COGNITIVE BENEFITS OF LEARNING ADDITIONAL LANGUAGES IN OLD ADULTHOOD? INSIGHTS FROM AN INTENSIVE LONGITUDINAL INTERVENTION STUDY

Maria Kliesch[1,2], Simone E. Pfenninger[3], Martijn Wieling[4], Elisabeth Stark[1,2], Martin Meyer[5,6,7,8]

[1] *Zurich Center for Linguistics, University of Zurich, Switzerland*
[2] *Chair of Romance Linguistics, Institute of Romance Studies, University of Zurich, Switzerland*
[3] *Department of English and American Studies, University of Salzburg, Austria*
[4] *Center for Language and Cognition Groningen, Faculty of Arts, University of Groningen, Groningen, The Netherlands*
[5] *Department of Comparative Language Science, University of Zurich, Switzerland*
[6] *Cognitive Psychology Unit, Department of Psychology, Alpen-Adria University, Klagenfurt, Austria*
[7] *Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, Switzerland*
[8] *University Research Priority Program (URPP) Dynamics of Healthy Aging, University of Zurich, Switzerland*

Corresponding author:
Maria Kliesch, maria.kliesch@uzh.ch
Universität Zürich
Kompetenzzentrum für Linguistik
Andreasstrasse 15
CH-8050 Zürich

Abstract

Second language (L2) learning has been promoted as a promising intervention to stave off age-related cognitive decline. While previous studies based on mean trends showed inconclusive results, this study is the first to investigate nonlinear cognitive trajectories across a 30-week training period. German-speaking older participants (aged 64-75 years) enrolled for a Spanish course, strategy game training (active control) or movie screenings (passive control). We assessed cognitive performance in working memory, alertness, divided attention and verbal fluency on a weekly basis. Trajectories were modelled using Generalized Additive Mixed Models to account for temporally limited transfer effects and intraindividual variation in cognitive performance. Our results provide no evidence of cognitive improvement differing between the Spanish and either of the control groups during any phase of the training period. We did, however, observe an effect of baseline cognition, such that individuals with low cognitive baselines increased their performance more in the L2 group than comparable individuals in the control groups. We discuss these findings against the backdrop of the cognitive training literature and Complex Dynamic Systems Theory.

*Keywords***:** older adults; language learning; longitudinal; intra-individual variation; cognitive training; working memory; GAMM

INTRODUCTION

Research has repeatedly identified great variability in cognitive performance throughout the lifespan, such that some individuals show stable cognitive performance even in the presence of structural and functional declines (Christensen 2001; Park and Reuter-Lorenz 2009; Wilson et al. 2002). While partly determined by genetic, demographic and (other) circumstantial factors (e.g. Fagan and Pihlstrøm, 2017; Hayat et al. 2016), a number of lifestyle factors and activities have been identified that may mitigate the adverse effects of structural and functional decline, such as an enriched environment, healthy nutrition (Mora 2013), physical activity (Bamidis et al. 2014) or making music (Mansens et al. 2018). The available epidemiological evidence further indicates that socially and mentally stimulating leisure activities (Lövdén et al. 2005; Stine-Morrow et al. 2014; Zuelsdorff et al. 2019) as well as activities that tap into multiple cognitive domains (Binder et al. 2016) reliably predict changes in cognitive performance in the third age. In this regard, the term "Third Age" is used to refer to the stage of life initiated by retirement that is characterized by a greater amount of free time, personal fulfillment and an active lifestyle (Pfenninger & Singleton, 2019). While it is undoubtedly important for the well-being of older individuals, attenuating cognitive decline is also crucial for society, as the share of older persons in the total population will increase significantly in the coming decades, not just within the EU.

In addition to being socially engaging, ecologically relevant and purposeful, learning a second language (L2) is a highly complex cognitive task. It involves the encoding, storage and retrieval of arbitrary relations between phonemes, words and their meanings, the concurrent maintenance and updating of syntactic and semantic information, application of patterns not present or even contradictory to the ones of the respective L1, attention to relevant new information, prediction of patterns, et cetera. In particular, L2 learning places high demands on attentional, verbal and working memory (WM) processes (Issa and Morgan-Short, 2018; Linck et al. 2014), all of which have been shown to be affected by age-related declines (e.g. Salthouse 2010). Since L2 learning has been found to engage an extensive brain network that is known to overlap with regions negatively affected by aging, some researchers have proposed L2 learning in the third age as a promising way to stave off age-related cognitive declines (Antoniou et al. 2013; Antoniou and Wright, 2017). While older adults have been reported to manifest a strong interest in learning new languages (Long et al. 2019), the focus on L2 program development has commonly ignored this interest group. Similarly, while cognitive benefits of *lifelong* bilingualism have received considerable scientific

attention, only few behavioral and neurocognitive studies have investigated how cognitive ability changes as a function of L2 learning *beginning* in the third age. Here, we address this question by analyzing the microdevelopment of cognitive performance as a function of a 30-week L2 training for older adults, taking into account both intra- and interindividual difference factors that moderate these transfer effects.

Cognitive benefits of L2 learning in old age

To the best of our knowledge, a total of eight longitudinal intervention studies have been conducted so far on healthy adults analyzing the transfer effects of foreign language learning on general cognitive functioning (see Table 1). In this context, "transfer effects" typically refer to the enhancement of domain-general cognitive abilities that are not explicitly trained in a given intervention (Guye and von Bastian, 2017), such as the improvement of WM capacities through L2 learning. Findings from the existing studies do not provide conclusive results, which two groups of researchers have attributed to pitfalls in study designs and methodological concerns (Berggren et al. 2020; van der Ploeg, Keijzer and Lowie 2020).

