American Psychological Association

Psychology and Aging

ISSN: 0882-7974

2020, Vol. 35, No. 2, 212–219
http://dx.doi.org/10.1037/pag0000439

Foreign Language Learning in Older Age Does Not Improve Memory or Intelligence: Evidence From a Randomized Controlled Study

Rasmus Berggren and Jonna Nilsson
Karolinska Institute

Yvonne Brehmer
Tilburg University and Karolinska Institute

Florian Schmiedek
DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany

Martin Lövdén
Karolinska Institute

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 some basic knowledge in the new language (Italian), but between-groups 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 basic studies in foreign languages in older age are likely to have no or trivially small effects on cognitive abilities. We place this in the context of the cognitive training and engagement literature and conclude that while foreign language learning may expand the behavioral repertoire, it does little to improve cognitive processing abilities.

Keywords: cognitive aging, cognitive training, language learning, memory, randomized controlled study

Supplemental materials: http://dx.doi.org/10.1037/pag0000439.supp

Learning a foreign language challenges many cognitive processes. For example, it requires learning a novel mental lexicon. This process involves the encoding, storage, and retrieval of arbitrary and multimodal relations between novel words and their meanings. The cognitive demands for acquiring such relations overlap with those involved in many other associative memory tasks. Word learning therefore shares many similarities, both at the cognitive and neural level, with general declarative memory processes (Davis & Gaskell, 2009; Ullman, 2004). When studying a foreign language in a formal learning setting, relationships between foreign words and their meanings are often intentionally studied many hours per week. This setting and the cognitive processes that this task demands resemble those involved in laboratory-based practice regimes and associative-memory tasks (e.g., the encoding of word-word pairs) designed by researchers to improve memory performance (e.g., Bellander et al., 2017).

Language comprehension and production also involve a demanding concurrent cognitive task of interpreting, retrieving, and combining syntactic and semantic information as the spoken and written language unfolds (Ullman, 2004). This task requires processes that are widely considered to serve working memory, including the concurrent maintenance and selectively updating/ma-

This article was published Online First February 3, 2020.

Rasmus Berggren and Jonna Nilsson, Aging Research Center, Karolinska Institute; Yvonne Brehmer, Department of Developmental Psychology, Tilburg University, and Aging Research Center, Karolinska Institute; Florian Schmiedek, Department for Education and Human Development, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany; Martin Lövdén, Aging Research Center, Karolinska Institute.

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 Martin Lövdén. We thank Marie Helsing, Anders Rydström, Jakob Norgren, Joanna Lindström, and William Fredborg for facilitating the completion of this study. The manuscript has been presented in a poster format at the Flux Congress 2018 in Berlin, Germany in August 2018. We have also posted this as a preprint at the OSF framework: https://osf.io/8y4ga/. European data protection laws prohibit us from putting the data in the public domain, but data can be requested from the authors or from the data protection office at Karolinska Institute (dataskyddsombud@ki.se) and subsequently transferred for well-defined analysis projects that are in line with the purposes covered by the original ethical approval. This requires a data use agreement, which effectively transfers the confidentiality obligations of the institution at which the original research was conducted (Karolinska Institute) to the institution of the recipient of the data.

Correspondence concerning this article should be addressed to Rasmus Berggren or Martin Lövdén, Aging Research Center, Karolinska Institute, Tomtebodavägen 18A, 171 77 Stockholm, Sweden. E-mail: rasmus.berggren@gmail.com or martin.lovden@ki.se

nipulation of information, the mental integration of different elements of information, and the attention to relevant but not to irrelevant information (Baddeley & Hitch, 1974; Kane & Engle, 2003; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000). Indeed, both experimental studies and work on individual differences indicate that performance on working memory tasks is strongly related to performance on a wide variety of language tasks (e.g., Carretti, Borella, Cornoldi, & De Beni, 2009; Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Gathercole & Baddeley, 1993). Working memory may be particularly taxed early in the acquisition of a foreign language because comprehension and production have to be carried out with a small mental lexicon and with restricted grammar knowledge that must be intentionally retrieved and combined (Linck, Osthus, Koeth, & Bunting, 2014; Ullman, 2004).

