Journal of Cognitive Enhancement
https://doi.org/10.1007/s41465-021-00235-2

ORIGINAL RESEARCH

The Pursuit of Effective Working Memory Training: a Pre-registered Randomised Controlled Trial with a Novel Varied Training Protocol

Liisa Ritakallio[1] · Daniel Fellman[1,2,3] · Jussi Jylkkä[1] · Otto Waris[1,4,5] · Nelly Lönnroth[1] · Reidar Nervander[1] · Juha Salmi[6,7,8] · Matti Laine[1,9]

Received: 26 April 2021 / Accepted: 11 November 2021

Abstract

Working memory (WM) training, typically entailing repetitive practice with one or two tasks, has mostly yielded only limited task-specific transfer effects. We developed and tested a new WM training approach where the task paradigm, stimulus type, and predictability of the stimulus sequence were constantly altered during the 4-week training period. We expected that this varied training protocol would generate more extensive transfer by facilitating the use of more general strategies that could be applied to a range of WM tasks. Pre-post transfer effects following varied training (VT group, $n = 60$) were compared against traditional training (TT group, training a single adaptive WM task, $n = 63$), and active controls (AC, $n = 65$). As expected, TT evidenced strong task-specific near transfer as compared to AC. In turn, VT exhibited task-specific near transfer only on one of the measures, and only as compared to the TT group. Critically, no evidence for task-general near transfer or far transfer effects was observed. In sum, the present form of VT failed to demonstrate broader transfer. Nevertheless, as VT has met with success in other cognitive domains, future studies should probe if and how it would be possible to design WM training protocols that promote structural learning where common features of specific tasks would be identified and utilised when selecting strategies for novel memory tasks.

Keywords Working memory · Memory training · Cognitive training · Varied training · Skill acquisition · Structural learning

✉ Liisa Ritakallio
liisa.ritakallio@abo.fi

1 Department of Psychology, Åbo Akademi University, Turku, Finland

2 Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden

3 Department of Applied Educational Science, Umeå University, Umeå, Sweden

4 Department of Child Psychiatry, University of Turku and Turku University Hospital, Turku, Finland

5 INVEST Research Flagship Center, University of Turku, Turku, Finland

6 Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland

7 Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland

8 Turku Institute for Advanced Studies, University of Turku, Turku, Finland

9 Turku Brain and Mind Center, Turku, Finland

Introduction

The importance of working memory (WM) for human cognition has been the driving force behind the active WM training research over the last decades (Klingberg et al., 2002; Morrison & Chein, 2011). WM training regimes have included repetitive computerised adaptive practise with the same task(s), and the pioneering studies showed surprisingly large and broad training improvements in cognitive functioning. However, many of these initial studies suffered from methodological shortcomings (Green et al., 2019; Melby-Lervåg et al., 2016; Morrison & Chein, 2011; Shipstead et al., 2012), and the most recent meta-analyses indicate that this type of training elicits more substantial performance improvement only on the trained tasks and their untrained variants (Gathercole et al., 2019; Holmes et al., 2019; Kassai et al., 2019; Melby-Lervåg et al., 2016; Norris et al., 2019; Sala & Gobet, 2017; Soveri et al., 2017), instead of more general enhancement of cognition. This is against the original assumption that this type of training increases WM

Published online: 22 December 2021

capacity and thereby improves cognitive functions relying on WM. Thus, the WM training effects, mainly seen only on the untrained variants of the trained task(s), appear to depend on other mechanisms.

Some researchers have proposed an alternative explanation for training-related improvements in WM that is based on cognitive skill learning, which fits well to the limited WM training effects obtained thus far (e.g. Fellman et al., 2020a; Gathercole et al., 2019; Laine et al., 2018). According to this view, repeated practice with a new and demanding WM task triggers the spontaneous development of a cognitive skill for performing that particular task. Recent evidence on the importance of task-specific strategies and their evolvement during repeated WM practice (Fellman et al., 2020a; Forsberg et al., 2020; Laine et al., 2018; Malinovitch et al., 2020; Waris et al., 2021a, b) supports the skill learning view, as the selection and application of a suitable strategy represent central components of learning a new cognitive skill (Chein & Schneider, 2012). As repeated practice with a limited set of WM tasks only develops skills to perform those specific tasks and their very closely related untrained variants, the challenge is to try to create training protocols that would be less susceptible to this "curse of specificity" that is characteristic of skill learning.

One possible avenue for broader transfer in WM training is to introduce variability in the training protocol. In contrast to constant training with the same task(s) that would lead to the emergence of a task-specific skill, varied training can help in identifying lawful relationships in a range of task variants (Schmidt & Bjork, 1992) that can yield more extensive performance improvements. In perceptual-motor learning, varied training has consistently been shown to facilitate transfer (e.g. Braun et al., 2009), and similar findings have also been reported for cognitive skill acquisition such as task-switching (Karbach & Kray, 2009; Sabah et al., 2019), problem-solving (Vakil & Heled, 2016), mental calculations (Sanders et al., 2002), and oral reading fluency (Reed et al., 2019).

There are other WM training studies that have employed more varied training protocols, but these studies have not included a comparison to a WM training group with a single, non-varied training task (e.g. Chein & Morrison, 2010; Richey et al., 2014). One exception is the study by Redick et al. (2020) who examined the role of proactive interference on WM training and transfer. They employed two WM training groups practising an operation span task with either varied (letters, words, digits) or non-varied content (letters), and an active control group. Only the training group with non-varied content showed some limited transfer to other serial short-term memory tasks with letters. The authors attributed these transfer results to the development of stimulus- or task-specific strategies. However, as Redick et al. (2020) point out, their verbal WM transfer tasks did not include other stimulus materials than letters, which might have masked possible task-general near transfer effects. Moreover, only stimuli and not task paradigms were varied, albeit optimal performance in different WM task paradigms can call for different strategies.

In the present study, we developed and tested a novel training protocol with varied WM tasks (varied training; VT), and contrasted it with traditional training (TT) that employed repetitive practice with a single adaptive WM task. Our VT protocol included several elements aimed to bolster structural learning of WM tasks, i.e. the development of more general rules (strategies) to solve these tasks. More specifically, the protocol called for rapid switching between a range of WM tasks that varied in terms of paradigm, stimulus materials, and predictability of stimulus sequence (fully or partly random). We expected that the VT protocol would promote the generation of a more abstract representation of the WM task space where, despite surface differences, similar strategies could be employed. Thus, the aim of this study was to test whether our novel VT paradigm elicits wider transfer effects within the memory domain, as compared to TT and active controls (AC). We hypothesised that, as compared to the AC and the TT groups, the VT group would show transfer also to other memory tasks than the untrained variants of the training tasks. In contrast, the TT group would show only task-specific near transfer compared to the AC group. These hypotheses as well as the study protocol were pre-registered in the Open Science Framework (see https://osf.io/c9ygt).

Materials and Methods

Participants

The study was conducted in accordance with the Helsinki Declaration and approved by the Ethics Committee of the Departments of Psychology and Logopedics, Åbo Akademi University. The participants were English-speaking adults in the age range of 18 to 50, recruited via the crowdwork website Prolific (http://www.prolific.co). Informed consent was obtained from all participants before enrolment. The participants were also informed of the contents of the study, the voluntary nature of participation, and the option to discontinue at any time without giving a reason. The participants were offered financial compensation for their participation (£50.52).

Procedure

All data collection was done online on the participants' home computer or similar. The study was run on our in-house online platform SOILE, which is developed for

