Training of resistance to proactive interference and working memory in older adults: A randomized double-blind study

SUPPLEMENTARY MATERIAL

Sandra V. Loosli, Rosalux Falquez,

Josef M. Unterrainer, Cornelius Weiller, Benjamin Rahm, Christoph P. Kaller

Methods

S1. Participants (Additional Information on Inclusion and Exclusion Criteria)

Participants were recruited from the lab's subject database as well as with flyers at public places. To ensure that none of the participants had a neurological or psychiatric disorder including mild cognitive impairment (MCI) or dementia, several steps were undertaken:

First, participants that were recruited from the lab's subject data base had already completed a screening for dementia and a questionnaire regarding neurological or psychiatric disorders for a previous study. Only subjects who scored in the normal range in this previous dementia screening and did not indicate any neurological or psychiatric disorders were recruited for the present study. Individuals who were newly recruited for the present study initially completed the same screening and questionnaires as the other participants and only those with scores in the normal range and no indications of pre-existing disorders were included in the final sample of the present study.

In a second step, all pretest scores and all training scores of Day 1 were z-transformed and checked for possible outliers within the sample ($z <> \pm 3$ SD). As the performance of all participants was within this range, no data were excluded (see also 'Data and Outlier Analysis' in the main manuscript).

Third, pretest scores and questionnaires which had published norms (Geriatric Depression Scale; GDS, Yesavage and Sheikh, 1986; Test of Nonverbal Intelligence; TONI, Brown et al., 1997; Stroop Test, Bäumler, 1985; see description below) were compared with the published norms. Regarding the GDS, all included participants scored in the normal or low borderline range at the pretest and follow-up session (score range from 0 to 6). In the interference condition of the Stroop task, all participants performed in the normal range (all > 65^{th} percentile of their respective age group). In the naming condition of the Stroop task, which can be considered as a measure of processing speed, two individuals performed at the 10^{th} or 16^{th} percentile (borderline performance), while all other participants performed

above the 46th percentile. In the TONI, all participants performed on average level (all ≥ 26th percentile when compared to the appropriate age group).

Altogether, these examinations confirmed that the final sample represents healthy older adults.

A total of 39 subjects were enrolled for participation. After screening of the medical parameters, five participants were excluded for reasons such as history of stroke (2), meningitis and amblyacousia (1), suspicion of dementia (1), or antidepressant medication (1). Of the remaining 34 participants, five did not complete the study due to illness during the course of the training (3), exhaustion (1), or for unknown reasons (1). In addition, four participants had to be excluded after data inspection: One participant misunderstood the instructions at the beginning of the training and often answered incorrectly or not to all conditions in the training tasks, one was excluded due to technical reasons (data of the last training day were not stored), one because of too many trials with missing responses on the first training day and one due to experimenter error. No participants had to be excluded due to outliers in the training or transfer tasks.

For three participants, not all data of the transfer tasks were available: One did not finish the posttest session because of strong back pain and therefore had no data for the Stroop task and the gF task, one participant did not follow the instructions of the paired-associates task in the pretest session. For another participant, no audio records for the verb-generation task in the posttest session and for the paired-associates in all sessions were stored due to technical malfunction. However, all analyses were run with and without these participants, and as the same results were found, they were included to increase power. The final sample consisted of 25 participants (14 in the high-PI group, 11 in the low-PI group).

S2. Experimental Tasks and Design

Pre- and posttest Tasks on Near and Far Transfer

Near Transfer Tasks

Verb-generation Task. In this semantic memory task (Persson and Reuter-Lorenz, 2008; Thompson-Schill et al., 1997) 32 nouns were presented visually one at a time on a computer screen with Presentation® software (Version 12.2, www.neurobs.com), while participants had to generate a related verb as fast as possible for each noun. For each trial, participants' oral responses were audio-recorded for subsequent analysis. Participants were instructed to say only the verb and not to read out loud the noun or to comment on the noun. The time window for the response was unlimited. After an answer was given, the experimenter pressed a mouse button to proceed with the next trial. A fixation cross appeared for 1000 ms between each stimulus

There were two conditions: One half of the nouns had many appropriate associated verbs, but no clear dominant response (high interference condition, e.g., kitchen with *to cook, to eat, to clean*, etc.), the other half had only few associated verbs, but one verb was clearly dominant (low interference condition, e.g., bed with *to sleep*). RTs for the high interference condition were expected to be longer, as the response had to be chosen from different competing alternatives, which involves higher selection demands (see Thompson-Schill et al., 1997).