Thus, foreign language learning places high demands on the efficiency of cognitive processes that are negatively affected in aging, such as associative memory and working memory processes (e.g., Naveh-Benjamin, 2000; Rönnlund, Nyberg, Bäckman, & Nilsson, 2005; Schaie, 1994). Some researchers have therefore proposed that learning a foreign language in older age is a promising way to battle critical aspects of cognitive decline (e.g., Antoniou, Gunasekera, & Wong, 2013). The proposed mechanisms of change that underlie these hypothesized effects of engagement in foreign language learning are similar to those behind the hypothesis that cognitive training (e.g., practicing computerized working memory tasks) may affect cognitive ability: practicing or engaging in demanding cognitive processes can lead to improvements in the efficiency of these processes themselves or to the development of cognitive skills, strategies, and knowledge that are broadly relevant for cognitive performance (Lövdén, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010; see also Stine-Morrow et al., 2014).

Whereas the effects of cognitive training on cognitive abilities, including language abilities (e.g., Carretti et al., 2009; Payne & Stine-Morrow, 2017), have been extensively investigated, there are few studies of the effects of language learning on cognitive performance in older adults. Those that have been reported show mixed results and are plagued with important methodological limitations, such as small sample sizes, nonrandom assignment to experimental groups, lack of active control groups, and suboptimal statistical analyses (see Antoniou & Wright, 2017 for review). We therefore launched a randomized controlled study, allocating older adults with Swedish as their native language to either participate in a foreign language (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 (relaxation training) that was included to control for expectancy effects. The language course emphasized both verbal communication exercises (during classes), which we assumed would be working memory demanding, and the study of new Swedish-Italian word pairs every week (at home). Before and after the 11-week intervention period, we assessed participants' performance on several cognitive abilities with multiple tests for each ability. We assessed performance on abilities that are expected to be directly involved in language learning (i.e., associative memory and working memory) as well as abilities that are probably less central in language learning (i.e., item memory, verbal intelligence, and spatial intelligence).

Method

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 (see Online Supplement 1 for full list of criteria that defined good health), cognitively unimpaired (as defined by a score on the Mine-Mental State Examination above 25), native Swedish speakers, eligible for MRI, available for the entire study period, unexperienced with participation in studies assessing cognitive functions, and have adequate hearing and vision (including color vision) and no substantial prior knowledge of any of the Romance languages. All participants had at least working knowledge in English and self-rated their prior knowledge in Italian as either "nonexistent" or "very poor." Due to logistical constraints, data were collected in four waves, with an equal proportion of individuals in each treatment group in each wave. A total of 169 participants met eligibility criteria for study inclusion and entered the study after providing informed consent (see Online Supplement 2 for a consort flowchart). The study was approved by the ethical review board in Stockholm (case 2015/2284–31/2) and conducted in accordance with the Declaration of Helsinki. After the first day of cognitive testing, 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. The allocation was 6:5 in favor of the language learning group in order to increase statistical power for investigating possible association between individual differences in vocabulary acquisition and cognitive performance in the language learning group. 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 (first author) 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 small overall; standardized mean difference between full sample (including dropouts) and final sample (excluding dropouts) at pretest ranged from −0.04 to 0.07 across the cognitive measures. The final sample consisted of 90 participants in the language learning group (M [SD] age = 69.2 [2.7] years; 52 females; M [SD] education = 16.0 [3.0] years) and 70 participants in the relaxation group (M [SD] age = 69.5 [2.8] years; 48 females; M [SD] education = 15.4 [3.5] years). Descriptive statistics for cognitive performance are reported in Table 1.

