research questions (adaptive vs. active control, adaptive vs. randomized, and adaptive vs. self-selected training), entered as sum contrasts (i.e., -1 vs. 1) with the intercept reflecting the grand mean of the gain scores.

We included two crossed random effects (Baayen, et al., 2008) in the models: the random effect of subject to account for random variability between participants, and the random effect of task to account for the fact that the paradigms we used in our study to assess WM and reasoning reflect only a sample of possible tasks that could be administered to measure these theoretical constructs (cf. von Bastian & Oberauer, 2013). Random effects can be assumed for intercepts (i.e., random variation around the overall mean of the dependent variable) and for slopes (i.e., additional random variation in the size of effects of all predictors). The results of a recent simulation study demonstrated that models with design-driven maximal random effects structure generalize best (Barr, Levy, Scheepers, & Tily, 2013). Given that each subject belonged to one group only, we included the random effect of subject for the intercept only, while we introduced the random effect of task for both intercept and slope. In one case (intermediate transfer to WM), the model with this random-effects structure did not converge. Following the recommendations by Barr et al. (2013), we chose to remove the random intercept of task for this model, leaving a random effect of task on the slope, and a random effect of subject on the intercept.

Model fitting was carried out using the statistics program R (R Core Team, 2014) with the package “lme4” (Bates, Maechler, Bolker, & Walker, 2014). Kenward-Roger approximation with the package “pbkrtest” (Halekoh & Højsgaard, 2014) was used to compute the degrees of freedom to derive information about the significance of the predictors. Results of the LME models are summarized in Tables 4 (fixed effects) and 5 (random effects). All reported *p*-values are two-tailed.

**Table 4. Parameter Estimates for Fixed Effects of the Linear Mixed-Effects Models Relating Effects of Training Algorithm to Training and Transfer Gains**

| Transfer range / Parameter | Estimate | *SE* | *t* | *p* |
|---|---|---|---|---|
| Training effects (complex span) | | | | |
| Intercept (grand mean) | 0.98 | 0.07 | 14.86 | **< .001** |
| Adaptive vs. active control | 0.97 | 0.19 | 5.13 | **.002** |
| Adaptive vs. randomized | -0.07 | 0.17 | -0.41 | .698 |
| Adaptive vs. self-selected | -0.17 | 0.16 | -1.03 | .343 |
| Intermediate transfer (working memory) | | | | |
| Intercept (grand mean) | 0.43 | 0.04 | 10.39 | **< .001** |
| Adaptive vs. active control | 0.20 | 0.12 | 1.62 | .174 |
| Adaptive vs. randomized | -0.18 | 0.15 | -1.20 | .292 |
| Adaptive vs. self-selected | -0.05 | 0.11 | -0.42 | .697 |
| Far transfer (reasoning) | | | | |
| Intercept (grand mean) | 0.28 | 0.06 | 4.92 | **.002** |
| Adaptive vs. active control | -0.07 | 0.11 | -0.70 | .507 |
| Adaptive vs. randomized | 0.02 | 0.12 | 0.18 | .866 |
| Adaptive vs. self-selected | -0.09 | 0.11 | -0.79 | .461 |

***Note.*** **Bold *p*-values indicate significant predictors ($p < .05$).**

**Table 5. Estimates for Random Effects of the Linear Mixed-Effects Models Relating Effects of Training Algorithm to Training and Transfer Gains**

| | *SD* | | |
|---|---|---|---|
| Random Effect | Training | Intermediate Transfer | Far Transfer |
| Subject | | | |
| Intercept | 0.46 | 0.25 | 0.04 |
| Task | | | |
| Intercept | 0.06 | - | 0.11 |
| Adaptive vs. active control | 0.18 | 0.07 | 0.16 |
| Adaptive vs. randomized | 0.09 | 0.16 | 0.21 |
| Adaptive vs. self-selected | 0.09 | 0.06 | 0.19 |
| Residual | 0.76 | 0.69 | 0.71 |

## TRAINING GAINS

The significant intercept ($b = 0.98$, $p < .001$) indicates that performance in the trained tasks generally increased from pre to posttest. The first contrast (active control vs. adaptive training) being significant shows that the adaptive training group's improvement in the trained tasks is larger than the one observed for the active control group ($b = 0.97$, $p = .002$). Hence, it can be concluded that there was a WM training effect that went beyond simple retest or non-specific intervention effects. The non-significant comparisons between adaptive and non-standard WM training procedures indicate that training gains were similar for all three training algorithms (see also Figure 2A).

