No evidence that foreign language learning in older age improves cognitive ability: A randomized controlled study

**Rasmus Berggren¹, Jonna Nilsson¹, Yvonne Brehmer^{1,2}, Florian Schmiedek³, &
Martin Lövdén¹**

*¹Aging Research Center, Karolinska Institutet and Stockholm University, Stockholm,
Sweden*

²Department of Psychology, Gothenburg University, Gothenburg, Sweden.

*³Department for Education and Human Development, German Institute for International
Educational Research (DIPF), Frankfurt am Main, Germany*

Correspondence about this article should be sent to Rasmus Berggren
(rasmus.berggren@ki.se) or Martin Lövdén (martin.lovdén@ki.se), Aging Research Center,
Tomtebodavägen 18A, 171 77 Stockholm, Sweden.

Abstract

Foreign language learning in older age has been proposed as a promising avenue for combatting age-related cognitive decline. We tested this hypothesis in a randomized controlled study in a sample of 160 healthy older participants (aged 65-75 years) who were randomized to 11 weeks of either language learning or relaxation training. Participants in the language learning condition obtained basic knowledge in Italian, but between-group differences in improvements on latent factors of verbal intelligence, spatial intelligence, working memory, item memory, or associative memory were negligible. We argue that this is not due to either poor measurement, low course intensity, or low statistical power, but that language learning in older age is likely to have no or trivially small effects on cognitive abilities. We place this in the context of the cognitive training literature and conclude that while skill acquisition expands the behavioral repertoire it does little to improve cognitive processing abilities.

Introduction

Learning a language challenges many cognitive processes. It requires the encoding, storage, and retrieval of arbitrary and multimodal relations between words and their meaning, as well as a demanding concurrent process of interpreting syntactic and semantic information as the spoken language unfolds (Ullman, 2004; Davis & Gaskell, 2009). Language learning thus places high demands on cognitive processes that are negatively affected in aging, such as associative memory and working memory processes (e.g., Rönnlund et al., 2005; Shaie, 1994). With the underlying hypothesis that engagement in demanding cognitive processes can lead to improvements in the efficiency of these processes themselves, and not only to the mere acquisition of task-specific skills and knowledge, some researchers have therefore proposed that learning a foreign language in older age is a promising way to battle cognitive decline (Antoniou, Gunasekera, & Wong, 2013).

Studies investigating the effects of language learning on cognitive performance in older adults are however few, show mixed results, and are plagued with important methodological limitations, such as small sample sizes, non-random assignment to experimental groups, lack of active control groups, and suboptimal statistical analyses (Antoniou & Wright, 2017). We therefore launched a randomized controlled study, allocating older adults to either participate in an Italian course ($n = 90$) or a relaxation course ($n = 70$), to test the hypothesis that foreign language learning in older age results in larger improvements of cognitive ability than participation in an active control condition, such as relaxation training.

Methods

Participants

Healthy adults aged 65 to 75 years were recruited through ads in a local newspaper. To be eligible, participants had to be in good health, native Swedish speakers, and have no prior knowledge of any of the Romance languages (see SOM1 for full list of inclusion and exclusion criteria). A total of 169 participants met eligibility criteria for study inclusion and entered the study after providing informed consent. The study was approved by the ethical review board in Stockholm (case number 2015/2284-31/2) and conducted in accordance with the Declaration of Helsinki. Participants were randomly allocated to either a language (Italian) learning or a relaxation training course in a 6:5 ratio, stratifying on sex, age, and word-word associative memory performance at pretest.

For obvious reasons, participants were not blind to treatment condition during the study, but were uninformed of the study hypothesis. The data analyst and data collector (RB) were aware of treatment condition for all participants, as well as the study hypothesis.

Five participants in the language-learning group and four in the relaxation group ended their participation early. Thus, 160 participants (94.7%) completed the study. The effects of dropout were overall small; standardized mean difference (SMD) between full sample (including dropouts) and final sample (excluding dropouts) at pre-test range from -0.04 to 0.07 across the cognitive measures.