RTs were defined as duration between the beginning of the visual presentation of the noun and the beginning of the verbal answer and were analyzed with an in-house developed software tool written in MATLAB (The Mathworks Inc., Natick, MA). Onsets were manually determined and the verbal answers were transcribed and rated. Only verbs without any other previous verbal response were counted as correct and evaluated further. Three different stimulus sets were used for pretest, posttest, and follow-up sessions.

Paired-associates Task. This task was used to measure PI resistance within a verbal episodic memory task (see e.g., Henson et al., 2002; Persson and Reuter-Lorenz, 2008). The task consisted of two parts, i.e., a study and a recall phase. In the study phase,

different word pairs were presented sequentially for 3000 ms on a screen with Presentation® software, the cue above and its associate below a centrally presented fixation cross. Between each word pair, a fixation cross was centrally presented for 1000 ms. Four cues appeared three times with the same associate (low interference condition, e.g., three times egg - train), another four cues appeared three times with a different associate (high interference condition, e.g., table – milk, table – child, table - village). Participants had to read and to memorize all word pairs. They were told that later they would have to remember only the most recently learned associate for each cue word (e.g., for the example above, train and village had to be remembered). In the recall phase, the eight cue words were shown sequentially and the according associate had to be said as fast as possible. Participants were told to say “next” if the associate could not be remembered. The time window to respond was unlimited, and the cue word stayed on the screen until a response was given. After the response, the experimenter pressed a mouse button, then a fixation cross was presented centrally for 1000 ms followed by the next cue word. In total, there were four different runs, i.e., 24 pairs with eight cues had to be learned and recalled four times, each time with different stimuli.

Responses in the control condition were expected to be faster compared to those in the interference condition, as in the latter, PI from the first two associates may lead to a delayed response and to more errors (Henson et al., 2002). As in the verb-generation task, oral responses were audio-recorded so as to determine onsets of the verbal responses and to transcribe and evaluate the answers (see above).

Three different stimulus sets were used for pre-, post- and follow-up sessions. No proper nouns and no ambiguous words were used. No antonyms were chosen as associations and it was not possible to form a new term when combining a cue with its association (e.g., as in apple – pie). Cues and associations were not repeated within the same version of the task.

Stroop Test (Bäumler, 1985). This task was used to measure prepotent response inhibition (see Friedman and Miyake, 2004). There were three conditions: First, 72 color

names printed in black ink on a paper sheet had to be read as fast as possible (reading condition; e.g., BLUE printed in black had to be named 'blue'). Second, 72 colored lines had to be named with the appropriate color (naming condition; e.g., a red color line had to be named 'red'). Third, 72 color names printed in colored ink were used. But the meaning of the word did not correspond with the ink color, and participants had to name the color of the ink instead of reading the word (interference condition; e.g., when GREEN was printed in blue ink, the participant had to say 'blue'). The duration to complete the whole sheet was measured with a stopwatch. Interference was calculated as the time difference between the naming and the interference condition. The same test version was used for all sessions.

Far Transfer Tasks.

Test of Nonverbal Intelligence (TONI; Brown et al., 1997). This task was used to measure transfer on abstract figural problem solving, i.e., fluid intelligence or gF (Brown et al., 1997). The task was to complete abstract sets of geometric patterns with a logically matching part out of four to six specified alternatives. Participants were shown 45 problem sets on a touch-sensitive computer screen with Presentation® software. The matching pattern had to be selected with a pen for touch-screens. After the answer was marked, "continue" had to be typed on, then, the next pattern set appeared. The time window to give an answer to a single problem was unlimited. The time for completion of the whole task was restricted to 10 minutes in which the participants had to solve as many problems as possible. Before the actual task was started, five practice trials were solved in order to ensure that the task was properly understood. The dependent variable consisted of the number of correctly solved problems within the time limit. Version A was used for pretest and follow-up sessions, version B was used for the posttest session.

Digit-Symbol Substitution Test (DSST; Tewes et al., 2000, from the German version of the Wechsler Intelligence Scale for Children – 3rd Edition). This task was used to assess transfer on psychomotor speed and was conducted as a paper-pencil test. A code table with nine numbers and corresponding nine symbols was displayed on a paper

sheet. Below, a series of 119 numbers ranging from 1-9 with an empty square below were depicted in random order. The participant's task was to draw the corresponding symbols in the squares below the numbers as fast as possible. The empty squares had to be filled out in sequence. The dependent measure was the number of symbols correctly assigned within two minutes. The same test version was used for all sessions.