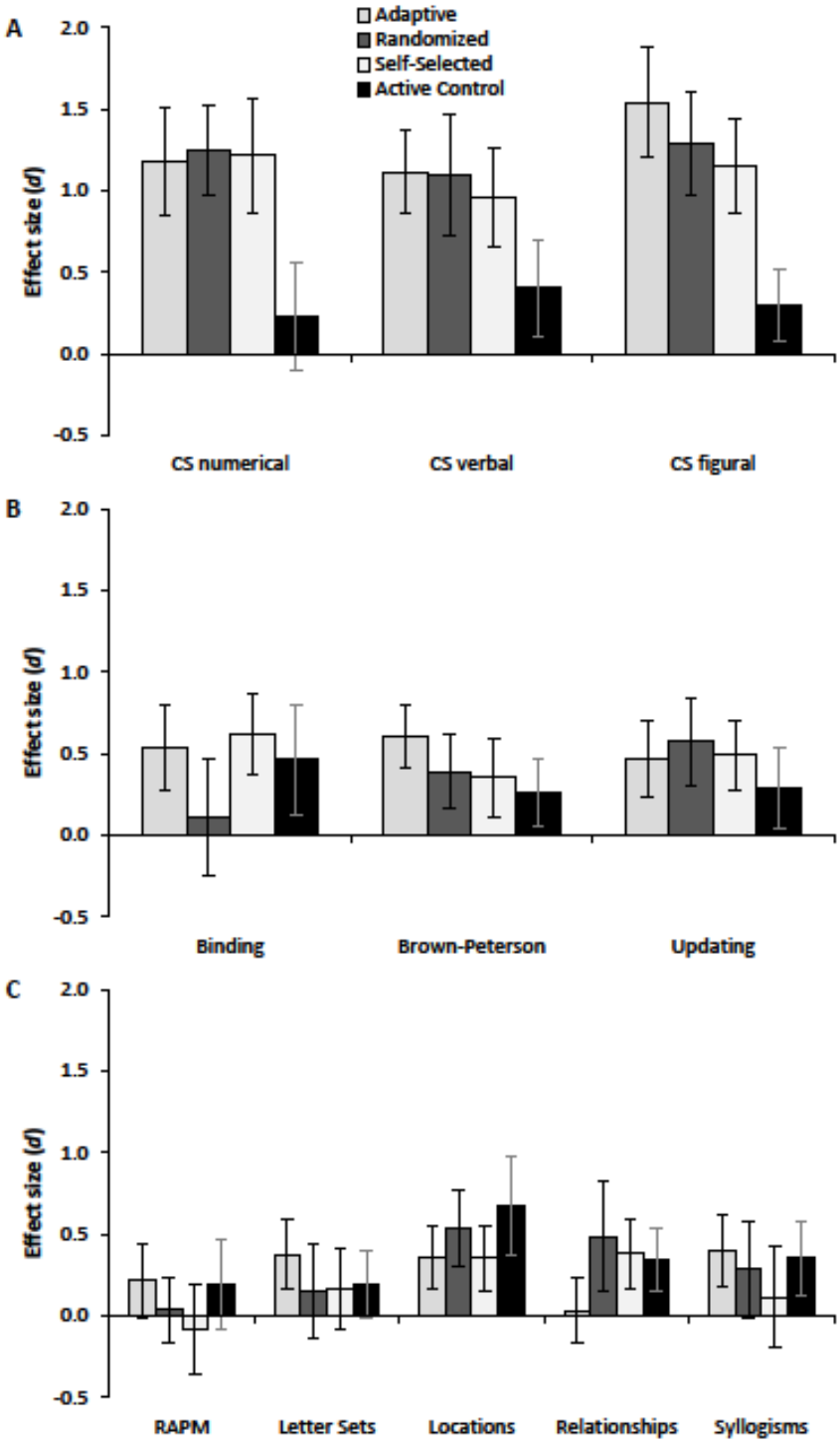


***Figure 2.*** **Gain scores in (A) the trained WM tasks, (B) structurally different WM tasks (intermediate transfer), and (C) reasoning tasks (far transfer). Error bars represent confidence intervals (95%).**