The final sample consisted of 90 participants in the language learning group (mean (SD) age = 69.2 (2.7) years, 52 females, mean (SD) education = 16.0 (3.0) years) and 70 participants in the relaxation group (mean (SD) age: 69.5 (2.8) years, 48 females, mean (SD) education: 15.4 (3.5) years).

At the start of the study, we aimed for a sample size of around 100 in the language learning group and 80 in the relaxation group. The aim of a larger sample size for the

language learning group was based on an interest in additionally modelling individual difference in this group. Although we planned to analyze the data with structural equation modelling, we performed simplified power calculations for a mixed analysis of variance (between-within interaction) using G*Power 3.1 (Faul et al., 2007) to get rough guiding estimates for deciding on our sample size: At an $\alpha = .05$ and assuming a correlation among repeated measures of .5, 200 participants are required to detect a group (2) X time (2) interaction of a small ($d = 0.2/f = 0.1$) magnitude with a power of 0.80. We considered effect sizes lower than this to be mostly relevant at the public health level and not of interest in the present work. Being confident that the correlations among repeated measures would be higher than .5 for most measures (observed pre-post correlations ranged between $r = .48$ to $r = .79$), we were satisfied with aiming for 80 in the relaxation group and with a slightly lower sample after dropout. With a final sample of 160 participants, the statistical power to detect an interaction of $d = 0.20$ at $\alpha = 0.05$, assuming a correlation among repeated measures of .5, is 71%, and with a correlation of .6, it is .80.

Procedures

The language and the relaxation training courses were designed for the purpose of this study. Both courses lasted for eleven weeks and were administered by an adult education center in central Stockholm that supplied facilities, course materials, and licensed teachers. Participants and the course teachers were not informed of the research hypotheses but were aware of the existence of the two treatment conditions.

Participants in the language learning condition met twice per week over eleven weeks for 2.5 hours each class (for comparison, a standard daytime class for older adults at this center comprise about 2.5 hours per week over eight weeks). During classes, participants focused on verbal communication exercises while following a course book with a pace of

approximately one chapter per week. In addition, each week participants were given lists of new Italian words to be learned, with the aim to facilitate vocabulary learning between sessions. The number of words on the list varied, with an average of 44 (SD = 11) words per week. The total of words for the course period was 485. The first session each week started with a pen-and-paper glossary test on the preceding week's words, in which the participants were presented with Swedish words and were instructed to fill in the Italian translation. In total, the course was almost three times as time consuming as an ordinary beginners course and additionally also included vocabulary learning in-between sessions.

Participants in the relaxation condition met once per week for one hour each class. During classes, participants laid on a yoga mat and focused on breathing techniques and relaxation exercises. No cardiovascular training or mindfulness exercises were conducted.

Cognitive performance was assessed two weeks before (pretest) and one week after the intervention (posttest). Participants were tested in groups of eight or less. Written and oral instructions as well as a practice run preceded each test. The cognitive test battery was completed over two consecutive days, where each testing session lasted for three hours, including breaks. The battery consisted of two tests of spatial intelligence (Raven's matrices (Raven, 1960) and WASI-II Matrix Task (Wechsler, 1999)), three tests of verbal intelligence (Analogies, Syllogisms, and Verbal Inference; Ekstrom et al., 1976), two tests of working memory (Numerical updating, n-back), and three tests of long-term associative memory and item memory using different types of stimuli (word-word, face-name, picture-picture). The testing battery was identical at pre- and post-test. Detailed descriptions of the tasks used can be found in SOM2. Pre-post correlations for the measures ranged between .48 and .79, indicating acceptable lower bounds for reliability. Detailed psychometric properties of all cognitive tests can be found in SOM3.

At the end of the language course, participants in this condition received a vocabulary test consisting of 110 words randomly sampled from the chapters completed during the course. Participants were presented with the Swedish word and was asked to fill in the Italian translation and were given two points per correct word (correct spelling), one point for any minor error (e.g., correct word but spelling error) and zero for major error (e.g., incorrect word) or absent answer, resulting in a maximum score of 220 points.

Statistical analyses

Outliers were determined using the outlier labelling rule (Hoaglin, 1986) with a g of 2.8. A total of four individual data points were deleted and treated as missing values.