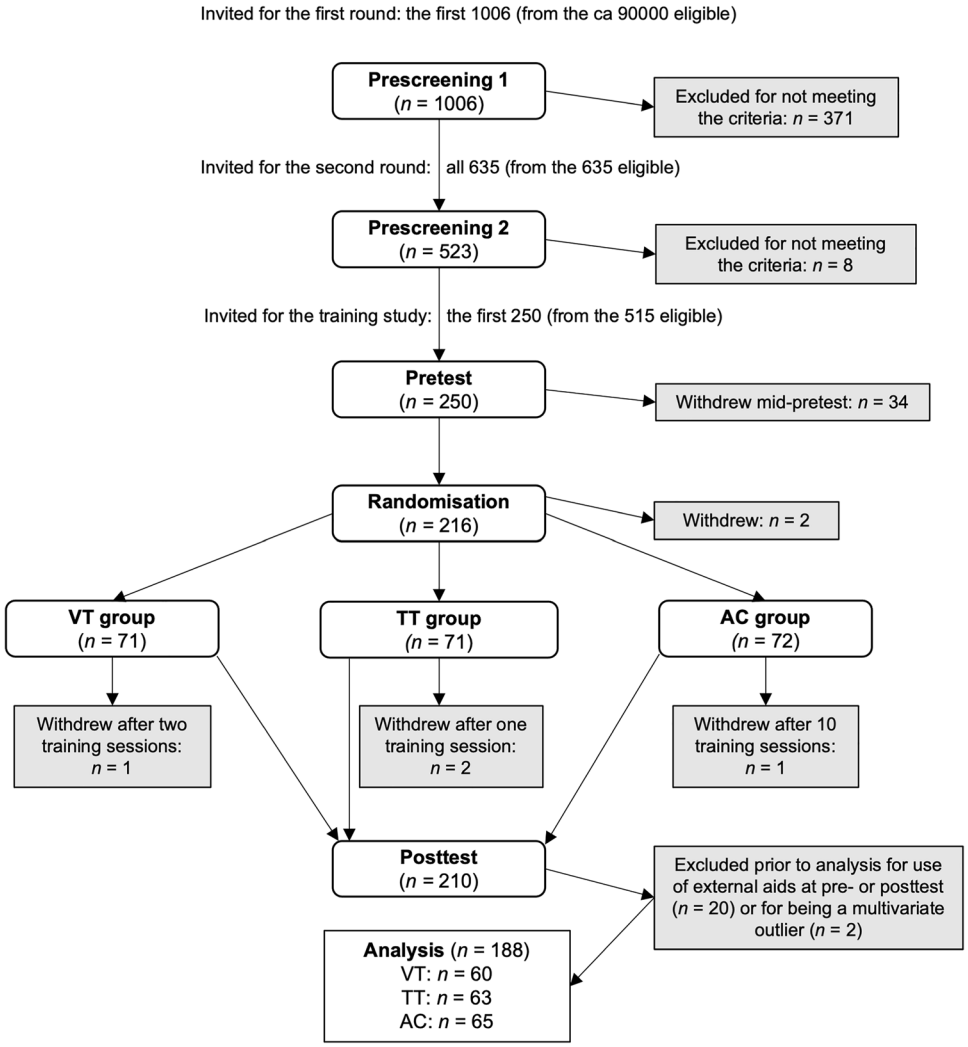

Fig. 1 Flowchart of the study design. VT, varied training; TT, traditional training; AC, active control

creating and distributing cognitive tasks and questionnaires. The data collection in this study was divided into two main stages over a 5-week period. The first phase of the study was prescreening, in which participants' eligibility was assessed and some background information was gathered. The second phase of the study was the actual training study, which consisted of a pretest, a training period, and a posttest (see Fig. 1 for an illustration of the study procedure).

From the original pool of ca 90,000 eligible participants, the first 1006 to sign up were recruited for the first round of prescreening. Of these, 371 participants were excluded for not meeting the screening criteria. Those who were older than 50 years, had a first language other than English, a psychiatric illness currently affecting their life, a neurological illness, a neurodevelopmental illness, poor and uncorrected eyesight, troubles reading or seeing the instructions, CNS medication or drug use (apart from alcohol and/or marijuana), or reported being intoxicated at the time of the study or heavy alcohol consumption on the night before, and those failing our simple attention checks or not reporting a functioning Prolific participant ID, were excluded after this stage (see Supplementary Information (SI), Supplement A, Table A1 for details).

From the remaining pool of 675 eligible participants, all were invited to continue onto the second prescreening round, which was subsequently completed by 523 participants. Of these, 8 participants were excluded for not meeting the screening criteria. Those who were extreme outliers at our cognitive prescreening tasks, or reported being intoxicated at the time of the study or heavy alcohol consumption on the night before, and those failing our attention check or not reporting a functioning Prolific participant ID, were excluded after this stage (see SI: Supplement A, Table A2 for details).

Of the remaining 515 participants, we recruited the first 250 (48.5%) who signed up for the actual WM training study. The first part of the WM training study comprised a pretest, which was successfully completed by 216 participants, whereas 34 discontinued. The 216 participants were then randomised into one of three training procedures: ***VT*** (71

participants), ***TT*** (71 participants), or ***AC*** (72 participants). Two participants discontinued right after the randomisation, two participants (TT $n = 2$) withdrew after the first training session, one participant (VT $n = 1$) after the second training session, and one participant (TT $n = 1$) after all 10 training sessions. The last part of the study was the posttest. All in all, 210 participants completed all of the aforementioned phases successfully. From the final sample of 188 participants who completed the study with acceptable data, 60 participants were in the VT group, 63 in the TT group, and 65 in the AC group (see Fig. 1 for an illustration of the study procedure). In the next section, we list more specific details of each phase.

Prescreening

The prescreening was spread over 1 week, and it consisted of two parts: prescreening 1 (ca 6 min) and prescreening 2 (ca 30 min).

Prescreening 1 The first round of prescreening consisted of several questionnaire items. The participants were first asked about their age, gender, education, health status, and alcohol and drug use. This was followed by two personality questionnaires. These were included for the sake of another study and will not be discussed here.

Prescreening 2 The second round of prescreening consisted of some questionnaire items and two cognitive tasks. The participants were first asked about their alcohol and drug use. This was followed by two cognitive tasks: one measuring fluid intelligence (the 16-item International Cognitive Ability Resource Sample Test; ICAR-16; Condon & Revelle, 2014), and the other attention (the antisaccade task; e.g. Kane et al., 2001). Although also a part of our screening protocol, these were mainly included for the sake of another study and will not be discussed here in detail.

Training Study

In week 2, the WM training study began. The first 250 eligible participants who completed the two prescreening phases were invited to the actual 4-week experiment. The training study was spread over weeks 2–5, consisting of a pretest session (ca 2 h 15 min), 10 training sessions (each ca 30 mins), and a posttest session (ca 2 h 15 min). In week 2, the participants completed the pretest and the two first training sessions. In weeks 3 and 4, the participants completed three training sessions per week. In week 5, the participants completed the last two training sessions and the posttest.

Pretest The pretest session consisted of a battery of cognitive tasks measuring WM and episodic memory. The participants also responded to questions on their strategy use after each task. Following task completion, they also filled in a number of questionnaires, including surveys on metacognitive and memory functions in their everyday life. Moreover, to ensure that potential confounds would not affect task performance, the participants were asked about possible prior experience with any of the tasks they just completed, as well as about their levels of motivation and alertness.

Training Period The participants were randomised into one of three training regimes. ***The VT group*** trained with altogether 20 continuously changing adaptive WM tasks during the course of their training. During each session, they trained for 5 min on altogether 6 different tasks (6×5 min = 30 min), and after each task, they responded to a question about their strategy use in that task. The idea was to create a training programme with several tasks that switched at a rather fast pace and where the tasks varied in paradigm (n-back, different spans, stimulus sequence recall), stimulus type (digits or spatial locations), and structure of the stimulus sequence (random or easily clusterable subseries). The aim was to promote flexible strategy use and achieve wider transfer effects to different WM and other cognitive tasks than achieved with traditional WM training (or training with a non-WM, general knowledge task that was administered to the controls). ***The TT group*** trained with a single adaptive WM task (n-back with digits) during their training. During each session, they trained for 30 min with this task, after which they responded to a question on their strategy use. Finally, ***the AC group*** trained with an adaptive general knowledge quiz task. During each session, they trained for 30 min with this task, after which they responded to a question on their strategy use.

Posttest The posttest session consisted of an identical battery of cognitive tasks as the pretest. The participants also responded to the same questions on their strategy use after each task. Following task completion, they also filled in a number of questionnaires, including a survey on strategy use in their everyday life. Moreover, the participants were asked about their levels of motivation and alertness.

Measures

Pre- and Posttest Tasks

All participants completed 11 cognitive tasks, including nine WM tasks and two episodic memory tasks, at both pre- and posttest. The 11 tasks represented six different paradigms: n-back tasks (with digits, letters, and colours), forward simple span tasks (with letters and colours), running memory tasks (with letters and colours), selective updating tasks (with digits and colours), episodic word list learning, and episodic word pair learning. All tasks, except for the n-back, word list learning, and word pair learning, included

practice trials before the actual tasks started. The tasks varied between ca 4 and 14 min in length, and the order of the tasks was randomised. In both training groups (VT and TT), especially the paradigm-wise overlap was expected to be an important factor in the occurrence of transfer (see, e.g., Soveri et al., 2017). Thus, we categorised the 11 tasks according to the paradigm-wise overlap with the VT training protocol: ***criterion-related training effects***, ***shared task-specific near transfer***, ***non-shared task-specific near transfer***, ***task-general near transfer***, and ***far transfer***.

Criterion-Related Training Effects One of our pre-posttest tasks, namely the n-back with digits, was also administered as a training task for both the VT and TT groups, but not for the AC group. This task therefore served as the criterion task for the VT and TT groups.