### TRANSFER GAINS

Performance in the intermediate (WM) and far (reasoning) transfer tasks generally increased from pre to posttest, indicated by the significant intercepts ($b$ = 0.43, $p$ < .001 and $b$ = 0.28, $p$ = .002). Figure 2B illustrates that there was a tendency of adaptive training yielding larger intermediate (WM) transfer gains than active control training, which, however, was not significant ($b$ = 0.20, $p$ = .174). There was also no significant difference between these two groups in reasoning gain scores ($b$ = -.07, $p$ = .507, see also Figure 2C). None of the contrasts examining differences between WM training algorithms were significant; hence, the type of WM training procedure did not modulate intermediate and far transfer effects.

In summary, the results showed that adaptive WM training led to larger gains in the trained tasks than active control training. However, there was no consistent evidence for transfer to structurally dissimilar WM tasks or to reasoning tasks. Furthermore, we observed no differences between adaptive and non-standard (i.e., randomized or self-selected) WM training procedures for neither training nor transfer gains.

### CONTROL TASK

Improvement in the open format trivia quiz for the active control group was tested against the conjoined experimental groups. As expected, the time (pretest vs. posttest) and group (experimental vs. active control) interaction was not significant, $F(1, 128) = 1.81$, $p = .181$, $\eta_p^2 = .01$.

## DISCUSSION

In this study, we tested the hypothesis that in order to be most effective, WM training should provide a task difficulty that continuously exceeds an individual's routine cognitive demands, and, thus, has to be adaptive (Lövden, et al., 2010). Previous evidence in favor of this hypothesis (Klingberg, 2010) was gained from studies comparing adaptive to low-level WM training (in which individuals constantly practice with low task difficulty), a design which confounds adapting difficulty to individual performance with variation in task difficulty. However, exposure to continuously varying task difficulty also requires the cognitive system to adjust its functional supplies to changing environmental demands, thereby potentially inducing cognitive plasticity.

To differentiate between these two factors, we compared adaptive to randomized instead of to low-level WM training. The main finding of our study is that we observed no differences between training procedures in terms of training and transfer effects. Thus, our results indicate that training with varying task difficulty is similarly effective as individually adaptive training. The fact that participants in the randomized training condition practiced on average on overall easier levels of task difficulty than those in the adaptive condition even indicates that training gains may indeed be driven by variability in rather than by continuous adaptation of task difficulty to individual performance.

Furthermore, there was no difference in training progress between adaptive and self-selected training, showing that the adaptive procedure applied in our study (i.e., a threshold of 80% correct before progressing to the next higher level of difficulty, and a threshold of 60% correct for moving to the next lower level of difficulty) matches what individuals themselves would define as an optimal modification of training task difficulty. This is in line with a recent study by Gibson and colleagues (2013) demonstrating that an adaptive algorithm operating in this range is more effective than one pushing for higher WM performance (i.e., requiring perfect performance for reaching the threshold).

To investigate whether adaptive training is superior to other training procedures in terms of motivation (and with it, trainees' compliance), we measured training enjoyment, effort, and perceived fit between task difficulty and cognitive ability (after each session) and current overall training motivation (at the beginning and halfway through the training period). Given that task difficulty was independent of individual performance in the randomized condition, it can be expected that the perceived fit is lower in this group compared to the two other WM training groups, which should not differ. This was precisely the case. Importantly, however, this lower perceived fit had no negative impact on the other motivational measures (enjoyment, effort, and overall training motivation). The only exception was that participants in the randomized training condition rated their enjoyment higher after the first and the last training session than participants in the other training groups (i.e., their ratings followed a U-shaped function, whereas the ratings of participants in the other groups remained roughly the same across sessions). It is unclear why the randomized training procedure was regarded more enjoyable in the first training session than the other two WM training procedures, as all groups started on the same level of task difficulty in that first training session. In sum, apart from the first and last training session, all three WM training procedures were perceived as similarly enjoyable and challenging and thus could be applied similarly well in practice. Moreover, these findings suggest that training and transfer effects cannot be attributed to differences in training motivation or effort between training groups alone.