Table 1

Overview of Previous Studies Investigating Transfer Effects of L2 Learning in Old Adulthood

Authors	Age of Participants	L1 & L2 to Be Learned	Groups	Group Assignment	Duration & Intensity	Cognitive Measure	N Repeated Measures	Result
Long et al. 2020	21-85	L1: English L2: Gaelic	LANG1: L2 beginner LANG2: L2 elementary LANG3: L2 intermediate	Non-random	1 week 14h	Test of Everyday Attention (TEA)	2 (before/after)	Selection bias: LANG3 > LANG1; LANG1 improved more than LANG3, but remained below LANG3's baseline level
Berggren et al. 2020	65-75	L1: Swedish L2: Italian	LANG: learning L2 in classrooms ACTV: relaxation training	Random	3 months 5h/week (ACTV only 1h/week)	Associative memory Item memory Working memory Verbal intelligence Spatial intelligence	2 (before/after)	No difference in baseline or group*time interaction
Valis et al. 2019 & Klimova et al. 2020	*M* = 71, *CI* = [69, 73]	L1: Czech L2: English	LANG: learning L2 in classrooms PASV: no training	Random	3 months 45min/week	MoCA	2 (before/after)	No difference between groups

Wong et al. 2019	60-85	L1: Chinese L2: English	LANG: learning L2 with software ACTV: playing games (sudoku, crosswords...) PASV: music appreciation	Random	6 months 5h/ week	Clinical Dementia Rating Alzheimer's Disease Assessment Scale - Cognitive Subscale Category Verbal Fluency	3 (before/after/3 months follow-up)	No group*time interaction
Bubbico et al. 2019	59-79	L1: Italian L2: English	LANG: learning L2 in classrooms PASV: no training	Random	4 months 2h/week	MMSE Speed attention Immediate and delayed verbal memory Executive functions	2 (before/after)	Selection bias: PASV > LANG at pre-training; LANG remained stable, PASV showed decrease in performance; No group difference in performance at post-training
Ware et al. 2017	63-90	L1: French L2: English	LANG: translating sentences from L2 to L1 in class using online dictionaries for help.	/	4 months 2h/week	MoCA	2 (before/after)	No change from pre to post

Ramos et al. 2016	60-80	L1: Spanish L2: Basque	LANG: learning L2 in classrooms PASV: no training	Non-random	8 months 5.5h/week	Switching ability	2 (before/after)	Selection bias: PASV > LANG; No group*time interaction
Bak et al. 2016	18-78	L1: English L2: Gaelic	LANG: learning L2 in classrooms ACTV: other types of courses PASV: no training	Non-random	1 week 14h	Test of Everyday Attention	3 (before/after/9 months follow-up)	LANG improved more than PASV in one of the subtasks; Younger participants in LANG group outperformed their older peers Only participants who kept practicting the L2 for more than 5h/week showed better performance at follow-up than at pre-training

Most of the studies in Table 1 compared the L2 training to control conditions, such as relaxation training (Berggren et al. 2020). Other studies used passive control conditions (PASV) for comparison (Bubbico et al. 2019; Ramos et al. 2016; Valis et al. 2019); yet others used both a passive and an active (ACTV) control group. Table 1 also provides an overview of the cognitive variables for which pre- and post-training performance was compared between groups. Out of the eight studies, two identified the hypothesized positive group*time interaction. Bubbico et al. (2019) found that the LANG group showed a stable performance in global cognition (i.e. MMSE scores) from pre- to post-measurement, while group PASV showed a decrease in this respect. Their LANG participants, however, were significantly older, less educated and performed significantly worse during pre-training than the passive control group, with scores in the Mini-Mental-Status Examination (MMSE) suggesting mild cognitive impairment in at least some of the LANG participants (see also Ware et al. 2017). The second study that found increased improvement in the LANG as compared to a PASV group (Bak et al. 2016) only did so for one of the subtasks of the Test of Everyday Attention. The training in this study, however, had a duration of one week only. At a 9-months follow-up, cognitive performance was improved in eight participants who had continued to study Gaelic for at least five hours a week. Hence, one week of L2 training alone was not sufficient to yield long-term changes in cognitive ability (see also Antoniou et al. 2013). The remaining studies with training durations of 3-8 months could either not identify any change in cognitive performance at all or did not find a group*time interaction (Berggren et al. 2020; Ramos et al. 2016; Valis et al. 2019; Ware et al. 2017; P. C. M. Wong et al. 2019). Importantly, three of the studies (Bubbico et al. 2019; Valis et al. 2019; Ware et al. 2017) applied either the Montreal Cognitive Assessment (MoCA) or the MMSE to assess cognitive changes, both of which are screening tools for cognitive impairment that manifest ceiling effects in healthy older adults and have been shown to be susceptible to practice effects, particularly between the first and second administrations (Cooley et al. 2015; Duff et al. 2007). Hence, it is possible that by using more sensitive assessment tasks and measures, such as reaction times and tasks assessing specific cognitive domains, differences in cognition between experimental groups could be observed. In three studies where this was done; however, the authors in question (Berggren et al. 2020; Ramos et al. 2016; P. C. M. Wong et al. 2019) could not find any transfer effects of L2 learning. For instance, Ramos et al. (2016) and Berggren et al. (2020) applied Bayesian statistics and found the null hypothesis to be more probable, which lead the authors of the latter study to conclude that “an entry-level language course aimed at

older healthy adults is unlikely to have any substantial effect on general cognitive ability" (p. 218).

Intraindividual Variability and Dynamic Systems

Research has shown that cognitive performance fluctuates within individuals and that it can vary across and even within days – a phenomenon that is exacerbated with increasing age (Martin and Hofer 2004). Fluctuations in cognitive ability can be attributed to levels of stress (Neupert et al. 2006), belief of competence (Neupert and Allaire 2012) and task motivation (Chiew and Braver, 2013). Similarly, negative affect linked with poor health has been associated with low cognitive performance on any given day (Strauss et al. 2002). These intraindividual differences make it difficult to interpret an individual's overall level of cognition based on performance recorded at one single occasion. In addition, and importantly for studies investigating the effects of cognitive training, Complex Dynamic Systems Theory (CDST) (Larsen-Freeman 1997, 2017; Larsen-Freeman and Cameron 2008; Verspoor, Lowie, and Van Dijk 2008) further predicts that even if the training situation is held constant for each participant, individuals likely vary in within-person processes over time, which can produce between-person differences in training outcomes (Könen and Karbach 2015).

As a consequence, by interpreting cognitive phenomena on mean trends and variance of group scores at 2-3 points in time, previous studies on the cognitive benefits of L2 learning may have underestimated the complexity of the developmental process and misinterpreted within-subject variability as group effects.