N-Back with Digits (NBD) In this adaptive updating task (Kirchner, 1958), digits ranging from 1 to 9 are presented on the screen one at a time. The task is to respond to whether the currently presented item corresponds to the item presented n items back. The participant responds to each stimulus by pressing on the designated "yes" or "no" button on the keyboard. The participants completed 12 blocks. The blocks consisted of n+20 trials, six of which were targets and 14 non-targets. The lowest level, 1-back, did not contain any lures, i.e. stimuli presented just before or after the target. To minimise familiarity-based responding, the higher levels, 2- to 12-back, contained four lures amongst the non-targets, two being presented just before and the other two just after a target. The trials in a block were displayed as follows: a blank screen for 450 ms, a stimulus displayed for 1500 ms, a blank screen for 450 ms, followed by the next stimulus. This task was adaptive in difficulty: the participant started at the easiest level, 1-back, and was able to reach 12-back at the highest. If the participant got 15–17 trials correct, the level for the next block remained the same. If 18 or more trials were correct, the level was increased by one. On the other hand, if 14 or fewer trials were correct, the level was decreased by one. The dependent variable was the average level of n the participant reached across the 12 blocks.

Shared Task-Specific Near Transfer Our two pre-posttest tasks categorised as shared task-specific transfer measures were untrained variants of the n-back task. These tasks measured task-specific near transfer for both the VT and TT groups. Due to the similarity in transfer type between these two groups, we decided to call this category *shared* task-specific near transfer.

N-Back with Letters (NBL) This task is the same as the NBD, the only difference being that the items are letters A to I instead of digits.

N-Back with Colours (NBC) This task is the same as the NBD, the only difference being that the items are coloured squares (red, green, blue, yellow, black, purple, orange, pink, and grey) instead of digits.

Non-shared Task-Specific Near Transfer We administered four WM measures from two different task paradigms. These tasks were untrained variants of a trained task for the VT group, while the TT group did not practise with these at all. Thus, these tasks measured task-specific near transfer for the VT group and task-general near transfer for the TT group. Due to the difference in transfer type between these two groups, we decided to call this category *non-shared* task-specific near transfer, after the VT group's transfer type.

Forward Simple Span with Letters (FSSL) This task is based on the classic simple span paradigm (Wechsler, 1997). In this WM task, letter sequences ranging from A to I of varying length are presented on the screen. The participant does not know beforehand when each sequence will end, but the task is always to recall the items in the order they are presented. The participant responds after each sequence by clicking the correct items in the correct order, on a row of horizontally aligned boxes with letters A to I shown on the screen. This task contained 6 trials, with sequence lengths 4–9, in a randomised order (we also included a sequence with 10 items but this sequence was not displayed due to a technical error). The stimulus presentation time was 1000 ms and the inter-stimulus interval 500 ms. The dependent variable was the total number of correctly recalled items in the correct order.

Forward Simple Span with Colours (FSSC) This task is the same as the FSSL, the only difference being that the items are coloured squares (red, green, blue, yellow, black, purple, orange, pink, and grey) instead of letters, and that there were altogether 7 trials, one of each length 4–10.

Running Memory with Letters (RML) This task is based on the paradigm by Pollack et al. (1959). In this WM task, letter sequences ranging from A to I of varying length are presented on the screen. The participant does not know beforehand when each sequence will end, but the task is always to recall the last 4 items in the order they are presented. The participant responds after each sequence by clicking the correct items in the correct order, on a row of horizontally aligned boxes with letters A to I shown on the screen. This task contained 8 trials, with sequence lengths 4–11, in a randomised order. The stimulus presentation time was 1000 ms and the inter-stimulus interval 500 ms. The dependent variable was the total number of correctly recalled items in the correct order.

Running Memory with Colours (RMC) This task is the same as the RML, the only difference being that the items are coloured squares (red, green, blue, yellow, black, purple, orange, pink, and grey) instead of letters.

Task-General Near Transfer We also administered two WM measures from the same task paradigm which were classified as task-general near transfer tasks. In other words, neither the VT nor the TT group had practised with this task paradigm. These tasks measured task-general near transfer for the VT and TT groups.

Selective Updating of Digits (SUD) This WM updating task is a slightly modified version of the task originally created by Murty et al. (2011). Five digits ranging from 0 to 9 are presented on the screen in a row of boxes. After 4000 ms, the digits disappear and a blank screen is presented for 100 ms. This is followed by an updating stage (lasting 2000 ms), in which a new row of boxes appears, with some of the boxes containing new digits and others being blank. The task is to recall the final sequence formed by the digits, taking into account the updates. This task contained 20 trials, and the order of trials was randomised for each participant. Half of the trials included only the initial sequence without any updates, while the other half included three updating stages. The dependent variable was the number of correctly recalled digits in the correct order on the updating trials. For more details, see Fellman et al. (2020b).

Selective Updating of Colours (SUC) This task is the same as the SUD, the main difference being that the items are coloured squares (red, green, blue, yellow, black, purple, orange, pink, and grey) instead of digits. The other difference between the versions is that here, the stimulus display times were a little longer: the original colour row disappeared after 7000 ms and the updating stage lasted 5000 ms.

Far Transfer Lastly, we administered two tasks that tap on processes different to WM measures, namely two episodic memory tasks. Following a common convention in WM training research (Jefferies et al., 2004; Klem et al., 2015), they are thus labelled as far transfer tasks, albeit being memory tests. These tasks measured far transfer for the VT and TT groups.

Word List Learning (WLL) In this episodic memory task, a list of 15 words was presented, one word at a time. The task was to memorise as many words as possible. After all the words had been displayed, the participant had to respond to a simple mathematical task (this served to minimise the role of WM in recall). This was a simple arithmetical operation task lasting ca 1 min in duration (e.g. 6+5–4+6 = ?), to which the participants responded by typing in their answer. Next, the participant was shown a page with 15 empty boxes, i.e. one for each word, and was asked to type in the words from the list, in any order. This was followed by a second round using the same word stimuli but a different mathematical task. There were two versions of this task, one with word list A and the other with word list B. Half of the participants received this task with list A at pretest and list B at posttest, and the other half vice versa. The reason for this was that we did not want the posttest result reflect learning from already presented words. The dependent variable was the total number of correct words (see SI: Supplement B for details on how the stimulus words were chosen).

Word Pair Learning (WPL) In this episodic memory task, a list of 10 word pairs (altogether 20 words) was presented, one word pair at a time. The task was to memorise as many word pairs as possible. After all the word pairs had been displayed, the participant had to respond to a simple mathematical task, similar to that described for WLL. Next, the participant was shown a page with 10 words, i.e. the first word of each pair, and 10 empty boxes (one for each missing word), and asked to type in the missing word next to their counterpart. This was followed by a second round using the same word stimuli but a different mathematical task. As with WLL, there were two versions of this task, one with word pair list A and the other with word pair list B. Half of the participants received this task with list A at pretest and list B at posttest, and the other half vice versa. The dependent variable was the total number of correct word pairs (see SI: Supplement B for details on how the stimulus words were chosen).

Varied Training

The VT group trained with altogether 20 WM tasks during their training period (for more details, see SI: Supplement C). This list of tasks included five different paradigms: n-back tasks (with digits and spatial locations), forward simple span tasks (with digits and spatial locations), backward simple span tasks (with digits and spatial locations), running memory tasks (with digits and spatial locations), and paired recall (with digits and spatial locations). Moreover, in half of the tasks, the stimulus sequences were random, while the other half contained also specially designed, easily clusterable subseries. The idea of this manipulation was that it may promote utilisation of grouping strategy in WM tasks. In other words, detection and chunking of regular subseries (e.g. ...7-1-5-**2-4-6-8**-1-9-3...) might make one prone to employ grouping strategy more broadly, irrespective whether the sequence includes apparent regularities or not. Grouping or chunking represents a potentially more general strategy that can be applied in a more structured way in different memory tasks (Dunlosky & Kane, 2007; Jones,

2012), and it is not limited to verbal materials (Oberauer et al., 2018). For information on the composition of the easily clusterable subseries, see SI: Supplement C, Table C2 for the digit tasks and SI: Supplement C, Table C3 for the tasks containing spatial locations.

The tasks lasted ca 5 min each, and each 30-min session contained 6 tasks. This means that the participant trained with each of the tasks altogether three times and ca 15 min (3 × 5 min) during the course of their training period. The presentation order of the tasks within each session was randomised. However, the placement of the tasks into specific sessions was fixed and followed a rotating order, ensuring that each task was practised with at approximately equal intervals (see SI: Supplement C, Table C1 for details). Each task was practised first in either session 1, 2, 3, or 4, for the second time in session 4, 5, 6, or 7, and finally in session 7, 8, 9, or 10.

Traditional Training

The TT group trained with a single WM task, adaptive n-back with digits, throughout their training period. Thus, they spent 10 sessions with the digit n-back, training in total ca 5 hours (10 x 30 min) with this task.

N-Back with Digits, TT training (NBD-TT) This task is similar to the pre- and posttest version (NBD), the main difference being that there are as many as 20 blocks in each session. The highest level that can be reached during the course of the training period is 15-back. The task is adaptive both within and between the sessions, following the same rules as the pre-posttest version. The starting level in the first training session was 1-back.