Although we planned to analyze the data with structural equation modeling, we performed simplified power calculations for a mixed analysis of variance (between-within interaction) using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) to get rough guiding estimates for deciding on our sample size: at $\alpha = .05$ and assuming a correlation among repeated measures of .5, 200 participants are required to detect a Group (2) × 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

Table 1
Descriptive Statistics for Cognitive Performance as a Function of Session and Treatment Group

			Language learning (N = 90)								Relaxation (N = 70)							
			PRE				POST				PRE				POST			
Domain	Test	Range	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	Pre-post correlation	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	Pre-post correlation
Spatial intelligence	Raven's	0–18	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	0–30	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
Verbal intelligence	Analogies	0–12	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
	Syllogisms	0–30	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	0–20	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
Working memory	*N*-back	0–1.0	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	0–1.0	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
Item memory	Face-name	0–1.0	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	0–1.0	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	0–1.0	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
Associative memory	Face-name	0–1.0	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	0–1.0	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	0–1.0	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

Note. WASI = Wechsler Abbreviated Scale of Intelligence.

correlations ranged between r = .48 and r = .79), we were satisfied with aiming for 80 in the relaxation group (to save money, while maintaining a larger sample in the language group for potential analyses of individual differences) 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 α = .05, assuming a correlation among repeated measures of .5, is 71%, and with a correlation of .6, it is .80.

Procedure

The language and the relaxation training courses were designed for the purpose of this study. Both courses lasted for 11 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 11 weeks for 2.5 hr each class (for comparison, a standard daytime class for older adults at this center comprise about 2.5 hr once per week over 8 weeks), while following a course book (Olsson & Braconi, 2005) with a pace of approximately one chapter (two to four pages) per week. Every chapter included a main text (a dialogue), which served as the basis of the verbal communication exercises that were focused on during the classes. Basic grammatical information and a list of words and their Swedish translations accompanied the main text. The dialogues during the verbal communication exercises (role play in groups and couples) involved ordinary events connected with leisure or tourism (e.g., ordering coffee at a restaurant, asking for directions). In addition, each week participants were instructed to learn a list of new Swedish-Italian word pairs associated with the current chapter 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. The standard content of the beginners' course at this center and the accompanying course book is an application of the first (A1) level of the common European framework of reference for languages. The course implemented in the context of this study was almost 3 times as time consuming as the ordinary course and included vocabulary learning in between sessions.

Participants in the relaxation condition met once per week for 1 hr each class. During classes, participants lay on a yoga mat and focused on breathing techniques and relaxation exercises. No cardiovascular training or mindfulness exercises were conducted.

Cognitive performance was assessed 2 weeks before (pretest) and 1 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 2 consecutive days, where each testing session lasted for 3 hr, 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,

French, Harman, & Dermen, 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 posttest. Detailed descriptions of the tasks used can be found in Online Supplement 3. 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 Table 1. A smaller subsample (of equal proportion in each treatment group) also took part in structural MRI, but these data are not reported here.

After the language course, immediately after the final cognitive testing session, participants in the language condition received a vocabulary test consisting of 110 words randomly sampled from the chapters completed during the course. Participants were not informed of the final vocabulary test prior to the final testing session. Participants were presented with the Swedish word and were asked to fill in the Italian translation. Participants were given 2 points per correct word (correct spelling), 1 point for any minor error (e.g., correct word but spelling error), and 0 points for major errors (e.g., incorrect word) or absent answers, resulting in a maximum score of 220 points.

Statistical Analyses

Data cleaning. Outliers were determined using the outlier labeling rule (Hoaglin, Iglewicz, & Tukey, 1986), using the interquartile range multiplied with a factor (*g*) of 2.8 to determine the cutoff. A total of four individual data points (i.e., a score on a test at one point in time) were deleted and treated as missing values.

Bayesian linear mixed models. Bayesian linear mixed models were estimated for each manifest variable. Dependent variables were standardized (to the pretest *SD*) to have a mean of 0 and a variance of 1. The model included between-subjects predictor group (relaxation = 0, language = 1), within-subject predictor time (pre = 0, post = 1), and the Group × 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 and can be interpreted analogous to the Group × Time interaction in a traditional 2 × 2 ANOVA, where positive values denote greater gains for the language learning group. For estimation purposes we used weakly informative normal (0, 1) priors for all fixed effects. We computed 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 a $\theta \sim$ half-normal (0, 0.2) informed prior, based on the expected effect size in the power analysis, and a $\theta \sim$ half-normal (0, 0.707) commonly used reference prior, combined with the likelihood estimate from the uninformed model. In addition to the Bayesian mixed models, we employed structural equation modeling (SEM) to test the hypothesis of differential gain for the two treatment groups. Analytical advantages of SEM include better accounting for measurement error, the ability to construct latent factors from observables, and in the present case, explicitly modeling the relationship between baseline performance and change.