S3. General Procedure

Enrolled participants were matched regarding age and sex, and, if available from a previous study of our lab in which some of the participants took part (Köstering et al., 2014), WM performance (*n*-back and recent-probes task), speed, and fluid intelligence. After matching, they were assigned to the high- or low-PI training condition. The rationale for this procedure was to minimize possible a-priori group differences in cognitive functions. The study design was double-blind, i.e., experimenters were not aware of the group membership of the participants, and participants neither knew that there were two different training conditions nor which group they belonged to. To realize this, one person of the study team (CPK) conducted the matching and the randomized assignment of the participants to the two training conditions, while two other members of the team (SVL, RF) conducted the training and the testing, but did not know to which group the participants belonged to.

The training was conducted in a quiet computer lab. There were ten training days (two weeks, Mondays to Fridays). In total, the whole training program was completed with four different groups whose training took place for practical reasons at different day times (e.g., in the morning or in the late afternoon). These groups included between four and thirteen participants each, equally distributed across high- and low-PI conditions. On the first day of the training, before participants started with training on the computerized tasks, a short introduction into memory functions was given and the general purpose of the training ("memory training", "improvement of cognitive functions") was explained. Then, the tasks were introduced and the training began.

On Day 1 and Day 10, all participants accomplished a high-PI version of the tasks, as these days served as test sessions for the training tasks. On Days 2 to 9, participants received either the high- or the low-PI versions, dependent on their group membership. The very first run of each task on Day 1 was regarded as practice and not analyzed further. Likewise, for Day 10, also the first run of each task was discarded to control for possible fatigue effects when comparing the effects of Day 10 with those of Day 1.

On each training day, participants first performed one run of the recent-probes task, then one run of the *n*-back task. Subsequently, a break of about 10 minutes followed, in which participants could walk around, talk, and drink something. Then, a second run of both tasks followed. Eye-relaxation exercises were used to rest for two to three minutes between the runs of both tasks. Task difficulty (e.g., WM load and amount of PI) was constant during all runs. Individual task sets were used for training on Days 2-9, while all participants received identical test sets on Days 1 and 10. At the beginning of each training day, participants had to type their birthday in a dialogue box to start their personalized sequence of task sets.

During all training sessions, two experimenters were present to help in case that any questions emerged. The total amount of time spent on the training tasks was about 32 minutes per day. However, including introduction and breaks, the overall intervention had a daily duration of about 60 minutes.

Pretest sessions were conducted on the weekend before the training started; posttests on the weekend directly after the second week of training. The follow-up session, which tested for possible long-term transfer, was conducted eight to nine weeks after the posttest session (M = 8.51 weeks, SD = 0.29, range = 8.14 – 9.29 weeks).

S4. Outlier Analysis in the Transfer Tasks

For transfer tasks with RTs as main dependent variable (verb generation and paired associates), trials with RT's $z > 3$ or < -3 were excluded from further analyses and the RTs of the remaining trials were median-aggregated. In the verb-generation task, only responses

with verbs were analyzed. In the paired associates, all usable responses were analyzed. In the verb-generation task, 11.4%, 18.1%, and 17.3% of all trials were excluded for the pretest, posttest, and follow-up sessions, respectively, due to non-analyzable data (e.g., participant spoke before generating a verb) or due to too slow responses. Likewise, in the paired associates, 21.6%, 26.2%, and 22.9% of trials were non-analyzable for the respective sessions. Statistical analyses with the data from the paired-associates test were also run using only correct trials, however the same effects resulted.

Results

S5. Immediate Transfer Effects

Performance in the transfer tasks at pretest, posttest and follow-up sessions for both groups is reported in Table S1 and displayed in Figure 4 in the main manuscript. Separate repeated-measurements ANOVAs with training group (high- vs. low-PI) as between-subjects factor and time (pretest vs. posttest) as within-subject factor were conducted for the three near-transfer tasks as well as for the two far-transfer tasks.

Near Transfer.

Verb Generation. The analysis revealed no main effect of time, $F(1, 22) = 0.09$, $p = .763$, $\eta_p^2 < .01$. However, there was a main effect of training group, $F(1, 22) = 9.18$, $p < .01$, $\eta_p^2 = .30$. The high-PI-training group had smaller interference effects than the low-PI group, which was – on a purely descriptive level – more pronounced at the post- than at the pretest session (see also Fig. 4), but the Time × Group interaction was not significant, $F(1, 22) = .97$, $p = .335$, $\eta_p^2 = .04$.

Paired Associates. There were no significant effects (main effect of time: $F(1, 21) = 1.27$, $p = .273$, $\eta_p^2 = .06$; main effect of group: $F(1, 21) = .44$, $p = .515$, $\eta_p^2 = .02$; Time × Group interaction: $F(1, 21) = .32$, $p = .580$, $\eta_p^2 = .02$).