Active Controls and Quiz Training

The AC group trained with a quiz game throughout their training period. Thus, they spent 10 sessions with the quiz, training in total ca 5 h (10 × 30 min) with this task.

Quiz Game (QG-AC) In this general knowledge quiz task, the participant answers multiple-choice questions, one at a time. There are 7 blocks in this task, each containing 20 questions. This task is adaptive in difficulty, and the participant starts at the easiest level, having two alternative choices to pick from. If the participant gets 15–17 questions correct, the difficulty level for the next block remains the same. If 18 or more trials are correct, the number of choices is increased by one. On the other hand, if 14 or fewer trials are correct, the number of choices is decreased by one (but never below the starting level). The maximum level in this task is four alternative choices. The pool includes approximately 850 questions, as we wanted the participants to encounter some of the questions more than once during their training period. The questions are from a broad range of categories, such as general knowledge, geography, history, politics, science and nature, books, films, music, and sports.

Strategy Use, Motivation, Alertness, and Training Expectations

Throughout the study, the participants were asked about their strategy use, as well as about their level of motivation and alertness. Moreover, before beginning their training period, they were asked to estimate how much their pre-post task performance would improve following their forthcoming training. Strategy use will be analysed and discussed in a separate article.

Expectations of Improvement In the very beginning of training session 1, the participants rated their expectations of improvement at each pre-posttest task following the training period. They were asked: "How well do you think you will perform on this task at posttest compared to pretest?", replying on a scale of 1 to 10 (1 = "The same level of performance as at pretest", 10 = "Very much better performance than at pretest"). At this point, the participants had been informed what type of training they would be engaging with but did not yet have experience with the training.

Motivation At the end of pre- and posttests, as well as at the end of training sessions 1, 5, and 10, the participants rated their level of motivation. They were asked: "How motivated were you to perform the tasks?", replying on a scale of 1 to 5 (1 = "Not at all motivated", 5 = "Very motivated").

Alertness At the end of pre- and posttests, as well as at the end of training sessions 1, 5, and 10, the participants were also asked to rate their level of alertness. They were asked: "How alert are you at the moment?", replying on a scale of 1 to 5 (1 = "Very tired", 5 = "Very alert").

Data Pre-processing

The data was processed before analyses to screen for (1) cheating, (2) multivariate outliers, and (3) listwise for univariate outliers, missing data and unreliable effort at pretest, and taking into account possible colour blindness in the case of colour stimuli. From the 210 participants that completed the study, we first excluded those who responded "Yes" to the item "Did you use external tools (for example, writing, taking notes, or drawing) to help you solve the tasks?" at either pre- or posttest. The participants were told that their honest response was critically important and would not affect their payment in any way. Following this, 20 participants were excluded from all analyses.

Table 1 Summary table of the reasons for excluding participants from the pre-posttest transfer analyses

Task	Domain	Missing data	Colour blindness	Unreliable effort	Number of participants (*N*) in analyses			
					N	VT	TT	AC
	Criterion task	2	0	3	183	58	62	63
NBD		2	0	3	183	58	62	63
	Shared TSNT	0	5	4	179	54	63	62
NBL		0	0	3	185	57	63	65
NBC		0	5	2	181	56	63	62
	Non-shared TSNT	1	5	0	182	57	63	62
FSSL		0	0	0	188	60	63	65
FSSC		0	5	0	183	57	63	63
RML		1	0	0	187	60	63	64
RMC		0	5	0	183	57	63	63
	TGNT	0	5	0	183	57	63	63
SUD		0	0	0	188	60	63	65
SUC		0	5	0	183	57	63	63
	FT	0	0	0	188	60	63	65
WLL		0	0	0	188	60	63	65
WPL		0	0	0	188	60	63	65

Note. NBD, n-back with digits; *NBL*, n-back with letters; *NBC*, n-back with colours; *FSSL*, forward simple span with letters; *FSSC*, forward simple span with colours; *RML*, running memory with letters; *RMC*, running memory with colours; *SUD*, selective updating of digits; *SUC*, selective updating of colours; *WLL*, word list learning; *WPL*, word pair learning; *TSNT*, task-specific near transfer; *TGNT*, task-general near transfer; *FT*, far transfer; *VT*, varied training group; *TT*, traditional training group; *AC*, active control group

Before conducting analyses on the background variables, we also excluded all participants who were multivariate outliers at the 11 pretest tasks. Multivariate outliers were predefined as scoring below the threshold of $p < .001$ in the Mahalanobis distance value ($\chi 2$ (11, 188) = 31.26; Tabachnick & Fidell, 2007). Two participants (TT group $n = 2$) exceeded this cut-off value on the tasks. Thus, the final sample included 188 participants (VT group $n = 60$, TT group $n = 63$, AC group $n = 65$).

Before conducting analyses on the pre-posttest task gains, we also excluded listwise the performance of those participants who had missing data at pre- or posttest, colour blindness in the case of colour stimuli, or unreliable effort at pretest, as well as those who were univariate outliers at the pretest tasks. Unreliable effort was present if a participant remained at the lowest level in the n-back tasks, or did not recall any items correctly in the other tasks. Univariate outliers were predefined as scoring three times the interquartile range above or below the 1st or the 3rd quartile. However, no participant exceeded this cut-off on any of the tasks. Thus, the final sample included 181–188 participants for the task-specific analyses and 179–188 participants for the domain-specific analyses (note that we performed a listwise exclusion in the domain-specific analyses if a given participant did not meet our inclusion criteria in any of the single tasks within a domain). Table 1 depicts a summary of the exclusions and the final sample included in the task-specific and the domain-specific pre-posttest analyses following this procedure.

Data Analysis

The data was analysed in the R Environment version 4.0.3 (R Core Team, 2017), using the "BayesFactor" package (Morey et al., 2018) for computing the Bayes Factors. The Bayes Factor (BF) approach allows the researcher to measure the evidence for the null hypothesis or for the alternative hypothesis on a continuous scale with a range of 1–∞. A BF with a value of 1 indicates no support for either hypothesis, whereas a value above 1 indicates evidence for the alternative hypothesis, and a value below 1 evidence for the null hypothesis (Jeffreys, 1961; Kass & Raftery, 1995). The interpretation of the BFs in this study followed the guidelines proposed by Kass and Raftery (1995), where BFs between 1 and 3 are defined as "weak evidence", BFs between 3 and 20 as "positive evidence", BFs between 20 and 150 as "strong evidence", and BFs > 150 as "very strong evidence". In each BF analysis, we used the default prior setting (i.e. Cauchy distribution using a scaling factor $r = .707$). Besides BFs, we also report estimates of between-group mean differences using a posterior distribution with 10 000 iterations coupled with their 95% credible intervals.

For assessing baseline comparability between the groups, we analysed the participants' gender distribution, age,

Table 2 Background characteristics for the three groups ($N = 188$)

Measure	Group			Pairwise group comparisons BF_{H1}		
	VT group	TT group	AC group	VT vs. TT	VT vs. AC	TT vs. AC
Sample size (*n*)	60	63	65			
Gender (F/M)*	34/26	35/28	39/26	1/4.50[a]	1/4.29[a]	1/4.09[a]
Age (*M, SD*)	32.43 (7.84)	32.35 (8.04)	32.28 (8.74)	1/5.19[b]	1/5.21[b]	1/5.29[b]
Years of education (*M, SD*)	15.18 (3.56)	16.78 (3.66)	16.48 (2.72)	2.79[b]	2.01[b]	1/4.66[b]

Note. Estimates are from 10 000 samples of the posterior distribution. *VT*, varied training group; *TT*, traditional training group; *AC*, active control group

[a]Bayesian Pearson chi-square test

[b]Bayesian ANOVA

*None of the participants in our final sample chose *other* as their response

education length, pretest performance, and expectations they had of their prospective improvement from pre- to posttest. Moreover, we analysed the levels of motivation and alertness at different time points in the study (at pretest, beginning, mid, and end of the training period, and at posttest) for ruling out potential confound effects (Boot et al., 2013). This was done pairwise with either Bayesian chi-square tests (gender) or Bayesian analysis of variance (ANOVA; all other variables). As we were interested in the differences between each of the group pairs (VT vs. TT, VT vs. AC, and TT vs. AC), we computed BFs between each pair.

For assessing group differences with respect to the improvements from pre- to posttest, we analysed change in performance in each of the eleven pre-posttest tasks (NBD, NBL, NBC, FSSL, FSSC, RML, RMC, SUD, SUC, WLL, and WPL). Pre-posttest improvements were examined with Bayesian analysis of covariance (ANCOVA)[1], where posttest performance served as the dependent variable, group as the between-subjects factor, and pretest performance as the covariate. BFs were computed for change in each pre-posttest task for each paired group comparison.