The present study

The present longitudinal intervention study takes into account previous methodological biases in line with the premises of CDST, the dynamic nature of cognitive and L2 performance and the potential moderating effect of socio-affective factors. The L2 training (LANG) conducted in this study consisted of a semi-computerized Spanish training including social interaction, which, on the one hand, allowed participants to advance at their own pace, and, on the other, gave them ample opportunity for oral, listening and communication practice. The training duration of 30 weeks and the intensity of 5 hours per week is based on Antoniou et al. (2013: 2694-2695), estimating that learning-related cognitive and structural changes "should be expected within six months of commencing language training, with training occurring for 1h per day, 5 days per week". To assess cognitive transfer effects of the L2 training, we

implemented two carefully controlled experimental conditions via an active (ACTV) and a passive control group (PASV) designed to account for the use of technology in the L2 training (ACTV), the social interaction, the entertainment/motivation factor and both training duration and intensity between groups (ACTV and PASV). Due to a number of methodological, practical and ethical considerations, group assignment was not random. For one, a relatively high level of time commitment and effort was required from participants over several months (2–5 hrs per week over 30 weeks, excluding travel time). Thus, random assignment to an experimental group not favoured by a participant would have been ethically problematic, and would have led to an unpredictably higher dropout rate, which in turn would likely have affected group dynamics and overall motivation. In addition, and in line with Carey and Stiles (2015), we argue that, for psychological treatment studies such as this one, groups need to be at least as equivalent in terms of commitment and motivation to the treatment as they are in terms of gender, age or other background variables, because socio-affective factors can be expected to be highly influential for the treatment outcome. Therefore, in the present study, gender, age, education, prior multilingualism, the number of regular activities, training motivation and overall wellbeing were controlled experimentally between groups. Furthermore, by performing voluntary group assignment, observational studies such as ours have the advantage of being conducted under realistic conditions of the target population, in which retirees are unlikely to participate in regular and time-consuming leisure activities that are not of their own choosing. As a consequence, participants in the present study applied directly for the training that most appealed to them, but they were not explicitly informed that their cognitive performance was going to be compared to that of the other groups. A cognitive transfer test battery assessing verbal fluency, working memory, divided attention and alertness as well as an assessment of training motivation and overall wellbeing were administered on a weekly basis for a total of 30 weeks. Conducting such a longitudinal study with enough data points is necessary in order to capture in detail the way each individual develops over time as well as individual fluctuations in cognitive performance. While such a design does not lend itself to generalizations, it is well suited to disentangle mechanisms that have differing time-courses, which in turn may help reconcile previous and inconclusive results on the effectiveness of L2 training on cognitive performance and the training duration after which improvements can be expected. The study was approved by the Ethics Committee

of the Philosophical Faculty of the University of Zurich. The data and codebook can be accessed on the Open Science Framework (OSF[1]).

Hypotheses

We hypothesize (1) the LANG training to tap into a wide range of cognitive capacities and therefore have a higher chance of overlapping with the assessed cognitive skills than the PASV training. If this were the case, participants from the LANG group would increasingly outperform those from group PASV in terms of the cognitive tasks. In order to fully address Antoniou et al.'s hypothesis (2013) of the unique characteristics of L2 learning, we also hypothesize (2) that the LANG training is more effective in eliciting transfer effects than the ACTV training, which comprised a cognitively challenging but non-linguistic training paradigm. Finally, in an exploratory approach, we examined whether the effectiveness of the LANG training is influenced by the baseline condition of cognitive performance in an attempt to shed light on the seemingly contradictory findings of previous research and in order to help explain the potential benefits for some but not all older adults. This is an open empirical question, where no specific hypotheses were formulated.

METHOD

Participants

Participants were recruited through study advertisements in local newspapers, lectures for senior citizens at the Zurich University of Applied Sciences, websites for third age universities and personal calls to third age leisure clubs, associations and help groups. Participants were first screened for eligibility in a telephone interview. Inclusion criteria included their age (between 64 and 75 years), retirement status, being neurologically and psychiatrically healthy, having no learning disabilities and no untreated severe hearing impairment, not being a professional musician, German or Swiss German mother tongue and no more than school knowledge of any language other than (Swiss) German. For the LANG training, participants were additionally required to possess a smartphone, tablet or computer that could be connected to the internet. For the ACTV training, participants needed an internet-ready computer or laptop. If these inclusion criteria were met based on self-reports in the telephone interview, individuals were scheduled for a screening session. At the beginning

[1] DOI 10.17605/OSF.IO/WCFJ3

of the screening session, participants provided written informed consent and completed a questionnaire to assess all relevant ID variables, such as age, retirement/employment status, health and other background variables. Participants were required to score at least 26 points in the Montreal Cognitive Assessment (MoCA, Nasreddine et al. 2005). For participation in the training, participants were given snacks and drinks during the weekly in-person get-togethers and were reimbursed a symbolic 150 CHF (approximately 150 USD) at the end of the (30-week) study. The final sample after screening consisted of 65 participants (see Table 2 for demographics). Two participants could not complete the 30 training sessions due to a cerebral stroke. Data from two further participants who completed the training was discarded because they suffered a minor stroke halfway through the training period.

Table 2

Study Characteristics of the Whole Sample and for Each Training Group Separately

Demographics	Training Group			
	All	LANG	ACTV	PASV
Sample size (F,M)	61 (26,35)	28 (14,14)	17 (7,10)	16 (5,11)
Average age (SD)	68.40 (2.92)	68.50 (2.83)	67.80 (2.86)	68.90 (3.19)
Education (SD)	12.60 (3.18)	12.30 (3.25)	12.60 (2.65)	13.20 (3.64)
Multilingualism BLP (SD)	28.80 (18.00)	29.60 (15.40)	30.00 (23.70)	26.00 (16.20)
No. of Activities (SD)	16.40 (2.95)	17.30 (2.73)	15.40 (2.92)	15.90 (3.05)
Socio-Affect (SD)	75.70 (19.10)	74.00 (18.80)	76.00 (21.00)	78.40 (17.20)

The three training groups did not differ significantly with respect to the ratio of female to male participants $\chi^2(2) = 1.48$, $p = .48$, age, $F(2,58) = 0.61$, $p = .55$, education, $F(2,58) = 0.44$, $p = .64$, multilingualism, $F(2,58) = 0.24$, $p = .79$ or number of regular activities $F(2,58) = 2.61$, $p = .08$.