Stroop. The analysis revealed no main effect of time, $F(1, 22) = 1.10$, $p = .306$, $\eta_p^2 = .05$, again indicating that interference did not decrease from pre- to posttest. There was no main effect of training group, $F(1, 22) = 2.14$, $p = .158$, $\eta_p^2 = .09$, also the interaction between time and group was not significant, $F(1, 22) = .06$, $p = .803$, $\eta_p^2 < .01$.

In sum, there was no improvement regarding interference scores in the near-transfer tasks, neither globally for both groups nor a differential improvement in one of the groups.

Far Transfer.

TONI. The analysis revealed a significant main effect of time, $F(1, 22) = 13.19$, $p < .01$, $\eta_p^2 = .38$, showing that there was a general improvement regarding gF scores (see Fig. 4). There was no main effect of group, $F(1, 22) = 1.30$, $p = .266$, $\eta_p^2 = .06$ and also no interaction between time and group, $F(1, 22) = .11$, $p = .746$, $\eta_p^2 = .01$, indicating that both groups improved similarly across time.

DSST. There was no transfer regarding speed, neither a general improvement, $F(1, 23) = 1.53$, $p = .229$, $\eta_p^2 = .06$, nor a differential improvement by one of the groups, $F(1, 23) = 0.03$, $p = .870$, $\eta_p^2 < .01$, and there was also no general group difference, $F(1, 23) = .17$, $p = .684$, $\eta_p^2 = .01$.

As for the near-transfer tasks, there were no differential effects in the far-transfer tasks. However, results revealed a global improvement in gF from pre- to posttest in both groups that was hence independent of the training on PI resistance.

S6. Long-term Transfer Effects

To investigate long-term transfer, further repeated-measurement ANOVAs were run to compare both experimental groups (High- vs. Low-PI; between-subjects factor group) regarding their performance in the pretest and the follow-up session (within-subject factor time) (see Table S1 and Fig. 4).

Near Transfer.

Verb Generation. The analysis revealed no significant effects (main effect of time: $F(1, 23) = .04$, $p = .847$, $\eta_p^2 < .01$; main effect of group: $F(1, 23) = .42$, $p = .521$, $\eta_p^2 = .02$; Time × Group interaction: $F(1, 23) = .01$, $p = .943$, $\eta_p^2 < .01$).

Paired Associates. There were no significant effects (main effect of time: $F(1, 21) = .64$, $p = .432$, $\eta_p^2 = .03$; main effect of group: $F(1, 21) = .09$, $p = .766$, $\eta_p^2 < .01$; Time × Group: $F(1, 21) = .02$, $p = .894$, $\eta_p^2 < .01$).

Stroop. The analysis revealed no significant effects (main effect of time: $F(1, 23) = .01$, $p = .927$, $\eta_p^2 < .01$; main effect of group: $F(1, 23) = 2.68$, $p = .116$, $\eta_p^2 = .10$; Time × Group: $F(1, 23) = .19$, $p = .669$, $\eta_p^2 = .01$).

In sum, there was no improvement in the near-transfer tasks from pretest to follow-up, not even main effects of time that would indicate a general reduction in interference across time.

Far Transfer.

TONI. The analysis revealed no main effect of time (although approaching a trend, $F(1, 23) = 2.84$, $p = .105$, $\eta_p^2 = .11$), no main effect of group, $F(1, 23) = .46$, $p = .506$, $\eta_p^2 = .02$, and no interaction, $F(1, 23) = .04$, $p = .841$, $\eta_p^2 < .01$. Scores first increased from pre- to posttest, then decreased from posttest to follow-up, however, were still higher than at the initial assessment (see Table S1 and Fig. 4).

DSST. There was a marginal main effect of time, $F(1, 23) = 3.99$, $p = .058$, $\eta_p^2 = .15$, indicating that there was a performance increment from pretest to follow-up. However, as in the other tasks, there was no group difference, $F(1, 23) = .08$, $p = .786$, $\eta_p^2 < .01$ and the interaction between group and time was also non-significant, $F(1, 23) = .09$, $p = .770$, $\eta_p^2 < .01$. In general, scores increased from the first to the second measurement, then again increased to the follow-up measurement (see Table S1 and Fig. 4)

Thus, despite marginal main effects of time in both far transfer tasks, indicating an increment from pretest to follow-up assessment, there were no differential effects regarding

these improvements. This is also reflected in the comparable effect sizes in both groups for changes between pretest and follow-up (see Table S1).