Latent change score models. We used structural equation modelling to test the hypothesis of differential gain in performance for the two experimental groups. We fit one latent change score model (McArdle & Nesselroade, 1994) for each of the five cognitive domains: Associative Memory, Item Memory, Spatial Intelligence, Verbal Intelligence, and Working Memory. Ability factors, representing the shared variance among multiple tests measuring the ability, were formed. One such factor (PRE) was formed from the observed variables at pretest and another factor (POST) was formed by the corresponding variables at posttest. The latent constructs of Associative Memory, Item Memory, and Verbal Intelligence were measured by three indicators at both pre- and posttest, while Spatial Intelligence and Working Memory were measured by two indicators. The factor loading and intercept of the first indicator variable was constrained to unity and zero, respectively, for identifiability purposes. A latent change score (Δ), representing the difference between pre- and post-test, was estimated. This allowed for change to be estimated as a latent variable, attenuating influences of measurement error and reducing task-specific variance in favor of task-general (ability) variance. Specifically, the variance of POST was constrained to zero, POST was

perfectly regressed onto PRE and Δ (setting the respective model parameters to one), and a covariance was specified between PRE and Δ . This specification captures the essential components of the latent change score model (Kievit et al., 2017) and Δ captures the latent change between PRE and POST. See SOM4 for a graphical illustration of the latent change score model.

We set these models up in a multi-group framework, where the models were estimated for the experiment group ($n = 90$) and the control group ($n = 70$) simultaneously. The parameter of primary interest is the mean latent change between PRE and POST, μ_{Δ} . To test the hypothesis that there is differential change between the groups, we compared a model where μ_{Δ} is equality-constrained across groups, to a model where μ_{Δ} is freely estimated in each group. If the equality-constrained model fits significantly worse than the freely estimated model, this would be evidence of differential change between the language learning and relaxation training groups. The threshold for statistical significance was $\alpha = .05$ for all tests.

Measurement invariance was evaluated in a multi-group longitudinal framework (Kievit et al., 2017). We tested for weak (“metric”), strong (“scalar”), and strict (“residual”) measurement invariance by imposing increasingly stricter constraints on model parameters (i.e., constrained factor loadings, item intercepts, and residual variances, respectively). Two constructs (Spatial intelligence and Working memory) had only two associated manifest variables, which leads to identifiability issues for the configural and weak models. Therefore, we began testing these models at the level of strong invariance. Baseline model fit was assessed by χ^2 test, CFI, and RMSEA. Nested model comparisons were assessed using the likelihood ratio χ^2 test.

All models displayed adequate fit at the strict measurement invariance level, both in terms of absolute model fit (CFI range: 0.974 – 1.000; RMSEA range: .001 - 0.065, all $\chi^2 p > .05$) and in terms of model-constrained nested comparisons for weak, strong, and strict

invariance (all $\Delta\chi^2 p > .05$; see SOM5 for details). Therefore, we proceed to test the hypothesis of differential change on the level of strict measurement invariance.

Bayesian linear mixed models. In addition to the latent change score models, Bayesian linear mixed models were estimated for each manifest variable. Dependent variables were standardized (to the pre-test SD) to have a mean of zero and a variance of one. The model included between-subject predictor Group (Control = 0, Experiment = 1), within-subject predictor Time (Pre = 0, Post = 1), and the Group x Time interaction (which we call θ in the following) as fixed effects, and random intercepts. Thus, θ captures the standardized differential mean pre-post change between treatment groups, d . We used weakly-informative Normal (0,1) priors for all fixed effects.

We compute one-sided Bayes factors (BFs) to assess the study hypothesis that language learning had a marginal net positive effect on cognition. BFs were approximated using $\theta \sim \text{half-normal}(0, 0.2)$ informed priors and $\theta \sim \text{half-normal}(0, 0.707)$ reference priors, combined with the likelihood of the uninformed model.

Statistical programs. All statistical calculations were performed in R (R Core Team, 2017). Latent variable models were estimated using the lavaan package (Rosseel, 2012) using full information maximum likelihood. Bayesian linear mixed models were estimated using the brms package (Bürkner, 2017).