Socio-affect was measured each week prior to the respective training and comprised one question on overall wellbeing and another on training motivation, the answers to which were indicated on a scale from 1-100 (see section Data Analysis). All three groups showed very high socio-affect (LANG: $M = 74.29$, $SD = 18.49$; ACTV: $M = 75.82$, $SD = 20.91$; PASV: $M = 78.37$, $SD = 17.10$). A Shapiro-Wilk test confirmed that socio-affect was not normally distributed ($p < 0.001$) and showed a clear ceiling effect, indicating that participants were very motivated and felt well. Significance testing via Kruskal-Wallis test did not show

any difference in mean socio-affect between groups ($\chi 2(2) = 0.53$, $p = 0.77$). A Generalized Additive Mixed Models (GAMM; see Data Analyses below) did not show normally distributed residuals after including this predictor due to the strong ceiling effect (see codebook on OSF). Hence, based on the comparability between groups and the potentially problematic residuals in the GAMM caused by this ceiling effect, socio-affect was assumed to be highly comparable between groups and was therefore not included in any of the subsequent analyses.

Training Procedure

The LANG group learned Spanish through the language learning software Duolingo (e.g., Vesselinov and Grego, 2012; Munday, 2016) and weekly classroom sessions. The ACTV group played an online real-time strategy game and met for strategy games on a weekly basis, while the PASV group met for social interaction and optional movie screenings.

We used the Duolingo School feature to monitor the number of experience points (XP) each individual obtained as we did not have access to the exact amount of time invested. Participants were requested to obtain 450XP per week, which corresponded to a usage of approximately 2-3 hours. If participants did not meet the target or surpassed it, they were reminded via email to practice or stop practicing, respectively, for the remainder of the week. The weekly classroom sessions were taught by a Spanish instructor and explicitly aimed at covering the areas that were deficient in Duolingo, i.e. oral practice, communication and grammar. Participants usually carried out L2 tasks in groups of 2-4 learners, while the instructor assumed the role of an observer during practice. The classes took place over two hours and included a number of L2 tests assessing L2 reception and production on both the lexical and morphosyntactic level (see Kliesch & Pfenninger, 2021 for a detailed description of the L2 training and tests).

In order to control for both the use of technology and the social exchange in the experimental group, the ACTV training also combined computerized and in-person training modalities. The computerized training consisted of playing “The Settlers Online”, a free, online real-time strategy game developed by Blue Byte GmbH and published by Ubisoft EMEA S.A.S. “The Settlers” is similar to “Rise of Nations”, which has been shown to yield significant cognitive improvement in older adults after 23.5 hours of training (Basak et al. 2008). Participants were instructed on how to play it by an experienced player and were requested to spend approximately three hours per week playing the game. Exact recording of

game time was not possible, because it was not tracked in the players' profile. However, all participants reported spending approximately 2-3 hours per week actively in the game. In addition, and in order to mirror to the classroom LANG course, participants in the ACTV group met on a weekly basis to play strategy board and card games, such as The Settlers, Cluedo, Scotland Yard, Dominion or El Dorado, each of which lasted approximately 2 hours.

The PASV group was intended as a passive control group that only mirrored the social interaction of the other two groups in order to ensure that all three groups were equally motivated. Participants in this group completed the cognitive test battery on a weekly basis and were invited to meet in a separate classroom to interact and watch a movie together. Movies were presented in German and only comprised feature films. Since we did not expect any cognitive benefits from watching movies, staying for the movies was voluntary, but participants commonly stayed for snacks and coffee. Any cognitive improvement encountered in this group can be interpreted as either a pure repetition effect or one enhanced by the effect of regular social interaction.

Cognitive Battery

The battery included five tasks measuring cognitive abilities that have been shown to decline as a function of age (e.g. Salthouse 2010) and that can reasonably be assumed and have been shown to be tapped by L2 learning. From these, we extracted seven cognitive variables of interest. Each task took participants approximately 1-5 minutes to complete. The assessed skills included verbal fluency, simple working memory, complex working memory, divided attention and alertness (see detailed description of the tasks in Kliesch & Pfenninger, 2021).

WM capacities have been shown to be a robust predictor of L2 development (Linck et al. 2014; Kliesch & Pfenninger, 2021) and are potentially increased in bilinguals compared to monolinguals (Grundy and Timmer 2017). Our WM composite consisted of a 2-Back and an Operation-Span task. A 2-Back task was used to gauge simple WM span. Letters were presented one at a time in upper- or lowercase at the center of the computer screen. The participants had to press a button whenever the letter was identical to the letter presented two items back (regardless of case) and press no button if it was different. Accuracy and reaction times (RT) were extracted as primary measures for this task; the latter was only calculated for correct hits and only for sessions that had a sufficient number of hits (i.e. less than 2SDs above/below mean). Complex WM, i.e. an individual's ability to store information while faced with an additional processing task, was measured via an operation span task adopted

from Lewandowsky et al. (2010). Participants evaluated mathematical equations (e.g. 1 + 10 = 12?) while simultaneously trying to remember associated sets of consonant letters. The primary measure of this task was accuracy of recall of consonants in the correct order.

Fluency and lexical organization in the L1 have been shown to be predictors of, and hence potentially affected by, L2 learning (Masrai and Milton 2015; Kliesch & Pfenninger, 2021). We administered the Regensburger Wortflüssigkeitstest [Regensburg Word Fluency Test] (Aschenbrenner, Tucha and Lange 2000), in which participants were instructed to write down as many words in German (L1) as possible within one minute, starting with a given letter (p, k, r or s) which alternated on a weekly basis, so that participants completed each version only once a month. The primary measure of interest was the total number of words produced.

The simultaneous allocation of attentional resources to linguistic form and meaning is often referred to as "divided attention" and has been found to be particularly challenging in L2 learning (Wong 2001). We used a simultaneity task to test participants' divided attention by instructing them to perform two tasks at the same time: 1) to use the computer mouse to follow a dot moving across the screen and 2) to complete a Stroop task (Stroop 1935) displayed at the center of the screen, for which they had to press a button whenever a color word was written in the color it described. The primary measures of interest were accuracy and RT (i.e. the time it took participants to press the key when color and word matched, and only while the mouse was positioned on top of the dot).