Table S1
Descriptives and Effect Sizes for Near and Far Transfer Measures

Task	Group	Pretest		Posttest		Follow-up		Cohen's *d* pre - post	Cohen's *d* pre – follow-up
		Mean	***SD***	***Mean***	***SD***	***Mean***	***SD***		
Verb Generation[a]									
	High-PI	304	166	216	96	311	340	0.63	-0.03
	Low-PI	354	279	402	210	370	174	-0.18	-0.07
Paired Associates[a]									
	High-PI	407	336	275	413	350	158	0.34	0.21
	Low-PI	442	337	372	230	372	232	0.23	0.23
Stroop[a]									
	High-PI	27.93	13.32	26.62	10.48	28.79	7.06	0.11	0.07
	Low-PI	35.39	13.57	32.25	9.56	34.07	11.42	0.26	-0.11
TONI[b]									
	High-PI	26.29	4.21	29.23	3.17	27.43	4.48	0.77	0.25
	Low-PI	25.00	5.24	27.18	4.29	26.45	4.41	0.44	0.29
DSST[b]									
	High-PI	62.14	13.95	64.29	14.55	64.57	16.18	0.15	0.16
	Low-PI	60.18	15.02	61.82	11.29	63.45	11.18	0.12	0.24

Note: SD = Standard Deviation; TONI = Test of Nonverbal Intelligence; DSST = Digit Symbol Substitution Test. [a] Interference score of reaction times in milliseconds (verb generation, paired associates), or in seconds (Stroop); [b] Number of correct responses.

S7. Power Analysis

Table S2
Required Sample Sizes to Obtain Significant Group x Time Interaction Effects

	Measure	η_p^2 of interaction effect	Correlation T1 – T2	Required *N*
Training	RP PI Accuracy	0.018	-.023	222
	NB PI Accuracy	0.017	.659	80
	RP PI RT-costs	0.012	.342	216
	NB PI RT-costs	0.020	.121	172
Transfer pre-post	Verb Generation[a]	0.042	-.430	130
	Paired Associates[a]	0.015	.405	156
	Stroop[a]	0.003	.571	562
	TONI[b]	0.010	.739	104
	DSST[b]	0.001	.854	576

Note: Required sample sizes were calculated using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). α was set to .05, Power was set to .80. T1 = First day of training or pretest; T2 = last day of training or posttest; RP = recent-probes; NB = *N*-back; PI = proactive interference; RT = reaction-times; TONI = Test of Nonverbal Intelligence; DSST = Digit Symbol Substitution Test. [a] RT interference score, [b] Number of correct responses.

Supplementary References

Bäumler, G. (1985). *Farbe-Wort-Interferenztest (FWIT) nach J. R. Stroop* [The Stroop-test]. Göttingen: Hogrefe.

Brown, L., Sherbenou, R. J. and Johnsen, S. K. (1997). *TONI-3: Test of Nonverbal Intelligence* (3rd ed.). Austin, TX: Pro-Ed.

Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39, 175-191. doi:10.3758/BF03193146

Friedman, N. P. and Miyake, A. (2004). The relations among inhibition and interference control functions: A latent variable analysis. *Journal of Experimental Psychology: General,* 133, 101-135. doi:10.1037/0096-3445.133.1.101

Henson, R. N. A., Shallice, T., Josephs, O. and Dolan, R. J. (2002). Functional magnetic resonance imaging of proactive interference during spoken cued recall. *Neuroimage,* 17*,* 543-558. doi:10.1006/nimg.2002.1229

Köstering, L., Stahl, C., Leonhart, R., Weiller, C. and Kaller, C. P. (2014). Development of planning abilities in normal aging: Differential effects of specific cognitive demands. *Developmental Psychology,* 50, 293-303. doi:10.1037/a0032467

Persson, J. and Reuter-Lorenz, P. A. (2008). Gaining control: Training executive function and far transfer of the ability to the ability to resolve interference [retracted]. *Psychological Science,* 19, 881-888. doi:10.1111/j.1467-9280.2008.02172.x

Tewes, U., Rossmann, P. and Schallberger, U. H. (2000). *HAWIK-III Hamburg-Wechsler-Intelligenztest für Kinder III* [Wechsler Intelligence Scale for Children (WISC-III; 1991)-German version]. Bern: Huber.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. and Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America,* 94*,* 14792-14797. doi:10.1073/pnas.94.26.14792

Yesavage, J. A. and Sheikh, J. I. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist*, 5, 165–172. doi:10.1300/J018v05n01_09