Finally, we evaluated whether we could replicate previous findings showing transfer effects to untrained WM tasks and reasoning after a similar adaptive complex span training (von Bastian & Oberauer, 2013). Despite the large training effects we observed in the present study, we found, however, no evidence for transfer effects. There are three major methodological differences between the present and our previous study that could potentially contribute to the diverging results: (1) our modifications to the adaptive algorithm, (2) a different activity in the active control condition, and (3) the lack of a follow-up assessment.

First, we modified the adaptive algorithm in several aspects due to design requirements. To keep comparability of single sessions between training conditions as high as possible, task difficulty was modified only once per session (i.e., after 100% of the trials per session). In contrast, in the previous study, task difficulty was adjusted within sessions, a procedure more typical in the training literature (e.g., Chein & Morrison, 2010; Dunning, Holmes, & Gathercole, 2013; Jaeggi, et al., 2008; von Bastian, Langer, et al., 2013). Furthermore, in the present study, task difficulty could increase or decrease, whereas it was only increased in the earlier study. Hence, it is possible that participants in the previous study could have reached higher levels of difficulty that were more challenging and thus induced larger magnitudes of transfer. To test this possibility, we ran ANOVAs comparing the span levels reached in the two studies across the 20 training sessions for the figural and the numerical complex span. We refrained from doing so for the verbal complex span as stimuli were letters in the present study and words in the earlier study. There were neither significant group effects (both $F$s < 1) nor significant linear trends for the group x session interactions (figural: $F(1, 54) = 1.46$, $p = .232$, $\eta_p^2 = .03$; numerical: $F(1, 54) = 1.64$, $p = .206$, $\eta_p^2 = .03$). Thus, the levels of difficulty achieved were about the same across the two studies, suggesting that our modifications to the adaptive algorithm (i.e., spacing and direction of difficulty adjustment) did not affect training progress and are therefore an unlikely explanation for the absence of transfer. However, further studies are needed to clarify how such modifications affect training and transfer gains.

The second difference between the previous and the present study concerns our choice of control intervention (perceptual matching and trivia quizzes, respectively). As we observed large improvements in processing speed after perceptual matching training in the earlier study (which strongly contributes to WM performance, see Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007; cf. von Bastian & Oberauer, 2013), we chose to use trivia

quizzes instead. Theoretically, such questions on general knowledge should demand only little WM and draw mainly on crystallized intelligence. Still, as we discussed in a recent review (von Bastian & Oberauer, 2014), there are two potential drawbacks of using trivia quizzes as a control condition. First, trivia quizzes could be more fun to do than complex span tasks. However, there were no differences in enjoyment ratings during training between the adaptive WM and the control training group (linear trend $F < 0$). Second, trivia quiz questions could evoke reasoning strategies (e.g., rejection of implausible answers) that would require – and hence, practice – relational integration processes; that is, the coordination of information elements into structures. Recent theories consider relational integration as crucial part of WM (e.g., Oberauer, 2010; Oberauer, Süß, Wilhelm, & Wittmann, 2003), and research has shown that such processes are highly related to fluid intelligence (e.g., Oberauer, Süß, Wilhelm, & Wittmann, 2008). We can only speculate whether such processes took place during active control training, but it could serve as an explanation for the active control group also showing some improvement in the transfer tasks. Arguably, however, WM demands can still be expected to be higher for complex span tasks than trivia quizzes. Furthermore, previous training studies using trivia quizzes as control activity were in fact successful in detecting transfer (e.g., Jaeggi, et al., 2014). Therefore, even though we cannot exclude that the active control group's improvements obscure transfer effects of WM training, we believe it is unlikely that they fully explain the lack thereof.

The third methodological deviation concerns the assessment of transfer effects. In our previous study (von Bastian & Oberauer, 2013), participants were tested twice for transfer: once immediately after training and once six months later. As we found no significant decrease in performance from post to follow-up assessment, we were able to evaluate transfer effects taking both points in time together, yielding larger statistical power to detect potential transfer effects. We cannot exclude that the addition of a follow-up assessment to our study would have resulted in observable transfer gains. The duration of the testing sessions is another feature of transfer assessment that has been recently discussed as one potential explanation for the inconsistencies observed in the training literature. Green, Strobach, and Schubert (2014) argue that long testing sessions could foster unwanted effects of fatigue, resource-depletion, or practice, thereby making it difficult to detect transfer. Even though our testing sessions were indeed relatively long (3 h), the fact that we used two different orders of test administration should control for such effects. In addition, the testing sessions in the previous study were considerably longer (4.5 h), making testing session duration an unlikely explanation for the absence of transfer.