Low processing speed, i.e. the speed at which cognitive information-processing steps can be completed, has been shown to predict successful L2 learning (Nelson et al. 2012) and was assessed here through a simple alertness task (see Zimmermann and Fimm 2002). A cross appeared on the screen every 3.5-5 seconds, and participants were instructed to press a button as quickly as possible whenever it did. The primary measure of this task was RT.

Data Analyses

All analyses were conducted in R 4.0.2, and generalized additive mixed models (GAMMs) were modelled using the 'mgcv' package, version 1.8.34 (Wood 2012) and visualized using the 'itsadug' package, version 2.4 (Van Rij et al., 2017).

First, all RTs were log-transformed to reduce the typical right skew. For Verbal Fluency, the ease between letters (*k*, *p*, *r*, *s*) was corrected statistically by subtracting the mean of the respective four versions from each individual score. Wellbeing and Training

Motivation were combined into one score called 'Socio-Affect' owing to their theoretical similarity and strong correlation of $r = 0.7$. In a similar vein, 2-Back and Operation Span accuracy were integrated into 'WM accuracy' after a principal components analysis confirmed that they loaded on the same factor. The distributions of all resulting RTs and accuracy scores were inspected visually. This indicated a ceiling effect for Divided Attention accuracy, so that only Divided Attention RTs were used for further analyses. Finally, all cognitive variables were z-transformed, and RTs were subtracted from 0. In this way, all cognitive variables could be entered into a single model (see section on GAMMs below), and variables could be interpreted consistently with positive values indicating better performance. Finally, the very first data point of each participant was considered a familiarization trial and therefore removed from further analyses[2].

In order to model the cognitive development within each experimental group and test whether trajectories differed between them, we used generalized additive modeling (GAM; Wood, 2006, 2017; see Wieling, 2018, for a tutorial) as our analysis method. GAM is a class of statistical models in which the relationships between the response and predictor variables are modeled by smooth (basis) functions. While the effect of each covariate can be non-linear, the amount of non-linearity (i.e. the wiggliness) of each predictor is controlled by penalizing more complicated basis functions more strongly than simpler basis functions. In addition, mixed GAMs (GAMMs) allow for random effects to be estimated in order to account for structural variability in the data resulting from repeated measures within each subject. Hence, GAMMs constitute an ideal tool to address not only whether there is an overall effect of L2 training on cognitive performance, but also whether cognitive trajectories follow different patterns across time. As residuals of the fitted GAMs generally followed a scaled-*t* distribution (i.e. longer-tailed than a normal distribution), we fitted most model using this distribution (cf. Wieling, 2018). Since the models created with GAM analyses do not include easily interpretable coefficients, visualization is an essential part of the statistical analysis process. We refer the reader to the data and online code on OSF[3] for more information on model creation and criticism.

[2] Results were similar when including the first time point.
[3] DOI 10.17605/OSF.IO/WCFJ3

RESULTS

Improvement of L2 Competence

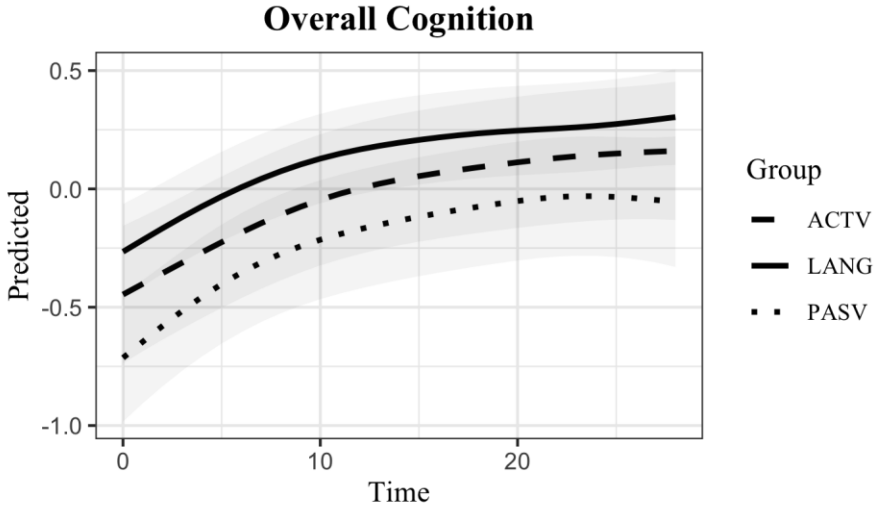
We found that L2 performance of the LANG learners significant improved in all of the seven measures of L2 proficiency. Significant growth in these capacities was observed during the first ten to twenty weeks (e.g. fluency or integrative skills) or consistently throughout the training (e.g. morphosyntactic accuracy). It can therefore be assumed that L2 learning did indeed take place. Detailed results on L2 trajectories and their predictors are reported in Kliesch & Pfenninger (2021).

Overall Training-Related Improvements on Cognitive Tasks

We found that overall cognitive performance increased over the course of the training in all training groups (see Figure 1). The fitted trajectories were obtained by creating a single GAMM that included normalized values for all cognitive measures as dependent variables as well as a smooth over time per training group (see Online Supplement SC1).

Figure 1

Fitted trajectories of overall cognitive performance in groups LANG, ACTV and PASV



Visual inspection of Figure 1 suggests that improvement in cognitive performance for all groups was highest in the first ten weeks and that there was a potential difference in baseline performance already at the beginning of the training. To confirm this, a GAMM was specified in which separate smooths over Time were created for groups ACTV and PASV (with group LANG forming the reference level), which were included as ordered factors (see Wieling,

2018) so as to obtain the actual differences in both intercepts and smooths over time (see Online Supplement SC2). This model confirmed a significant difference in the intercept between group LANG and PASV (p = .03), such that group LANG outperformed group PASV on average by 0.35 SDs (see Table 3 for model summary). Using this GAMM to predict group performance at the first measurement point showed that group LANG already outperformed group PASV by 0.40 SDs at the beginning of the study despite experimental controlling of background variables between groups.