For a further, more general assessment of group differences with respect to the improvements from pre- to posttest, we also analysed change in performance in each of the four transfer domains (shared task-specific transfer, non-shared task-specific transfer, task-general transfer, and far transfer), as categorised from the perspective of the VT group. As before, pre-posttest improvements were examined with Bayesian ANCOVA. To weigh the tasks within the transfer domains equally, we used the mean of standardised *z* scores for the tasks in the domain. BFs were computed for change in each transfer domain for each paired group comparison.

[1] Even though in our pre-registration we had originally planned to conduct a different path of analysis, reviewer comments prompted us to switch to ANCOVA in order to account for the detected pretest group differences and to avoid reliability issues with gain scores.

Results

Demographic Variables, Pretest Task Performance, Training Gain Expectations, Motivation, and Alertness

As depicted in Table 2, the results showed that the groups were comparable with respect to their gender distribution, age, and education length. Moreover, the three groups showed comparable pretest performance (see SI: Supplement D), expectations of improvement (see SI: Supplement E, Table E1), motivation at pre- and posttest (see SI: Supplement E, Table E2), and alertness at pre- and posttest (see SI: Supplement E, Table E3), as there was no positive evidence for differences between the groups on these variables (BF_{H1}s < 3). The groups also showed comparable levels of training motivation, albeit some group differences in training alertness were observed at some assessment points: the AC group had higher ratings of alertness than the VT group during the first ($BF_{H1} = 22.57$) and fifth session ($BF_{H1} = 8.36$), and higher ratings of alertness than the TT group during the first ($BF_{H1} > 150$), fifth ($BF_{H1} = 12.34$), and tenth session ($BF_{H1} = 7.92$). However, average training alertness (*M* of the three training period ratings) showed very weak correlations with pre-post gains (VT group, $r = -.14$ to .09; TT group, r −.16 to .18; AC group, $r = -.13$ to .19), suggesting that these group differences did not have any substantive impact on the transfer outcomes.

Task-Specific Pre-posttest Gains

The pre-posttest improvements in the three groups' performance were examined at task-level utilising Bayesian ANCOVAs, testing for differences in posttest performance for each task with pretest performance as a covariate, in each paired group comparison (see Table 3 and Fig. 2). Starting with the criterion task (NBD) (see Fig. 2A), the TT group fared better in comparison to both the VT group and the AC group. There was very strong evidence for a difference

Table 3 Parameter estimates for the taskwise Bayesian ANCOVAs on the improvements from pretest to posttest

Task	Domain	VT group vs. TT group[a]		VT group vs. AC group[a]		TT group vs. AC group[b]	
		M_{diff} [95% HDI]	BF_{H1} ± error %	M_{diff} [95% HDI]	BF_{H1} ± error %	M_{diff} [95% HDI]	BF_{H1} ± error %
NBD	Criterion task	−0.40 [−0.59, −0.22]	**> 150 ± 1.38**	0.16 [0.01, 0.30]	1.31 ± 2.08	0.57 [0.40, 0.74]	**> 150 ± 0.99**
NBL	Shared TSNT	−0.26 [−0.43, −0.08]	**9.52 ± 2.23**	0.06 [−0.08, 0.21]	1/5.00 ± 2.21	0.33 [0.16, 0.48]	**> 150 ± 2.18**
NBC	Shared TSNT	−0.20 [−0.38, −0.02]	1.66 ± 1.78	0.07 [−0.08, 0.22]	1/4.55 ± 4.58	0.28 [0.12, 0.44]	**41.19 ± 1.53**
FSSL	Non-shared TSNT	1.78 [0.63, 3.02]	**8.24 ± 0.78**	0.95 [−0.15, 2.07]	1/1.67 ± 2.46	−0.76 [−1.79, 0.33]	1/2.78 ± 0.85
FSSC	Non-shared TSNT	1.63 [0.04, 3.30]	1/1.05 ± 1.21	0.95 [−0.65, 2.60]	1/3.57 ± 0.89	−0.67 [−2.21, 0.98]	1/4.76 ± 9.17
RML	Non-shared TSNT	0.69 [−0.22, 1.53]	1/2.13 ± 1.54	0.37 [−0.52, 1.22]	1/5.26 ± 1.04	−0.23 [−1.03, 0.60]	1/6.25 ± 2.22
RMC	Non-shared TSNT	0.57 [−0.33, 1.49]	1/3.23 ± 1.48	0.92 [0.03, 1.78]	1.15 ± 1.31	0.33 [−0.46, 1.09]	1/4.76 ± 3.50
SUD	TGNT	1.26 [0.19, 2.37]	1.67 ± 0.98	0.40 [−0.74, 1.53]	1/5.88 ± 0.94	−0.76 [−1.86, 0.36]	1/2.94 ± 0.97
SUC	TGNT	1.32 [−0.07, 2.76]	1/1.28 ± 0.94	0.45 [−1.02, 1.91]	1/5.88 ± 0.80	−0.76 [−2.10, 0.58]	1/3.85 ± 0.92
WLL	FT	−0.17 [−1.00, 0.74]	1/6.67 ±1.83	0.04 [−0.80, 0.87]	1/7.14 ± 1.80	0.11 [−0.68, 0.93]	1/7.14 ± 0.87
WPL	FT	0.20 [−0.48, 0.90]	1/5.88 ± 0.86	0.07 [−0.65, 0.83]	1/7.14 ± 1.07	−0.10 [−0.68, 0.46]	1/6.25 ± 6.39

Note. Bolded values indicate Bayes factors of 3 or greater. Estimates are from 10 000 samples of the posterior distribution; $M_{diff,}$ mean group differences; *HDI*, highest density interval; *NBD*, n-back with digits; *NBL*, n-back with letters; *NBC*, n-back with colours; *FSSL*, forward simple span with letters; *FSSC*, forward simple span with colours; *RML*, running memory with letters; *RMC*, running memory with colours; *SUD*, selective updating of digits; *SUC*, selective updating of colours; *WLL*, word list learning; *WPL*, word pair learning; *TSNT*, task-specific near transfer; *TGNT*, task-general near transfer; *FT*, far transfer; *VT*, varied training group; *TT*, traditional training group; *AC*, active control group

[a]Positive values represent greater performance in the VT group

[b]Positive values represent greater performance in the TT group

between TT and VT (BF_{H1} > 150 ± 1.38%), as well as between TT and AC (BF_{H1} > 150 ± 0.99%). Only weak evidence for a difference in NBD between VT and AC was obtained, with VT having fared slightly better.

As for training-induced improvements in the shared task-specific near transfer tasks (see Fig. 2B–C), the TT group showed superior pre-posttest gains on both tasks. In NBL, there was very strong evidence for a difference between the TT and AC groups (BF_{H1} > 150 ± 2.18%) as well as strong evidence for a difference between TT and VT (BF_{H1} = 9.52 ± 2.23%). In NBC, there was also strong evidence for a difference between TT and AC (BF_{H1} = 41.19 ± 1.53%). However, we observed positive evidence *against* the effect of group between the VT group and the AC group in both NBL and NBC (BF_{H1} < 1/3). Only weak evidence for a difference between VT and TT in NBC was obtained, with TT performing slightly better.

In regard to the non-shared task-specific near transfer tasks (see Fig. 2D–G), the VT group showed more improvement as compared to the TT group but not the AC groups on one of the four tasks. Specifically, there was positive evidence for a difference in FSSL when the VT group was compared against the TT group (BF_{H1} = 8.24 ± 0.78%). However, there was positive evidence *against* the effect of group between the VT and TT groups in RMC, between VT and AC in FSSC and RML, and between TT and AC in FSSC, RML, and RMC (BF_{H1} < 1/3). Only weak evidence either for or against a difference in posttest performance while controlling for pretest was obtained between the VT and TT groups in FSSC and RML, between VT and AC in FSSL and RMC, and between TT and AC in FSSL (BF_{H1} range < 3–> 1/3).

Concerning task-general near transfer (see Fig. 2H–I), no positive evidence for group differences was detected. Only weak evidence either for or against a difference in posttest performance while controlling for pretest was obtained between VT and TT in SUD and SUC, as well as between TT and AC in SUD (BF_{H1} range < 3–> 1/3). Moreover, there was positive evidence *against* the effect of group between VT and AC in both SUD and SUC, as well as between TT and AC in SUC (BF_{H1} < 1/3). Finally, we observed no group differences in either far transfer task (see Fig. 2J–K). In fact, all far transfer analyses revealed positive evidence *against* group differences (BFs < 1/3).

Domain-Specific Pre-posttest Gains

The pre-posttest improvements in the three groups' performance were also examined at domain-level utilising Bayesian ANCOVAs, testing for differences in posttest performance for each domain with pretest performance as a covariate, in each paired group comparison (see Table 4 and Fig. 3).