## CONCLUSION

The absence of transfer effects in this study questions the potential of adaptive complex span WM training to induce transfer effects in general and change in reasoning ability in particular, given that other studies using similar training paradigms did not detect far transfer to reasoning either (Chein & Morrison, 2010; Colom et al., 2010; Harrison, et al., 2013; Licini, 2014). Notwithstanding the absence of transfer effects, our findings contradict the assumption that WM training has to be adaptive to individual performance in order to yield training-induced gains in cognitive performance, as the experimental training manipulation had neither an effect on practiced (for which effects would be expected to be strongest) nor on untrained WM and reasoning tasks (intermediate and far transfer). Rather, the present data set suggests that exposing participants to varying levels of difficulty is sufficient for challenging the flexibility of the cognitive system by exceeding routine demands (cf. Lövden, et al., 2010) and thereby inducing performance improvements.

## AUTHOR NOTE

This research is part of the first author's doctoral thesis and was supported by grants to the first author from the Forschungskredit of the University of Zurich and the Suzanne and Hans Biäsch Foundation for Applied Psychology. We thank André Locher for programming the training tasks in Tatool, and Sabrina Guye, Veronica Heusser, Melanie Künzli, Eszter Montvai, Katharina Vogt, Helen Wirz, and Marc Züst for their assistance with collecting the data.

## REFERENCES


Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven advanced progressive matrices test. *Educational and Psychological Measurement, 54*(2), 394-403.

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven advanced progressive matrices test. *Journal of Psychoeducational Assessment, 17*, 354-361.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (in press). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-014-0699-x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 290-412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7).

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Retrieved from http://lme4.r-forge.r-project.org/book/

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben : Need for cognition: a scale measuring engagement and happiness in cognitive tasks. *Zeitschrift für Sozialpsychologie, 25*, 147-154.

Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI). Manual* (2nd ed.). Göttingen: Hogrefe.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49-59.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in Human Neuroscience, 6*(63), 1-7. doi: 10.3389/fnhum.2012.00063

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*, 12-21. doi: 10.1080/17470215808416249

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131. doi: 10.1037/0022-3514.42.1.116

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17*(2), 193-199. doi: 10.3758/PBR.17.2.193

Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence, 40*, 531-542. doi: 10.1016/j.intell.2012.07.004

Colom, R., Quiroga, M. Á., Shih, P. C., Martínez-Molina, A., Román, F. J., Requena, L., & Ramírez, I. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence, 38*, 497-505. doi: 10.1016/j.intell.2010.06.008

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., . . . Jaeggi, S. M. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence, 41*, 712-727. doi: 10.1016/j.intell.2013.09.002

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 739-786. doi: 10.3758/BF03196772

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory (NEO-FFI). Professional manual*. Odessa, FL: Psychological Assessment Resources.

Cousineau, D. (2005). Confidence intervals in within-subjects designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42-45.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19*, 450-466.

Deci, E. L., & Ryan, R. M. (n.d.). Intrinsic Motivation Inventory Retrieved 06/13, 2013, from http://selfdeterminationtheory.org/questionnaires/10-questionnaires/50

Dunning, D. L., Holmes, J., & Gathercole, S. E. (2013). Does working memory training lead to generalized improvements in children with low working memory? A randomized controlled trial. *Developmental Science, 16*(6), 915-925. doi: 10.1111/desc.12068

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

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.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309-331. doi: 10.1037/0096-3445.128.3.309

Gibson, B. S., Gondoli, D. M., Kronenberger, W. G., Johnson, A. C., Steeger, C. M., & Morrisey, R. A. (2013). Exploration of an adaptive training regimen that can target the secondary memory component of working memory capacity. *Memory & Cognition, 41*(5), 726-737. doi: 10.3758/s13421-013-0295-8

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

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models - the R package pbkrtest. *Journal of Statistical Software, 59*(9), 1-30.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24*(12), 2409-2419. doi: 10.1177/0956797613492984