Table 3

Model Fit of the GAMM Predicting Intercept and Development of Overall Cognition Based on Training Type

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.12	0.10	1.26	0.21	
isACTV01	-0.16	0.16	-1.02	0.31	
isPASV01	-0.35	0.16	-2.19	0.03	*
	edf	Ref.df	F	p-value	
s(Time)	3.59	4.59	17.19	< 0.001	***
s(Time):isACTV01	1.00	1.00	0.01	0.91	
s(Time):isPASV01	1.36	1.63	0.21	0.66	
s(Time,subject)	53.97	547.00	1.41	< 0.001	***

Note: Terms not marked “isACTV0” and "isPASV0" refer to the reference level for group LANG. All other terms constitute ordered difference smooths that capture the difference between trajectories LANG-PASV and LANG-PASV.

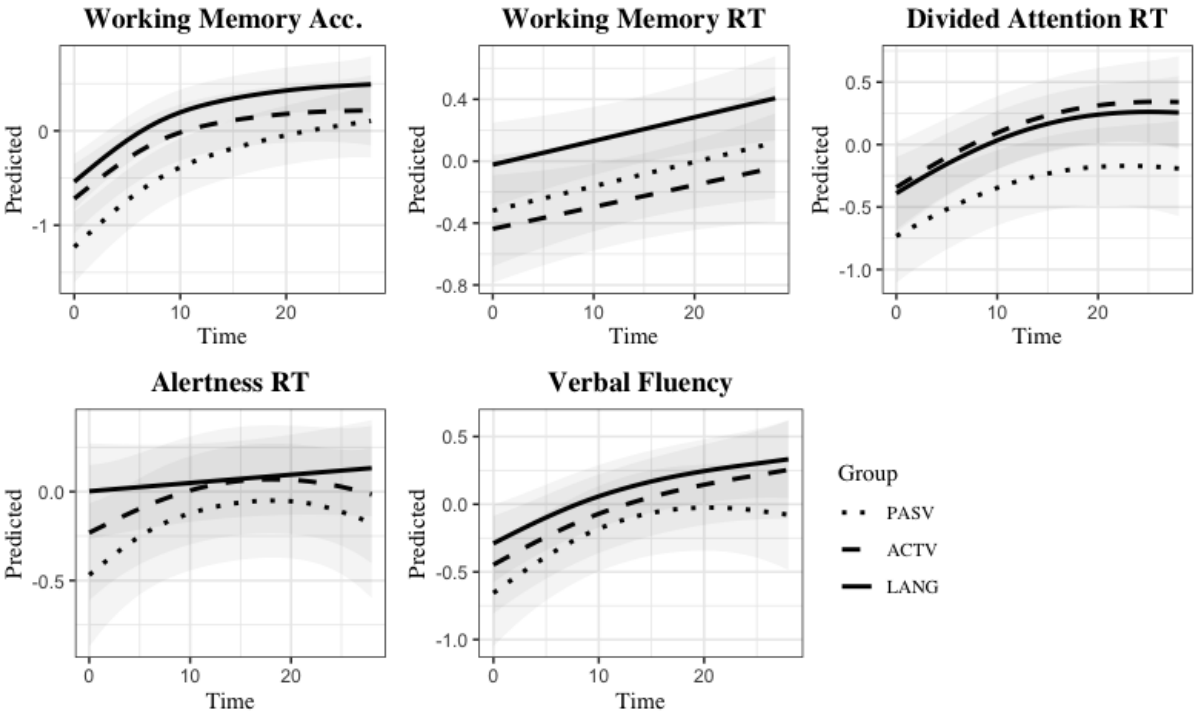
At the same time, none of the difference smooths over Time were significant, which means that the shape of the cognitive trajectories over time for groups ACTV and PASV did not significantly deviate from that of group LANG.

Training-Related Improvements on Individual Tasks

To examine potential group differences for each cognitive variable, a GAMM was created that included smooths over Time for each combination of cognitive variable and group, again using ordered factors (see Figure 2 for visualized trajectories and Online Supplement SC3).

Figure 2

Smooths over time per experimental group for each cognitive measure, with 95% confidence intervals plotted in gray. RTs were log transformed and inverted to make interpretation consistent across all measures (higher = better).



The model indicated that group LANG significantly outperformed group PASV in terms of Working Memory Accuracy and Divided Attention RT, and outperformed group ACTV regarding WM reaction time (see Online Supplement ST1 and Figure SF2 for individual trajectories). The respective smooths over Time show that participants improved on all cognitive measures except for Alertness RT, and that apart from the difference in intercept, there was no difference between groups regarding the shape of trajectories over time. See Table Online Supplement ST2 for descriptive statistics on each of the variables collected on a weekly basis.

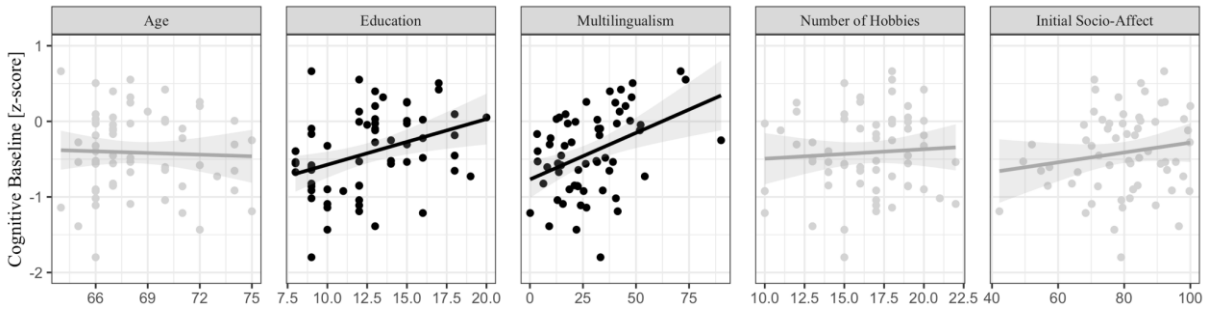
Individual Differences in Training-Related Improvements Based on Baseline Performance

Even though groups were comparable in terms of their demographics, group LANG outperformed group PASV in terms of overall cognition, WM accuracy and divided attention RT from the beginning and throughout the training (see above). In order to assess whether any of the background variables explained the difference in baseline performance, each

learner's performance was estimated at time point 0, using a GAM per subject with Time as the only predictor. This was done in order to account for intraindividual and training-independent fluctuations in cognitive performance at the onset of the study, which – if ignored – could have led to an over- or underestimation of an individual's actual baseline performance. Afterwards, this score was averaged over all cognitive variables to obtain a global measure for baseline performance and we subsequently correlated this measure with the background variables Age, Education, Multilingualism (BLP score), Number of Hobbies and Initial Socio-Affect (mean over first three weeks). Pearson's correlation showed a significant positive relationship between cognitive baseline across all three groups with both education ($r = .35$, $p = .01$) and multilingualism ($r = .41$, $p < .001$).