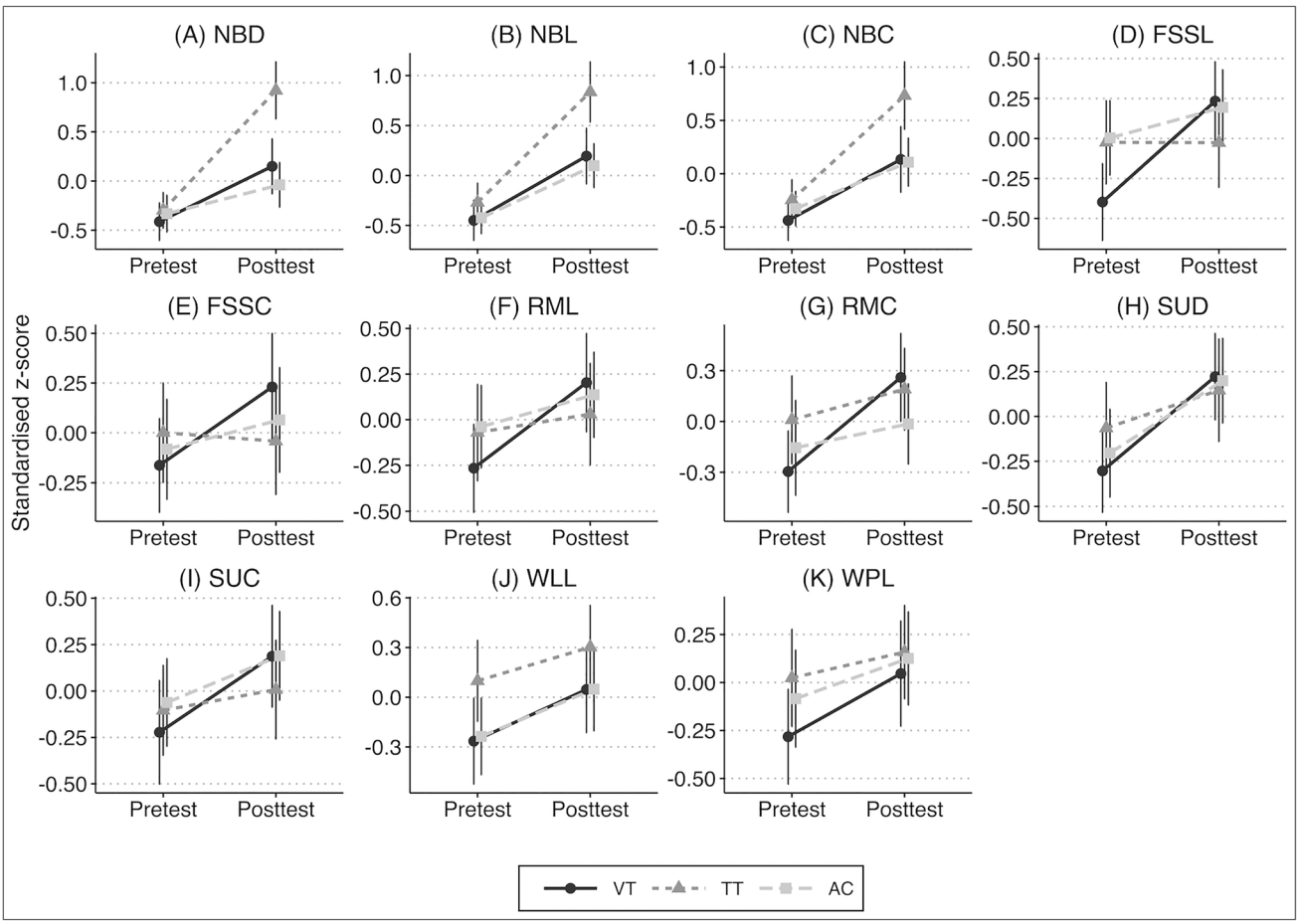

Fig. 2 (A–I) Standardised task-level pre-posttest improvements grouped by intervention. VT, varied training group; TT, traditional training group; AC, active control group. NBD, n-back with digits; NBL, n-back with letters; NBC, n-back with colours; FSSL, forward simple span with letters; FSSC, forward simple span with colours; RML, running memory with letters; RMC, running memory with colours; SUD, selective updating of digits; SUC, selective updating of colours; WLL, word list learning; WPL, word pair learning. Error bars represent 95% confidence intervals. The participants' performances were standardised within their respective task across our two measurement points (this was done for illustrative purposes)

As for training-induced improvements in the shared task-specific near transfer tasks (see Fig. 3A), the TT group showed superior pre-posttest gains in this domain compared to both the VT group ($BF_{H1} = 8.41 \pm 2.43\%$) and the AC group ($BF_{H1} > 150 \pm 1.11\%$). However, we observed positive evidence *against* the effect of group between the VT group and the AC group ($BF_{H1} > 1/3.70 \pm 1.47\%$), indicating that the VT participants' performance in the untrained n-back tasks was not improved following intervention.

In regard to the non-shared task-specific near transfer domain (see Fig. 3B), the VT group showed more improvement as compared to the TT group but not the AC group. Specifically, there was strong evidence for a difference when the VT group was compared against the TT group ($BF_{H1 =} 24.40 \pm 2.35\%$) and weak evidence when AC served as the reference ($BF_{H1} = 2.50 \pm 2.59\%$). However, there was positive evidence *against* the effect of group between the TT and AC groups in the non-shared task-specific domain ($BH_{H1} = 1/3.70 \pm 20.41\%$).

Concerning the task-general near transfer domain (see Fig. 3C), no positive evidence for group differences was detected. Only weak evidence either for or against a difference in posttest performance while controlling for pretest was obtained between VT and TT, as well as between TT and AC (BF_{H1} range < 3–> 1/3). Moreover, there was positive evidence *against* the effect of group between VT and AC ($BF_{H1} < 1/3$). Finally, we observed no group differences in the far transfer domain (see Fig. 3D). In fact, all far transfer analyses revealed positive evidence *against* group differences (BFs < 1/3).

Discussion

In the present study, we developed and tested a novel WM training protocol that was based on the varied training principle that has elicited broader transfer than repetitive

Table 4 Parameter estimates for the domainwise Bayesian ANCOVAs on the improvements from pretest to posttest

Domain	Tasks	VT group vs. TT group[a]		VT group vs. AC group[a]		TT group vs. AC group[b]	
		M_{diff} [95% HDI]	BF_{H1} ± error %	M_{diff} [95% HDI]	BF_{H1} ± error %	M_{diff} [95% HDI]	BF_{H1} ± error %
Shared TSNT	NBL, NBC	−0.17 [−0.29, −0.05]	**8.41 ± 2.43**	0.05 [−0.04, 0.15]	1/3.70 ± 1.47	0.24 [0.12, 0.35]	**> 150 ± 1.11**
Non-shared TSNT	FSSL, FSSC, RML, RMC	0.18 [0.07, 0.29]	**24.40 ± 2.35**	0.13 [0.02, 0.24]	2.50 ± 2.59	−0.04 [−0.13, 0.04]	1/3.70 ± 20.41
TGNT	SUD, SUC	0.13 [0.03, 0.23]	2.20 ± 1.55	0.04 [−0.07, 0.15]	1/5.56 ± 2.30	−0.08 [−0.18, 0.02]	1/2.04 ± 1.23
FT	WLL, WPL	0.03 [−0.09, 0.15]	1/6.25 ± 1.81	0.02 [−0.10, 0.13]	1/6.67 ± 2.92	−0.02 [−0.11, 0.08]	1/6.67 ± 3.26

Note. Bolded values indicate Bayes factors of 3 or greater. Estimates are from 10,000 samples of the posterior distribution; M_{diff}, mean group differences; *HDI*, highest density interval; *TSNT*, task-specific near transfer; *TGNT*, task-general near transfer; *FT,* far transfer; *NBD*, n-back with digits; *NBL*, n-back with letters; *NBC*, n-back with colours; *FSSL*, forward simple span with letters; *FSSC*, forward simple span with colours; *RML*, running memory with letters; *RMC*, running memory with colours; *SUD*, selective updating of digits; *SUC*, selective updating of colours; *WLL*, word list learning; *WPL*, word pair learning; *VT*, varied training group; *TT*, traditional training group; *AC*, active control group