Results

The scores on the post-study vocabulary test were visually inspected and deemed to be approximately normally distributed around a mean of 115.6 ($SD = 42.0$; $range = 6-214$), suggesting that participants in the language learning group did indeed acquire a basic Italian vocabulary.

Results from the hypotheses testing with the latent change score model showed that constraining the mean latent change (μ_{Δ}) to equality across groups did not lead to a significant

decrease in model fit for any of the five abilities; $\Delta\chi^2$'s range from 0.04 - 2.56. $\Delta\chi^2 p$ range: .110 - .841 (for detailed information see SOM6). For plotting purposes, we calculated unit-weighted Z-scores for each cognitive domain by calculating the standardized change for each manifest variable and averaging them across each cognitive domain. The distributions of unit-weighted Z-scores are presented in Figure 1.

Results from the Bayesian linear mixed model on each of the individual cognitive tests showed that the net positive effect d [95% HDI; BF] ranges from $d = -0.24$ [95% HDI: -0.51 - 0.03; $BF_{01} = 4.12$] for WASI to $d = 0.18$ [95% HDI: -0.09 - 0.46; $BF_{01} = 0.57$] for Syllogisms. Bayesian triplots are available in SOM7.

Table 1. Posterior standardized differential mean change, SD, 95% credibility interval (CI) and Bayes factors. Estimates obtained using uninformative normal(0,1) priors. Bayes factors were calculated using half-normal(0, σ) priors with $\sigma = 0.2$ (informed prior) and $\sigma = 0.707$ (reference prior).

Domain	Test	Posterior Mean	Posterior SD	95% CI	BF ₀₁ half-normal(0, σ)	
					$\sigma = 0.2$	$\sigma = 0.707$
Associative memory	Face-Name	0.02	0.15	-0.27 – 0.32	1.50	4.22
	Picture-Picture	0.04	0.13	-0.21 – 0.30	1.45	4.18
	Word-word	-0.03	0.12	-0.26 – 0.21	2.24	7.02
Item memory	Face-Name	-0.17	0.14	-0.45 – 0.11	3.23	10.48
	Picture-Picture	-0.02	0.16	-0.33 – 0.29	1.76	5.09
	Word-word	-0.13	0.13	-0.40 – 0.14	2.95	9.48
Working memory	N-Back	0.10	0.16	-0.22 – 0.41	1.06	2.64
	Numerical updating	-0.05	0.10	-0.26 – 0.15	2.94	9.63
Verbal intelligence	Analogies	0.02	0.11	-0.19 – 0.23	1.79	5.50
	Syllogism	0.18	0.14	-0.09 – 0.46	0.57	1.26
	Verbal inference	-0.06	0.11	-0.27 – 0.14	3.15	10.37
Spatial intelligence	Raven's	0.05	0.15	-0.24 – 0.34	1.32	3.64
	WASI	-0.24	0.14	-0.51 – 0.03	4.12	13.73

Figure 1. Distribution of unit-weighted change scores. Dots and line segments denote the median and 1st and 3rd quartiles.

Discussion

Language learning has been suggested as a promising intervention for improving cognitive performance in older age (Antoniou, Gunasekera & Wong, 2013; Antoniou & Wright, 2017). We tested this hypothesis in a randomized controlled study with reasonable power to detect relatively small effects. Latent modelling provided no evidence that language learning in older age improves cognitive abilities relative to relaxation training.

Bayes factors using informed ($d = 0.2$) priors generally support the null hypothesis of no positive net benefit of language learning (BFs range from $BF_{01} = 1.06$ for N-back, to BF_{01}

= 4.12 for WASI), with the exception of Syllogism showing weak evidence in favor of a positive net effect ($BF_{01} = 0.57 = BF_{10} = 1.75$). Evidence in favor of the null is even stronger when considering the reference prior ($d = 0.707$; Table 1). Together with the differential change estimated (SMD [95% HDI] range from -0.24 [-0.51 – 0.03] to .18 [-0.08 – 0.46]), we find no support that language learning in older age improves cognitive functioning to any substantial extent. The observed pre-post change in both groups is on the same magnitude as what have been reported in meta-analysis of cognitive training in older adults using no-contact passive control groups (e.g. Au et al., 2015) suggesting that the observed pre-post changes are merely re-test effects.