Figure 3

Regression plots of overall cognitive baseline performance with background variables. Black lines and dots mark relationships that showed a significant correlation.



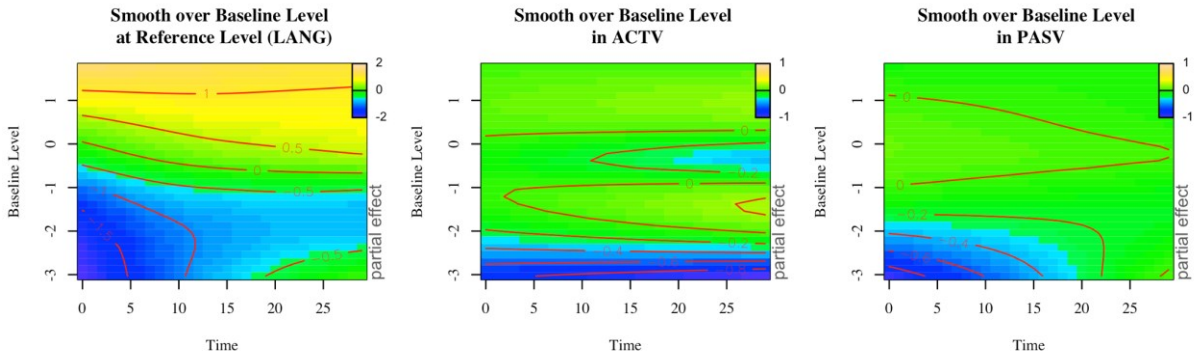
Given this baseline bias between groups, we created a new exploratory GAMM that modeled an interaction (i.e. a tensor product smooth) of Time and Baseline per group (as binary curves so as to include the uncertainty of the intercept in the main effects; see Wieling (2018), and Online Supplement SC4). Since two baseline scores were below 3.5SDs, we considered them as outliers and excluded the scores from the respective subject and task from this analysis[4], which affected Alertness RT of one subject and Working Memory RT of another (see Online Supplement SF3 for histogram of baseline levels). The resulting GAMM showed significant

[4] Results were similar when including the outliers, but the Time x Baseline interaction showed a slightly higher F-score (and lower p-value) for group PASV (including outliers: $F = 2.72$, $p = .01$; excluding outliers: $F = 2.00$, $p = 0.04$).

interactions of Time and Baseline level for all three groups (see Online Supplement ST3). Visualization of the respective interactions can be found in Figure 4.

Figure 4

Tensor product smooth for the interaction of Time and Baseline Level per group. Color coding is used to represent model predictions, with yellow indicating higher and blue representing lower cognitive scores.



As shown in the left panel of Figure 4, at the reference level (LANG), the difference in performance remained relatively stable throughout the training for participants with above-average baseline performance, which points to a potential ceiling effect. Since the contour lines become more vertical from top to bottom, this suggests that participants with low baselines increased their performance more than those with high baseline performance. The middle and the right panel of Figure 4, in contrast, visualize how this effect differed in groups ACTV and PASV, suggesting that participants with baseline performance below -2SDs improved their performance less in the ACTV and the PASV training conditions than in the LANG condition (i.e. by a difference of up to 0.8 SDs). In the PASV condition, however, this difference disappears during the second half of the training. The models with and without Baseline as predictor were compared by removing the above-mentioned baseline outliers from both datasets and adding selection penalties to both models, so as to be able to perform REML comparison (see Online Supplement SC5). The model that included Baseline as a predictor fit the data significantly better (fREML = 7575.74) than the one without (fREML = 8235.67, $p < .001$).

Finally, in order to investigate the influence of baseline performance on training benefit in each of the cognitive measures, a GAMM was created that included the interaction

of Time and Baseline for each combination of task and training type modelled as binary curves (see Online Supplement SC6). The model summary can be found in Table 5.

Table 4

Model Fit of the GAMM Predicting Intercept and Development of Each Cognitive Tasks Based on the Interaction Between Time and Baseline Performance per Training Type

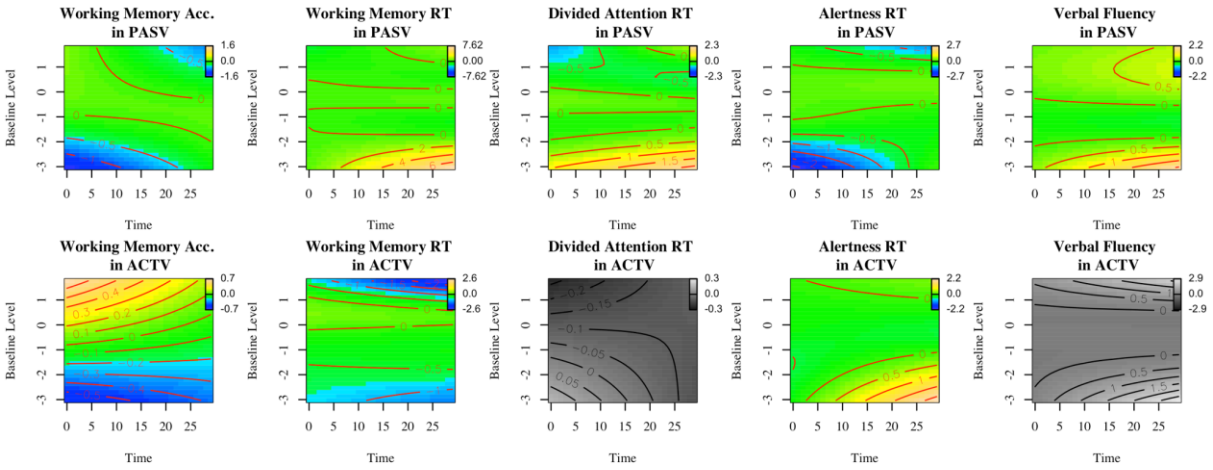
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.03	0.05	0.56	0.58	
	edf	Ref.df	F	p-value	
te(Time,initLevel):TaskssWM	13.75	16.76	19.53	< 0.001	***
te(Time,initLevel):TasksrtWM	5.47	6.42	60.47	< 0.001	***
te(Time,initLevel):TasksrtDivAtt	10.80	13.19	45.88	< 0.001	***
te(Time,initLevel):TasksrtAlert	12.46	14.54	35.83	< 0.001	***
te(Time,initLevel):TaskssRWT	7.73	10.01	39.93	< 0.001	***
te(Time,initLevel):isACTVissWM	4.00	4.00	3.38	< 0.01	**
te(Time,initLevel):isACTVisrtWM	9.01	9.64	4.59	< 0.001	***
te(Time,initLevel):isACTVisrtDivAtt	4.00	4.00	0.38	0.83	
te(Time,initLevel):isACTVisrtAlert	7.00	7.86	2.40	0.01	*
te(Time,initLevel):isACTVissRWT	8.03	8.95	1.65	0.10	
te(Time,initLevel):isPASVissWM	5.93	6.62	3.48	< 0.01	**
te(Time,initLevel):isPASVisrtWM	9.33	9.83	4.47	< 0.001	***
te(Time,initLevel):isPASVisrtDivAtt	8.07	8.99	2.04	0.03	*
te(Time,initLevel):isPASVisrtAlert	7.66	8.61	2.31	0.03	*
te(Time,initLevel):isPASVissRWT	8.01	8.88	3.36	0.00	***
s(Time,userCode)	96.85	547.00	1.33	< 0.001	***