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. Latent factors represent the shared variance among multiple tests purporting to measure the same ability. 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 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 posttest, 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., 2018), and Δ captures the latent change between PRE and POST. See Online Supplement 4 for a graphical illustration of the latent change score model.

We set these models up in a multigroup framework, where the models were estimated for the language group ($n = 90$) and the relaxation 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, analogous to a Time × Group interaction in a traditional 2 × 2 ANOVA. The threshold for statistical significance was $\alpha = .05$ for all tests.

Measurement invariance was evaluated in a multigroup longitudinal framework (Kievit et al., 2018). 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 χ^2 $p > .05$) and in terms of model-constrained nested comparisons for weak, strong, and strict invariance (all $\Delta\chi^2$ $p > .05$; see Online Supplement 5 for details). Therefore, we proceeded to test the hypothesis of differential change on the level of strict measurement invariance.

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

Descriptives

Means, standard deviations, and psychometric properties of the cognitive variables, as a function of treatment condition and session, are presented in Table 1. The scores on the poststudy vocabulary test were visually inspected and deemed to be approximately normally distributed with a mean of 115.6 (SD = 42.0; range = 6–214), suggesting that the average participant in the language learning group did indeed acquire a basic Italian vocabulary.

Bayesian Analysis

Results from the Bayesian linear mixed model on each of the individual cognitive tests are presented in Table 2. The findings showed that the net positive effect θ (95% CI; BF), analogous to a Time × Group interaction in a traditional 2 × 2 ANOVA, ranged from θ = −0.24 (95% CI [−0.51, 0.03]; BF_{01} = 4.12) for WASI to θ = 0.18 (95% CI [−0.09, 0.46]; BF_{01} = 0.57) for syllogisms. Bayesian triplots can be seen in Online Supplement 6.

Latent Change Score Model

Results from 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 ranged from 0.04 to 2.56, $\Delta\chi^2$ p range = .110–.841 (for detailed information see Online Supplement 7). This test can be interpreted analogous to a Time × Group interaction in a traditional 2 × 2 ANOVA. For illustration purposes, we created 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. The mean pre-post change is comparable in both groups for all five cognitive domains, in line with our previous results of no differential change between groups.

Discussion

Language learning has been suggested as a promising intervention for improving cognitive performance in older age (Antoniou et al., 2013; Antoniou & Wright, 2017). We tested this hypothesis in a randomized controlled study. Latent modeling 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 relative to relaxation (BFs ranged from BF_{01} = 1.06 for n-back to BF_{01} = 4.12 for WASI), with the exception of syllogisms 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 (standardized mean difference [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 verbal intelligence, spatial intelligence, associative memory, item memory, or working memory to any noticeable extent on the group level. The observed pre-post change in both groups is on the same magnitude as what have been reported for the no-contact (i.e., passive) control groups in meta-analysis of working memory training (e.g., Karbach & Verhaeghen, 2014), suggesting that the observed pre-post changes are merely retest effects. We therefore conclude that an entry-level language course aimed at older healthy adults is unlikely to have any substantial effect on memory or reasoning performance.

We argue that these findings are not due to low statistical power or poor measurement. Statistical power was adequate (≈ .80) to detect quite small effects (d = 0.2), which is also clear from the precision of our estimates. Cognitive abilities were statistically represented as latent factors of multiple cognitive tasks that 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, working memory) as well as abilities that are less likely to be central in

Table 2
Posterior Standardized Results of the Bayesian Linear Mixed Model

Domain	Test	Differential mean change	SD	95% CI	BF_{01} half-normal (0, σ) σ = .2	BF_{01} half-normal (0, σ) σ = .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
	Syllogisms	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

Note. BF = Bayes factors; WASI = Wechsler Abbreviated Scale of Intelligence. Estimates obtained using uninformative normal (0, 1) priors. Bayes factors were approximated using half-normal (0, σ) priors with σ = .2 (informed prior) and σ = .707 (reference prior), using the likelihood of the uninformed model.