[a]Positive values represent greater performance in the VT group

[b]Positive values represent greater performance in the TT group

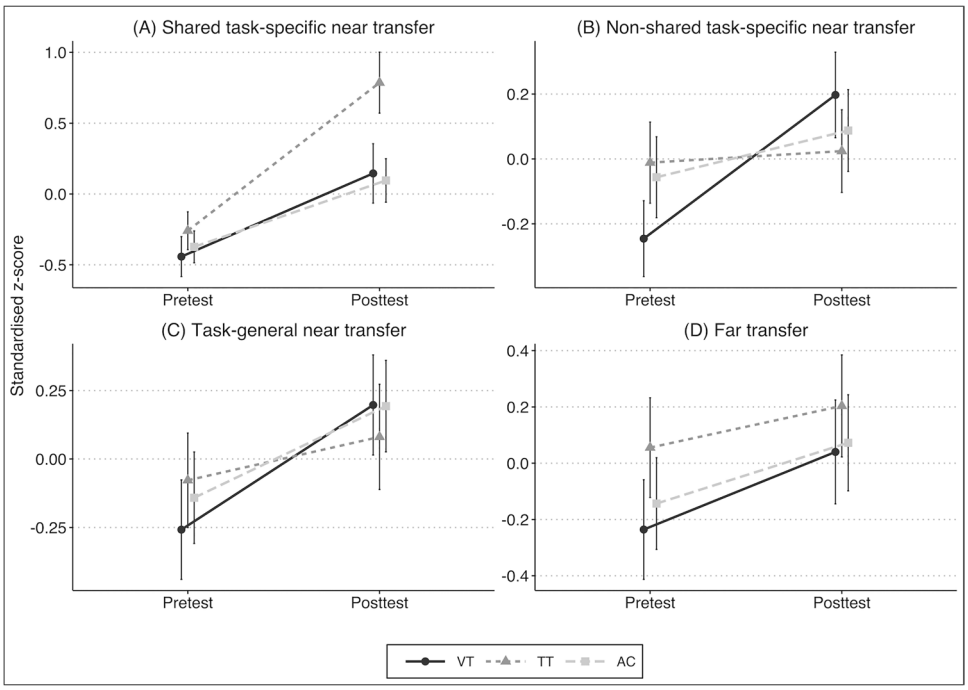

Fig. 3 (A–D) Standardised domain-level pre-posttest improvements grouped by intervention. VT, varied training group; TT, traditional training group; AC, active control group. Shared task-specific near transfer (n-back with letters, n-back with colours); non-shared task-specific near transfer (forward simple span with letters, forward simple span with colours, running memory with letters, running memory with colours); task-general near transfer (selective updating of digits, selective updating of colours); far transfer (word list learning, word pair learning). Error bars represent 95% confidence intervals. Before averaging the scores in a given domain, the participants' performances were standardised within their respective task across our two measurement points (this was done for illustrative purposes)

practice with the same tasks in other cognitive domains. Our VT protocol entailed quickly shifting WM tasks employing variable paradigms, stimuli, and stimulus sequence predictability. Inspired by the success of earlier VT research in other domains (Braun et al., 2009; Karbach & Kray, 2009; Reed et al., 2019; Sabah et al., 2019; Sanders et al., 2002; Vakil & Heled, 2016), we expected that VT would encourage the development of more general

rules (strategies) to solve WM tasks, and that this would result in the generalisation of training effects stretching beyond the trained task and its untrained variants. For this purpose, we ran a pre-registered randomised controlled trial with methodological and statistical rigor, comparing VT training ($n = 60$) to traditional training with a single adaptive WM task (TT; $n = 63$), as well as to an active control group practising with a general knowledge quiz task (AC; $n = 65$). As hypothesised, after the 4-week intervention, the TT group evidenced task-specific near transfer when compared with the other two groups. The VT group also showed some evidence for task-specific near transfer, but, against our first hypothesis, yielded no evidence for task-general near transfer or far transfer. We discuss these findings below.

Concerning the TT group, our hypothesis stated that it would exhibit only task-specific near transfer after training when compared to the AC group. This hypothesis received clear support for both untrained n-back transfer tasks, i.e. n-back with letters and with colours. On the letter, but not the colour variant of n-back, the TT group was also superior to the VT group, even though the VT group also had an n-back task in their training protocol. It may be that the relative similarity and/or the overlearned nature of the stimuli in the trained n-back task vs. the untrained letter version (alphanumeric, i.e. digits and letters) led to the TT group's stronger transfer on letter vs. colour n-back. All in all, these findings concur with the meta-analysis on transfer after n-back training (Soveri et al., 2017) and demonstrate once again the very limited transfer following traditional repetitive WM training. At the same time, these results confirm that our web-based experiment worked as expected.

As for the VT group, our hypothesis was that, as compared to the AC and the TT groups, the VT group would show transfer also to other memory tasks than the untrained variants of the training tasks. Starting from task-specific near transfer, the VT group surpassed the TT but not the AC group only on one such transfer task, namely forward simple span with letters. On three other measures tapping task-specific near transfer for the VT group, no evidence for group differences in pre-post gains was found. Again, we can only speculate whether the simplicity as well as the relative similarity and/or the overlearned nature of the stimuli in the trained vs. the transfer-positive task (digit and letter forward simple span with digits vs. letters) contributed to this single transfer effect. This limited task-specific near transfer suggests that the time devoted to training a given task paradigm was too short for the evolvement of effective task-specific skills in the VT group. The pattern is similar to the intermediate WM training results by Fellman et al. (2020a), where three sessions devoted to traditional digit n-back training resulted in task-specific near transfer to letter n-back but not to n-back with colours or boxes. In turn, at posttest following the full 12-session training with digit n-back, all three untrained n-back tasks used in their study evidenced task-specific near transfer.

The critical hypothesis was that VT training would yield generalisation beyond task-specific near transfer. This hypothesis failed to gain support. As regards task-general transfer, there was merely weak support for the VT group's success in one of the two relevant measures, selective updating of digits, but only when compared to the TT group and not in comparison to the AC group. As speculated above when discussing the task-specific near transfer findings, one could conjecture that transfer may emerge easier with overlearned stimuli (digits), but we should have seen a difference between the VT group and the active controls as well, besides a clear difference between the VT and TT groups. Concerning far transfer that we tapped with two verbal episodic memory tasks, no group differences in pre-post gains emerged. Thus, there was no evidence for wider transfer with the present VT protocol. The lack of wider transfer following varied WM training is in line with the results reported by Redick et al. (2020) who systematically varied the number of stimulus types (letters + digits + words vs. only letters) in their training study.

Why did our VT protocol fail to provide broader transfer that has nevertheless been seen in intervention studies concerning several other cognitive domains (Braun et al., 2009; Karbach & Kray, 2009; Reed et al., 2019; Sabah et al., 2019; Sanders et al., 2002; Vakil & Heled, 2016)? First, the aim with VT was to encourage the development of more general rules (strategies) to solve WM tasks, but the trained paradigms and stimuli may have been too dissimilar for more abstract WM task representation and general rules such as clustering (as such a broadly applicable strategy, e.g. Dunlosky & Kane, 2007; Jones, 2012) to emerge. For example, strategies that were developed for n-back may have been quite different from those employed in running memory. In this respect, we refer to Laine et al. (2018) for a particularly effective strategy tailored for n-back that would not be feasible for example with span tasks. Thus, one can raise the question as to whether there would be a "sweet spot" concerning the similarities between the trained tasks (neither too similar nor dissimilar) that would best facilitate structural learning instead of the mere development of task-specific skills. Furthermore, the selective updating paradigm, the critical measure for task-general transfer, is yet a different task and its relationship to the trained WM tasks may not be apparent. The same may be true for the far transfer tasks. Second, the participants in the VT group were not explicitly instructed to search for general solutions that could be applied to a variety of WM tasks. Instead, any structural learning would have been driven solely by the participant. Third, our very recent analyses of self-reported strategies indicate

that the most active strategy generation phase takes place within the first few minutes into the memory task, after which strategy use stabilises (Waris et al., 2021a, b). Thus, the metacognitive phase, during which the rules for the trained WM tasks were generated and selected (Chein & Schneider, 2012), may have encompassed only a small portion of the time allotted to training. If these brief periods during VT did not include reflections on similarities between the tasks and their consequences for strategy selection, most of the VT participants' strategic resources would have been spent on developing task-specific routines. Strategy considerations may have occupied VT participants even less if the task paradigm is central for determining the way one tries to solve a task (cf. Gathercole et al., 2019): while the VT protocol included altogether 20 WM tasks, these tasks encompassed only 5 different paradigms. It might also be that the VT protocol involved too many tasks with too little time per task, rendering VT practice more like task exposure than task training.

The present study joins a rather long list of methodologically stringent individual studies and meta-analyses (e.g. Melby-Lervåg et al., 2016; Soveri et al., 2017) that report only very narrow improvements following WM training. These improvements are primarily seen on the trained task and its untrained variants, reflecting the "curse of specificity" that has been noted in skill learning research. At the same time, varied training within the skill learning framework has met with some success in other cognitive domains (Braun et al., 2009; Karbach & Kray, 2009; Reed et al., 2019; Sabah et al., 2019; Sanders et al., 2002; Vakil & Heled, 2016), and we see no fundamental reason as to why it could not work for WM tasks as well. Mnemonic feats indicate that systematic implementation of suitable strategies can dramatically improve performance on individual memory tasks. While no single mnemonic strategy would suit all tasks, it is important to point out that some strategies are more generalisable than others. As we noted earlier, grouping or chunking represents a strategy that can be applied to a variety of memory tasks (Dunlosky & Kane, 2007; Jones, 2012; Oberauer et al., 2018). Thus, even though this study on VT in the working memory domain failed to find any substantive evidence for broader transfer, we believe that it is worth considering whether different implementations of this principle in WM training could be more successful. For this purpose, we pointed above to several potential shortcomings that one could try to amend in future studies. The key issue would be to create favourable conditions for "learning to learn" so that common features of specific tasks are extracted and exploited for efficient adaptation in novel tasks (Braun et al., 2010). How and if this could be done in WM training remains to be seen.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41465-021-00235-2.