We argue that these findings are not due to poor measurement, low statistical power, or low intervention dosage. Cognitive abilities were statistically represented as factors of multiple cognitive tasks which exhibited good psychometric properties. We sampled cognitive abilities broadly, including abilities that are likely to be directly involved in language learning (i.e., Associative Memory and Working Memory) as well as abilities less likely to be central in language acquisition (i.e., Verbal Intelligence, Spatial Intelligence, and Item Memory). Statistical power was appropriate to detect quite small effects due to the generous sample size. This is also clear from the good precision of our estimates, which allows us to conclude that our language intervention is unlikely to have any substantial effect on cognitive performance. Further, a vocabulary test administered at the end of the intervention confirmed that participants in the language learning course had acquired a basic Italian vocabulary. The extent of the language training was more intensive than many previous intervention studies in older adults, comprising up to 55 hours of teacher-led classes (three times the typical beginner's course) and learning up to 485 new words between classes. Although we cannot rule out that extending the scope and length of the intervention would result in improved cognitive performance, we conclude that taking a quite intensive

beginner's course in a foreign language in older age is unlikely to improve other cognitive performance than language proficiency in the trained language.

The present results align with a general theme of limited generalization of benefits of cognitive training interventions when the measures of cognitive performance are not tapping into learning-specific skills, strategies, and knowledge (Melby-Lervåg, Redick, & Hulme, 2016; Sala & Gobet, 2017; Simons et al., 2016). It may be that future results will show that cognitive interventions are helpful for some older adults, such as those that are cognitively under-stimulated, akin to selective benefits of physical exercise intervention for less physically active individuals. We conclude that while skill acquisition expands the behavioral repertoire of humans, it rarely affects cognitive processing ability.

Author contributions

ML and RB developed the study concept. JN and YB contributed to study design. RB collected the data and performed the data analysis with critical input from ML and FS. RB, ML and JN drafted the manuscript, and all authors critically revised the manuscript for intellectual content. All authors approved the final version of the manuscript for submission.

Acknowledgements

This research was supported by a program grant from FORTE (2013-2277) and a “distinguished younger researcher” grant from the Swedish Research Council (446-2013-7189) to ML. We would like to thank Marie Helsing, Anders Rydström, Jakob Norgren, Joanna Lindström and William Fredborg for facilitating the completion of this study.

References

- Antoniou, M., Gunasekera, GM., & Wong, PC. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: a hypothesis for future research. *Neuroscience Biobehavioral Review* (37), 2689-2698.
- Antoniou, M., & Wright, SM. (2017). Uncovering the Mechanisms Responsible for Why Language Learning May Promote Healthy Cognitive Aging. *Front. Psychol.* 8:2217, doi: 10.3389/fpsyg.2017.02217
- Au, J., Sheehan, E, Tsai, N., Duncan, GJ., Buschkuehl, M., Jaeggi, SM. (2015) Improving fluid intelligence with training on working memory: a meta-analysis. *Psycho Bull Rev*, 22, 366-377.
- Bürkner, PC. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01
- Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.
- Davis, MH., & Gaskell, MG. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Phil. Trans. R. Soc. B* 364. 3773–3800, doi:10.1098/rstb.2009.0111
- Ekstrom, RB., French, JW., Harman, HH., & Dermen, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton. NJ: Educational Testing Service.
- Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. doi:10.3758/bf03193146
- Giovanni Sala, G., & Fernand Gobet, F. (2017) Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, 26(6). pp 515–520.
- Hoaglin, DC., Iglewicz, B. & Tukey, JW. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, 81(396).
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. URL <https://www.R-project.org/>.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2). 1-36. URL <http://www.jstatsoft.org/v48/i02/>.
- Kievit, R., Brandmaier, AM., Ziegler, G., van Harmalen, AL., de Mooij, SSM., Moutoussis, M., et al. (2017). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, <https://doi.org/10.1016/j.dcn.2017.11.007>
- McArdle, JJ., & Nesselroade, JR. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.). *Life-span developmental psychology: Methodological contributions*, pp 223–267. Hillsdale. NJ: Erlbaum.
- Melby-Lervåg, M., Redick, TS., & 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, 512–534. doi:10.1177/1745691616635612
- Raven, JC. (1960). Guide to the standard progressive matrices: Sets A. B. C. D and E. London. England: H. K. Lewis.
- Rönnlund, M., Bäckman, L., & Nilsson, LG. (2005). Stability, Growth, and Decline in Adult Life Span Development of Declarative Memory: Cross-Sectional and Longitudinal Data From a Population-Based Study. *Psychology and Aging*, 20(1). DOI: 10.1037/0882-7974.20.1.3