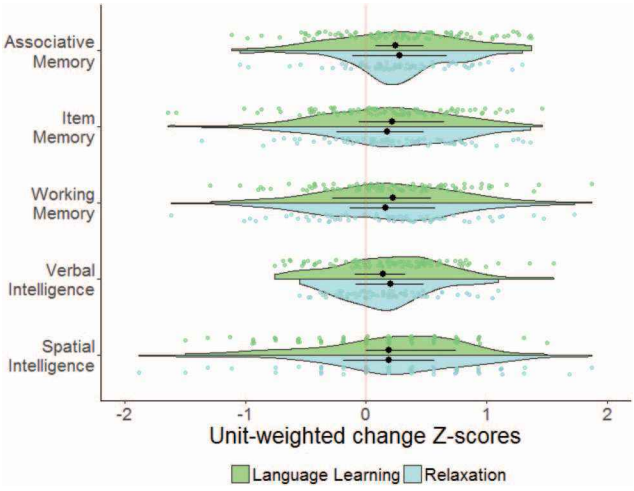

Figure 1. Distribution of unit-weighted change (posttest–pretest) scores. Green dots denote individual change scores of participants in the language group, and blue dots denote change scores for the participants in the relaxation group. Black dots and line segments denote the median and first and third quartiles.

language learning (i.e., reasoning abilities). The task selection can nevertheless be challenged for not capturing all aspects of cognitive performance that could be expected to be affected by language learning. For example, the use of complex updating tasks to assess working memory performance, as opposed to classic complex tasks, may be criticized based on reports of low correlations between the two task types (e.g., Kane, Conway, Miura, & Colflesh, 2007). We note, however, that others have shown that latent factors of performance on updating tasks and complex span tasks are highly correlated and that the factors predict reasoning ability equally well (Schmiedek, Hildebrandt, Lövdén, Wilhelm, & Lindenberger, 2009; Schmiedek, Lövdén, & Lindenberger, 2014; Wilhelm, Hildebrandt, & Oberauer, 2013). Furthermore, detailed task analyses show a substantial processing overlap between the two task paradigms (e.g., Ecker, Lewandowsky, Oberauer, & Chee, 2010), and updating tasks appear to predict language performance at least as well as complex span tasks (e.g., Carretti et al., 2009). Thus, we trust the updating tasks to have provided a reliable and valid measure of working memory ability in the present study.

We further argue that the intervention dosage was not unreasonably low. Although our ability to assess how much of the language that participants actually acquired is limited because we did not comprehensively assess language proficiency, a vocabulary test administered at the end of the intervention confirmed that the average participant in the language learning course had acquired basic Italian vocabulary. We further note that the extent of the language studies was more intensive than many previous intervention studies in older adults. It comprised up to 55 hr of teacher-led classes focusing on verbal communication exercises that previous research suggests are demanding working memory (e.g., Linck et al., 2014; Ullman, 2004). The time in class was 3 times longer than the typical beginners' course at the adult education center that administered the course. The participants also studied up to 485 new words between classes, and the importance of this was emphasized by weekly vocabulary tests. The course and the accompanying book cover at least the first (A1) level of the common European framework of reference for languages. The amount of intervention received also differed substantially between treatment conditions (5 hr per week for language learning vs. 1 hr per week for relaxation training). While we are aware that this discrepancy may be seen as a limitation, this choice was made in order to maximize the effect of the language learning, while still facilitating a believable relaxation training intervention. We deemed 1 hr per week of relaxation to be insufficient to plausibly alter cognitive functioning, while 5 hr per week would be untenable as relaxation training. Crucially, this intensity imbalance would rather serve to *increase* posttest differences, of which we find none. Comparable intensity (e.g., by extending the relaxation training to 5 hr per week) would likely further diminish any observed posttest difference.