Jaeggi, S. M., Buschkuehl, 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. doi: 10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition, 42*(3), 464-480. doi: 10.3758/s13421-013-0364-z

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W., J. (2010). The relationship between n-back performance and matrix reasoning - implications for training and transfer. *Intelligence, 38*(6), 625-635. doi: 10.1016/j.intell.2010.09.001

Karbach, J., Strobach, T., & Schubert, T. (in press). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology*. doi: 10.1080/09297049.2014.899336

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. doi: 10.1177/0956797614548725

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*, 317-324. doi: 10.1016/j.tics.2010.05.002

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., . . . Westerberg, H. (2005). Computerized training of working memory in children with ADHD - a randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(2), 177-186. doi: 10.1097/00004583-200502000-00010

Klumb, P. L. (2001). Knoten im Taschentuch: Der Einsatz von Gedächtnishilfen im Alltag. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 33*(1), 42-49. doi: 10.1026//0049-8637.33.1.42

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

Licini, C. (2014). *Verbesserung der Lernfähigkeit durch gezieltes Arbeitsgedächtnistraining [Improvement of the Ability Learn Through Working Memory Training].* (Unpublished master's thesis). University of Zurich, Zurich, Switzerland.

Lövden, 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. doi: 10.1037/a0020080

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

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61-64.

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

Noack, H., Lövden, M., & Schmiedek, F. (2014). On the validity and generality of transfer effects in cognitive training research. *Psychological Research, 78*(6), 773-789. doi: 10.1007/s00426-014-0564-6

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General, 134*(3), 368-387. doi: 10.1037/0096-3445.134.3.368

Oberauer, K. (2006). Is the focus of attention in working memory expanded through practice? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 197-214. doi: 10.1037/0278-7393.32.2.197

Oberauer, K. (2010). Declarative and procedural working memory: Common principles, common capacity limits? *Psychologica Belgica, 50*(3&4), 277-308.

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

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence, 36*, 641-652. doi: 10.1016/j.intell.2008.01.007

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raven, J. C. (1990). *Advanced Progressive Matrices: Sets I, II*. Oxford, U.K.: Oxford Psychologists Press.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., . . . 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. doi: 10.1037/a0029082

Rheinberg, F., Vollmeyer, R., & Bruns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen. *Diagnostica, 47*(2), 57-66.

Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Frontiers in Human Neuroscience, 6*(166). doi: 10.3389/fnhum.2012.00166

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

Schmiedek, F., Lövden, 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*(27), 1-10. doi: 10.3389/fnagi.2010.00027

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414-429. doi: 10.1037/0096-3445.136.3.414

Schweizer, S., Hampshire, A., & Dalgleish, T. (2011). Extending brain-training to the affective domain: Increasing cognitive and affective executive control through emotional working memory training. *PLoS One, 6*(9), e24372. doi: 10.1371/journal.pone.0024372

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

Smith, G., Del Sala, S., Logie, R. H., & Maylor, E. A. (2000). Prospective and retrospective memory in normal ageing and dementia: A questionnaire study. *Memory, 8*(5), 311-321. doi: 10.1080/09658210050117735

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., . . . Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41*, 638-663. doi: 10.1016/j.intell.2013.07.013

Stepankova, H., Lukavsky, J., Buschkuehl, M., Kopecek, M., Ripova, D., & Jaeggi, S. M. (2014). The malleability of working memory and visuospatial skills: A randomized controlled study in older adults. *Developmental Psychology, 50*(4), 1049-10559. doi: 10.1037/a0034913

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*, 261-288.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., . . . Gabrieli, J. D. E. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS One, 8*(5), e63614. doi: 10.1371/journal.pone.0063614

von Bastian, C. C., Langer, N., Jäncke, L., & Oberauer, K. (2013). Effects of working memory training in young and old adults. *Memory & Cognition, 41*(4), 611-624. doi: 10.3758/s13421-012-0280-7

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods, 45*(1), 108-115. doi: 10.3758/s13428-012-0224-y

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language, 69*, 36-58. doi: 10.1016/j.jml.2013.02.002

von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: A review. *Psychological Research, 78*(6), 803-820. doi: 10.1007/s00426-013-0524-6

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*(433), 1-22. doi: 10.3389/fpsyg.2013.00433