- Schaie, KW. (1994). The course of adult intellectual development. *American Psychologist*, 49(4). 304-313.
- Simons, DJ., Boot, WR., Charness, N., Gathercole, SE., Chabris, CF., Hambrick, DZ., Stine-Morrow, EA. (2016). Do "Brain-Training" Programs Work? *Psychol Sci Public Interest*, 17(3):103-186.
- Ullman, MT. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, 92, 231–270 doi:10.1016/j.cognition.2003.10.008
- Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio. TX: Psychological Corporation.

Supplementary online material**Table of contents**

SOM1	Inclusion criteria.	2
SOM2	Detailed task descriptions.	3-4
SOM3	Psychometric properties	5
SOM4	Graphical illustration of latent change score model	6
SOM5	Nested model comparison for configural, weak, strong, strict measurement invariance.	7
SOM6	Likelihood ratio χ^2 test of equality constrained Δ model vs unconstrained Δ model.	8
SOM7	Bayes factors and triplots	9

SOM1. Inclusion criteria

General inclusion criteria

- $65 \leq \text{age} \leq 75$ years at study start
- General availability during the study period (maximum of two scheduled absent classes prior to study start)
- Adequate hearing
- Normal eye-sight or, if needed, glasses or contact lenses
- No prior experience of participation in studies of cognitive function
- Swedish as mother tongue
- Monolingual
- No or minimal prior knowledge of any Romance languages (French, Spanish, Italian, Portuguese, Latin). Prior knowledge was self-assessed, where sufficient prior knowledge for exclusion was on the level of being able to ask and understand rudimentary questions (e.g. asking for directions, or asking what time it is, and then understanding the answer).
- MMSE score ≥ 25
- No color-blindness
- No use of psychotropic medication less than 6 months prior to study start
- Not sought medical care for any of the following diagnoses within the last 5 years, or received medical treatment within 1 year:
 - Attention deficit hyperactivity disorder (ADHD)
 - Bipolar disorder (BPD)
 - Major depressive disorder (MDD)
 - Generalized anxiety (GAD)
 - Panic disorder (PD)
 - Post-traumatic stress disorder (PTSD)
 - Psychotic symptoms (e.g., hallucinations, delusions)
 - Social phobia (SP)
 - Obsessive compulsive disorder (OCD)
- No prior history of diagnosed
 - Tinnitus
 - Dyslexia
 - Stroke
 - Epilepsy
 - Alzheimer's Disease
 - Dementia
 - Multiple sclerosis

For MR eligibility

- No metallic or ferromagnetic implants, including pacemaker, heart stimulator or stent
- No claustrophobia
- Being right-handed
- Weight ≤ 120 kilograms
- No self-reported or diagnosed hearing impairment
- No life history of concussion, loss of consciousness (longer than 10 minutes).

SOM2. Detailed task descriptions.