We cannot rule out that extending the scope and length of the language intervention would result in improved cognitive performance. However, to simply suggest that the intervention was not sufficient (e.g., "long enough" in terms of duration, "taxing enough" in terms of intensity, etc.) is not meaningful nor helpful for the scientific process since the word itself is defined as "enough to bring about change." Instead, the exact circumstances under which change would be predicted need to be carefully specified to arrive at testable hypotheses. Unfortunately, current theories are not formulated to allow for predictions of what adjustments would be needed to facilitate improvements in cognitive functioning. In fact, a lack of well-specified and testable theories represents a challenge for work on cognitive interventions in general (e.g., Lindenberger, Wenger, & Lövdén, 2017). Nevertheless, we can conclude from the present results that completing an intensive entry-level foreign language course in older age is unlikely to substantially improve any of the cognitive functions tested here.

It is worth addressing the fact that we have not framed this study in terms of the bilingual advantage hypothesis, which predicts that bilinguals experience later onset of cognitive decline and better cognitive performance on certain tasks, particularly those involving executive functioning (e.g., inhibitory control and task switching; e.g., Bialystok, Craik, & Luk, 2012). The proposed mechanism is that bilinguals need to switch between two languages, suppressing one in favor of the other, which would exert a cognitive load similar to that of long-term cognitive engagement, which over time would make bilinguals better than monolinguals at focusing on task-relevant information and more resilient to distracting information. The first reason for not considering the bilingual hypothesis here is that the empirical support for the hypothesis has recently come under scrutiny, questioning the reliability of the previously reported executive advantage of bilinguals (e.g., Paap & Greenberg, 2013; Paap, Johnson, & Sawi, 2015; Lehtonen et al., 2018). The second reason is that the entry-level language course in the present study cannot be reasonably expected to result in a level of secondary language proficiency that would necessitate competition with the native language and, thus, to exert the required executive load. Instead, the present language intervention aimed to function as a type of cognitive intervention by taxing abilities such as associative and working memory, in the absence of mastery, which is in line with previous proposals (e.g., Antoniou et al., 2013). In the context of the bilingual hypothesis and achieving competition between the native and second lan-

guage, it may be helpful to ask what degree of language learning would be required to achieve this and how realistic such an intervention would be in older adults with a limited time span. Taken together, the findings presented here cannot and should not inform the bilingual hypothesis.

It is possible that the present study sample was particularly resistant to the effects of language learning on cognition. Study participants were of above-average educational attainment, and their decision to take part in the study may be interpreted as a general willingness to engage in cognitive activities. As such, it can be speculated that the study participants had already accrued the proposed cognitive benefits of cognitive engagement and that findings would have differed in a sample of cognitively understimulated participants, akin to selective benefits of physical exercise intervention for less physically active individuals.

The results of the present study demonstrate that an entry-level language course aimed at older healthy adults is unlikely to have any substantial effect on general cognitive ability. Although interpretation of the results of cognitive interventions remain debated, the present results align with a general theme of limited generalization of benefits of cognitive 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). Thus, as it currently stands, foreign language learning in old age should not be recommended for improving cognition but for learning a skill that is invaluable for communication with people who do not share one's native language.

References

Antoniou, M., Gunasekera, G. M., & Wong, P. C. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: A hypothesis for future research. *Neuroscience and Biobehavioral Reviews, 37,* 2689–2698. http://dx.doi.org/10.1016/j.neubiorev.2013.09.004

Antoniou, M., & Wright, S. M. (2017). Uncovering the mechanisms responsible for why language learning may promote healthy cognitive aging. *Frontiers in Psychology, 8,* Article 2217. http://dx.doi.org/10.3389/fpsyg.2017.02217

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8,* 47–89. http://dx.doi.org/10.1016/S0079-7421(08)60452-1

Bellander, M., Eschen, A., Lövdén, M., Martin, M., Bäckman, L., & Brehmer, Y. (2017). No evidence for improved associative memory performance following process-based associative memory training in older adults. *Frontiers in Aging Neuroscience, 8,* Article 326. http://dx.doi.org/10.3389/fnagi.2016.00326