Acknowledgements We wish to thank Daniel Wärnå and Janne Hakala for programming the cognitive tasks. MSOffice and the R Environment "ggthemes" package (Arnold et al., 2021) were used for creating the artwork.

Author Contribution ML developed the study concept and LR, DF, JJ, OW, and JS contributed to the study design. Testing and data collection were conducted by LR with the assistance of NL and RN. Data analysis and interpretation were conducted by LR and DF. LR, JS, and ML drafted the manuscript. All authors approved the final version of the manuscript for submission.

Funding Open access funding provided by Abo Akademi University (ABO). This work was financially supported by the Academy of Finland (grant no 323251 to ML), Rehabilitation Foundation Peurunka (LR), The Folkhälsan Foundation - Professor Jan-Magnus Jansson's Fund (LR), TOP Foundation (grant no 20200699 to LR), and The Finnish Concordia Fund (grant no 20210058 to LR).

Availability of Data and Material The datasets analysed during the current study are available in the Open Science Framework repository, https://osf.io/jbeac.

Code Availability Not applicable.

Declarations

Ethics Approval The study was approved by the Ethics Committee of the Departments of Psychology and Logopedics, Åbo Akademi University, and conducted in accordance with the Helsinki Declaration.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Conflict of Interest The authors declare no competing interests.

References

Arnold, J. B., Daroczi, G., Werth, B., Weitzner, B., Kunst, J., Auguie, B., Rudis, B., Wickham, H., Talbot, J., & London, J. (2021). ggthemes: extra themes, scales and geoms for 'ggplot2'. Retrieved from https://cran.r-project.org/web/packages/ggthemes/index.html.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*(4), 445–454. https://doi.org/10.1177/1745691613491271

Braun, D. A., Aertsen, A., Wolpert, D. M., & Mehring, C. (2009). Motor task variation induces structural learning. *Current Biology, 19*(4), 352–357. https://doi.org/10.1016/j.cub.2009.01.036

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research, 206*(2), 157–165. https://doi.org/10.1016/j.bbr.2009.08.031

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*, 193–199. https://doi.org/10.3758/PBR.17.2.193

Chein, J. M., & Schneider, W. (2012). The brain's learning and control architecture. *Current Directions in Psychological Science, 21*(2), 78–84. https://doi.org/10.1177/0963721411434977

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *Quarterly Journal of Experimental Psychology, 60*(9), 1227–1245. https://doi.org/10.1080/17470210600926075

Fellman, D., Jylkkä, J., Waris, O., Soveri, A., Ritakallio, L., Haga, S., Salmi, J., Nyman, T. J., & Laine, M. (2020a). The role of strategy use in working memory training outcomes. *Journal of Memory and Language, 110*, 104064. https://doi.org/10.1016/j.jml.2019.104064

Fellman, D., Salmi, J., Ritakallio, L., Ellfolk, U., Rinne, J. O., & Laine, M. (2020b). Training working memory updating in Parkinson's disease: A randomised controlled trial. *Neuropsychological Rehabilitation, 30*(4), 673–708. https://doi.org/10.1080/09602011.2018.1489860.

Forsberg, A., Fellman, D., Laine, M., Johnson, W., & Logie, R. H. (2020). Strategy mediation in working memory training in younger and older adults. *Quarterly Journal of Experimental Psychology, 73*(8), 1206–1226. https://doi.org/10.1177/1747021820915107

Gathercole, S., Dunning, D., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language, 105*, 19–42. https://doi.org/10.1016/j.jml.2018.10.003

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., Bingel, U., Chein, J. M., Colzato, L. S., Edwards, J. D., Facoetti, A., Gazzaley, A., Gathercole, S. E., Ghisletta, P., Gori, S., Granic, I., Hillman, C. H., Hommer, B., …, & Witt, C. M. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement, 3*, 2–29. https://doi.org/10.1007/s41465-018-0115-y

Holmes, J., Woolgar, F., Hampshire, A., & Gathercole, S. E. (2019). Are working memory training effects paradigm-specific? *Frontiers in Psychology, 10*, 1103. https://doi.org/10.3389/fpsyg.2019.01103

Jefferies, E., Lambon Ralph, M. A., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of Memory and Language, 51*(4), 623–643. https://doi.org/10.1016/j.jml.2004.07.005

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology, 3*, 167. https://doi.org/10.3389/fpsyg.2012.00167

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*(2), 169–183. https://doi.org/10.1037/0096-3445.130.2.169

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science, 12*(6), 978–990. https://doi.org/10.1111/j.1467-7687.2009.00846.x

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin, 145*(2), 165–188. https://doi.org/10.1037/bul0000180

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

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*(1), 146–154. https://doi.org/10.1111/desc.12202

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

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. *Scientific Reports, 8*(1), 4045. https://doi.org/10.1038/s41598-018-22396-5

Malinovitch, T., Jakoby, H., & Ahissar, M. (2020). Training-induced improvement in working memory tasks results from switching to efficient strategies. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01824-6

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/10.1177/1745691616635612

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Bayes factor: Computation of Bayes factors for common designs. Retrieved from https://cran.r-project.org/package=BayesFactor.

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

Murty, V. P., Sambataro, F., Radulescu, E., Altamura, M., Iudicello, J., Zoltick, B., Weinberger, D. R., Goldberg, T. E., & Mattay, V. S. (2011). Selective updating of working memory content modulates meso-cortico-striatal activity. *NeuroImage, 57*, 1264–1272. https://doi.org/10.1016/j.neuroimage.2011.05.006

Norris, D. G., Hall, J., & Gathercole, S. E. (2019). Can short-term memory be trained? *Memory and Cognition, 47*(5), 1–12. https://doi.org/10.3758/s13421-019-00901-z

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., Schweppe, 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

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

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Redick, T. S., Wiemers, E. A., & Engle, R. W. (2020). The role of proactive interference in working memory training and transfer.

Psychological Research Psychologische Forschung, 84(6), 1635–1654. https://doi.org/10.1007/s00426-019-01172-8

Reed, D. K., Zimmermann, L. M., Reeger, A. J., & Aloe, A. M. (2019). The effects of varied practice on the oral reading fluency of fourth-grade students. *Journal of School Psychology, 77*, 24–35. https://doi.org/10.1016/j.jsp.2019.10.003

Richey, J. E., Phillips, J. S., Schunn, C. D., & Schneider, W. (2014). Is the link from working memory to analogy causal? No analogy improvements following working memory training gains. *PLoS ONE, 9*(9), e106616. https://doi.org/10.1371/journal.pone.0106616

Sabah, K., Dolk, T., Meiran, N., & Dreisbach, G. (2019). When less is more: Costs and benefits of varied vs. fixed content and structure in short-term task switching training. *Psychological Research, 83*(7), 1531–1542. https://doi.org/10.1007/s00426-018-1006-7

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science, 26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Sanders, R. E., Gonzalez, D. J., Murphy, M. D., Pesta, B. J., & Bucur, B. (2002). Training content variability and the effectiveness of learning: An adult age assessment. *Aging, Neuropsychology, and Cognition, 9*(3), 157–174. https://doi.org/10.1076/anec.9.3.157.9614

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. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

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 and Review, 24*(4), 1077–1096. https://doi.org/10.3758/s13423-016-1217-0

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson/Allyn & Bacon.

Vakil, E., & Heled, E. (2016). The effect of constant versus varied training on transfer in a cognitive skill learning task: The case of the Tower of Hanoi Puzzle. *Learning and Individual Differences, 47*, 207–214. https://doi.org/10.1016/j.lindif.2016.02.009

Waris, O., Fellman, D., Jylkkä, J., & Laine, M. (2021a). Stimulus novelty, task demands, and strategy use in episodic memory. *Quarterly Journal of Experimental Psychology, 74*(5), 872–888. https://doi.org/10.1177/1747021820980301

Waris, O., Jylkkä, J., Fellman, D., & Laine, M. (2021b). Spontaneous strategy use during a working memory updating task. *Acta Psychologica, 212*, 103211. https://doi.org/10.1016/j.actpsy.2020.103211

Wechsler, D. (1997). *Wechsler Memory Scale* (3rd ed.). The Psychological Corporation.