Domain	Task	Task description	Dependent measure	Max score
Item-associative memory task	Word-word/ Face-name/ Picture-picture	During encoding phase 40 word/face-name/picture pairs were shown, one at a time, in the middle of the screen for 6000 ms. During item recognition phase, 40 singular words/faces/pictures were presented, one at a time, and the subject was to indicate whether the word/face/picture was present as part of a pair in the encoding phase (20 targets) or not (20 foils). During associative recognition phase, 40 word/face-name/picture pairs were presented on screen, one at a time, and the subject was to indicate whether the pair was an already-seen pair from the encoding phase (20 targets) or a new pair using re-arranged items from the encoding phase (20 foils).	Proportion hits – proportion false alarms	1.0
Spatial intelligence	Raven's progressive matrices	The task involved completing as many of the 18 trials within a 12-minute time limit. Each trial consisted of increasingly complex visual geometric designs with a missing piece. The task was to select the missing piece from eight alternatives.	Number correct within time limit	18
	WASI-II matrices	The task involved completing as many of the 30 trials within a 15-minute time limit. Each trial consisted of increasingly complex visuospatial objects with a missing piece. The task was to select the missing piece from five alternatives.	Number correct within time limit	30
Verbal intelligence	ETS kit verbal inference	The task involved completing as many of the 20 verbal inference trials as possible within a 12-minute time limit. The task involved selecting the valid conclusion given the information given in a verbal statement. Example trial: One year a particular farmer's stand of wheat yielded 40 bushels per acre. 1) The farmer's land is extremely fertile 2) The farmer has raised wheat on his land 3) The weather that year was unfavourable for growing wheat 4) Forty bushels per acre is a high yield 5) The field would be more suitable for some other crop	Number correct within time limit	20
	BIS analogies	The task involved completing as many of the twelve analogies within a 5-minute time limit. The task involved finding completing analogies. Example trial: Newspaper is to text as speaker is to: 1) sound 2) song 3) news 4) listening 5) stereo	Number correct within time limit	12
	Syllogisms	The task involved completing as many of the 30 syllogism trials within a 20-minute time limit. The task on each trial was to decide whether the conclusion that followed the premises was logically valid or not. Absurd examples were chosen to emphasize the formal structure of the task. Example trial: All trees are fish. All fish are horses. Therefore all trees are horses.	Number correct within time limit	30

		1) valid conclusions 2) invalid conclusions		
Working memory	N-back	The test consisted of four runs of single 2-back and four runs of single 3-back. The stimuli set consisted of shapes, which was used during training. 192 stimuli were shown, of which 64 were targets (33%). Completion time was approximately 15 minutes.	(Number hits / total targets) + (number correct rejections/ total foils)	1.0
	Numerical updating	Four single digits (ranging from 1 to 9) were presented simultaneously in a row of four cells situated horizontally for 3000 ms. After an inter-stimulus interval of 250 ms, a sequence of eight updating operations were presented in a second row of four cells below the first row. These updating operations were additions and subtractions within a range -9 to +9. Each updating operation was applied to a different cell from the one a step earlier in the sequence, so that no two updating operations had to be applied to one cell in a sequence. Those updating operations had to be applied to the digits memorized from the corresponding cells above and the updated results had to be memorized. Ten trials were included in the test. Completion time was approximately 10 minutes.	Proportion correct	1.0

SOM3. Psychometric properties. Means, SDs, skewness and kurtosis for all cognitive variables at pre- and post-test.