Bialystok, E., Craik, FI, & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16,* 240–250. http://dx.doi.org/10.1016/j.tics.2012.03.001

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80,* 1–28. http://dx.doi.org/10.18637/jss.v080.i01

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences, 19,* 246–251. http://dx.doi.org/10.1016/j.lindif.2008.10.002

Daneman, A., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19,* 450–466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 422–433. http://dx.doi.org/10.3758/BF03214546

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London, 364,* 3773–3800. http://dx.doi.org/10.1098/rstb.2009.0111

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 170–189. http://dx.doi.org/10.1037/a0017891

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Faul, F., Erdfelder, E., Lang, A. G., & 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. http://dx.doi.org/10.3758/BF03193146

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. New York, NY: Psychology Press.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81,* 991–999.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the *N*-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 615–622. http://dx.doi.org/10.1037/0278-7393.33.3.615

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132,* 47–70. http://dx.doi.org/10.1037/0096-3445.132.1.47

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science, 25,* 2027–2037. http://dx.doi.org/10.1177/0956797614548725

Kievit, R., Brandmaier, A. M., Ziegler, G., van Harmalen, A. L., de Mooij, S. S. M., Moutoussis, M., . . . the NSPN Consortium. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience, 33,* 99–117. http://dx.doi.org/10.1016/j.dcn.2017.11.007

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin, 144,* 394–425. http://dx.doi.org/10.1037/bul0000142

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21,* 861–883. http://dx.doi.org/10.3758/s13423-013-0565-2

Lindenberger, U., Wenger, E., & Lövdén, M. (2017). Towards a stronger science of human plasticity. *Nature Reviews Neuroscience, 18,* 261–262. http://dx.doi.org/10.1038/nrn.2017.44

Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin, 136,* 659–676. http://dx.doi.org/10.1037/a0020080

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 223–267). Hillsdale, NJ: Erlbaum.

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,* 512–534. http://dx.doi.org/10.1177/1745691616635612

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1170–1187. http://dx.doi.org/10.1037/0278-7393.26.5.1170

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences, 29,* 1017–1045. http://dx.doi.org/10.1016/S0191-8869(99)00251-2

Olsson, M. R., & Braconi, P. E. (2005). *Buon viaggio!*. Stockholm, Sweden: Bilda Förlag.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66,* 232–258. http://dx.doi.org/10.1016/j.cogpsych.2012.12.002

Paap, K. R., Johnson, H. A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 69,* 265–278. http://dx.doi.org/10.1016/j.cortex.2015.04.014

Payne, B. R., & Stine-Morrow, E. A. L. (2017). The effects of home-based cognitive training on verbal working memory and language comprehension in older adulthood. *Frontiers in Aging Neuroscience, 9,* Article 256. http://dx.doi.org/10.3389/fnagi.2017.00256

Raven, J. C. (1960). *Guide to the standard progressive matrices: Sets A, B, C, D and E.* London, UK: H. K. Lewis.

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.2.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rönnlund, M., Nyberg, L., Bäckman, L., & Nilsson, L. G. (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,* 3–18. http://dx.doi.org/10.1037/0882-7974.20.1.3

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36. http://dx.doi.org/10.18637/jss.v048.i02

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

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49,* 304–313. http://dx.doi.org/10.1037/0003-066X.49.4.304

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1089–1096. http://dx.doi.org/10.1037/a0015730

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology, 5,* Article 1475. http://dx.doi.org/10.3389/fpsyg.2014.01475

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17,* 103–186. http://dx.doi.org/10.1177/1529100616661983

Stine-Morrow, E. A. L., Payne, B. R., Roberts, B. W., Kramer, A. F., Morrow, D. G., Payne, L., . . . Parisi, J. M. (2014). Training versus engagement as paths to cognitive enrichment with aging. *Psychology and Aging, 29,* 891–906. http://dx.doi.org/10.1037/a0038244

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92,* 231–270. http://dx.doi.org/10.1016/j.cognition.2003.10.008

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI).* San Antonio, TX: Psychological Corporation.

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

Received October 15, 2018
Revision received November 7, 2019
Accepted December 1, 2019 ■