Domain	Test	Experiment (N = 90)										Control (N = 70)									
		PRE					POST					PRE					POST				
		Mean	SD	Skewness	Kurtosis		Mean	SD	Skewness	Kurtosis	Pre-post correlation	Mean	SD	Skewness	Kurtosis		Mean	SD	Skewness	Kurtosis	Pre-post correlation
<i>Spatial intelligence</i>	Raven's	6.98	2.63	0.09	3.43		7.32	2.71	-0.12	2.46	0.55	6.26	2.65	0.11	3.09		6.45	2.73	-0.26	2.68	0.63
	WASI	20.99	2.68	-0.82	3.60		21.64	2.23	-0.15	3.00	0.57	20.42	2.70	-0.68	2.88		21.74	2.30	-0.27	2.73	0.60
<i>Verbal intelligence</i>	Analogies	6.09	2.49	0.24	2.40		6.29	2.52	0.29	2.33	0.79	5.13	2.02	0.25	2.93		5.27	2.09	0.21	2.64	0.74
	Syllogism	18.23	4.43	-0.04	3.10		19.62	3.88	0.02	3.66	0.60	17.61	3.41	0.67	3.06		18.28	3.62	-0.04	2.72	0.61
	Verbal inference	11.98	3.32	-0.36	2.10		12.56	3.68	-0.54	2.76	0.77	10.59	3.88	-0.10	2.00		11.41	3.80	-0.19	1.90	0.81
<i>Working memory</i>	N-back	0.72	0.08	-0.49	3.68		0.74	0.08	-0.14	2.51	0.53	0.71	0.08	0.05	1.90		0.72	0.08	-0.26	2.27	0.41
	Numerical Updating	0.28	0.11	0.07	2.23		0.30	0.10	0.06	2.21	0.76	0.24	0.11	0.10	2.32		0.27	0.10	0.09	2.21	0.76
<i>Item memory</i>	Face-name	0.73	0.16	-0.73	3.76		0.77	0.15	-0.90	3.77	0.53	0.70	0.14	-0.33	2.70		0.77	0.13	-0.75	3.44	0.63
	Picture-picture	0.87	0.11	-1.03	3.50		0.89	0.10	-0.77	2.58	0.50	0.86	0.11	-0.91	3.53		0.88	0.12	-1.65	6.74	0.52
	Word-word	0.67	0.18	-0.59	2.60		0.67	0.21	-0.98	4.04	0.66	0.61	0.19	-0.56	2.87		0.64	0.20	-0.89	3.24	0.68
<i>Associative memory</i>	Face-name	0.33	0.19	0.51	3.25		0.39	0.20	0.01	3.03	0.54	0.34	0.21	-0.48	2.58		0.39	0.21	-0.25	2.13	0.59
	Picture-picture	0.47	0.25	-0.19	2.37		0.55	0.25	-0.30	2.18	0.65	0.45	0.23	-0.19	2.62		0.52	0.26	-0.49	2.60	0.73
	Word-word	0.55	0.26	-0.37	2.50		0.61	0.26	-0.38	2.18	0.68	0.52	0.26	-0.40	2.39		0.59	0.24	-0.35	2.08	0.74

SOM4. Graphical illustration of a latent change score model.

Note: Three indicators (X, Y, Z) were used for verbal intelligence, associative memory and item memory, while two indicators (X, Y) were used for spatial intelligence and working memory.

SOM5. Nested model comparison for configural, weak, strong, strict measurement invariance. Models with 2 indicators (Working memory, Spatial intelligence) are unidentified at the configural and weak level of invariance. We therefore started testing at the level of strong measurement invariance.

Domain	Comparison	$\Delta\chi^2$	ΔDf	<i>p</i>
Associative memory	Configural vs weak	4.67	6	.589
	Weak vs strong	1.79	6	.938
	Strong vs Strict	8.18	9	.517
Item memory	Configural vs weak	4.20	6	.650
	Weak vs strong	11.18	6	.083
	Strong vs Strict	15.02	9	.090
Verbal intelligence	Configural vs weak	9.61	6	.142
	Weak vs strong	7.93	6	.243
	Strong vs Strict	14.30	9	.112
Working memory	Strong vs Strict	2.52	6	.866
Spatial intelligence	Strong vs Strict	3.64	6	.725

SOM6. Likelihood ratio χ^2 test of equality constrained Δ model vs unconstrained Δ model. p-values are calculated from the χ^2 distribution on 1 degree of freedom, $\chi^2(1)$, with a critical value of 3.84. Non-significant findings suggest that there is no evidence of differential pre-post change between treatment groups.

Domain	$\Delta\chi^2$	<i>p</i>
Associative memory	0.04	.841
Item memory	1.90	.168
Verbal inference	0.10	.748
Working memory	0.04	.836
Spatial intelligence	2.56	.110

15 SOM7. Bayesian triplots (prior, likelihood and posterior) and Bayes factors for
 16 informed prior, $N \sim \text{half-normal}(0, 0.2)$ and reference prior, $N \sim \text{half-normal}(0, .707)$. Bayes
 17 factors were approximated using the maximum likelihood estimate obtained from the
 18 uninformed model.

19
20