Note: Terms not marked “isACTV” and "isPASV" refer to the reference level for group LANG. All other terms constitute binary curve smooths that capture the difference between trajectories LANG-PASV and LANG-ACTV.

The model shows that baseline levels had an effect on cognitive development that was significantly different between groups LANG and PASV and between LANG and ACTV for most of the cognitive measures. The interactions are visualized in Figure 5.

Figure 5

Tensor product smooth for the interaction of Time and Baseline Level per group and cognitive measure. Color coding is used to represent model predictions, with yellow indicating higher and blue representing lower cognitive scores. Plots in gray scale represent non-significant relationships.



As shown in Figure 5 and in line with the overall pattern, participants with baseline levels below -2SDs showed less improvement in WM accuracy when they were in the ACTV or PASV training conditions compared to the LANG training – an effect which, however, disappeared in group PASV after 20 weeks. A similar effect can be observed for Alertness RT in group PASV. At the same time, participants who had low baseline Alertness RT showed stronger improvement in the ACTV group after week 10 than subjects with comparable baselines in group LANG, and participants with above-average baselines improved their WM accuracy more if they were in group ACTV as compared to LANG. Interestingly, low-baseline participants from group LANG were outperformed increasingly over time by group PASV in terms of WM and Divided Attention RTs and Verbal Fluency.

DISCUSSION

The present study is the first longitudinal intervention study with dense measurements to investigate linear and non-linear far transfer effects of L2 learning to cognitive functions known to decline as a function of age (WM, alertness, divided attention, verbal fluency). In line with Antoniou et al. (2013) and Antoniou and Wright (2017), we hypothesized semi-computerized L2 training (LANG) in older adults to result in stronger transfer effects to

cognitive abilities than a semi-computerized strategy game training (ACTV) and a social/passive control condition (PASV). On average, group LANG manifested higher overall cognitive performance than group PASV throughout the training. However, our GAMM analysis provided no evidence that, at any time during the 30 weeks of training, L2 training resulted in an increased improvement of cognitive abilities compared to either ACTV or PASV, neither on the global level nor on the level of specific cognitive tasks. Importantly, since four of the five cognitive measures were RTs and count measures, improvement was possible in all three training groups, and there does not appear to be a ceiling effect in any of the five measures (see Online Supplement SF1). Using baseline levels as an additional predictor in order to control for differences in initial cognition and to assess potential individual differences in training gain, we found a significant interaction of baseline cognitive performance and time. The interaction suggests that individuals with low initial cognitive abilities (< -2SDs) manifested stronger cognitive improvement in the LANG than the ACTV or PASV conditions. In particular, the LANG training appeared to yield higher gains in WM accuracy for individuals with low initial performance, an effect which, however, disappeared after approximately 20 weeks. The other skills showed inconclusive or even contradictory effects. Interestingly, the overall baseline cognitive level was predicted by the amount of prior multilingualism, such that individuals with higher previous L2 experience (other than Spanish) had better baseline performance than those with little to none.

These findings provide evidence against cognitive benefits of L2 learning in old adulthood being as universal and profound as originally assumed (Antoniou et al. 2013). That is, at least compared to, for example, dance interventions, which have been shown to improve global cognition after as little as one hour per week over a time span of 24 weeks (Hackney et al. 2015; Meng et al. 2020). For the majority of our participants, improvement was independent of training type and can be characterized as a typical practice effect. At the same time, our results indicate a potential cognitive benefit of L2 training in learners with low baseline performance, since those type of individuals showed a stronger improvement in the LANG training condition than in the ACTV and PASV training conditions. In line with this, Wong et al. (2019a) found attentional resources in older adults suffering from mild cognitive impairment (MCI) to be improved through L2 and game training but not through music appreciation (i.e. PASV), which confirms a potential cognitive benefit of L2 training for this cohort, in particular. Tigka et al. (2019), however, did not find cognitive development in either general cognitive functioning, attention, verbal learning, memory, visuo-perceptual

ability or executive function to differ between older adult L2 learners suffering from MCI following an 18-months L2 training and a passive control group. Accordingly, for the present study, it has to be borne in mind that (1) at least between groups LANG and PASV, the beneficial effect of the L2 training effect disappeared after 20 weeks of training; (2) even though statistically significant, the effect was based on a small subsample of our learners and may not be generalizable, especially as this analysis was of an exploratory nature; and (3) the effect appeared to be limited to measures of WM and alertness. Therefore, future research will be necessary to assess the robustness of cognitive benefits in individuals with reduced cognitive abilities.

In contrast, cognitive development of older learners with high cognitive performance is likely to resemble that of younger adults, for whom cognitive improvement through L2 learning has been shown to be very limited even if conducted with extreme intensity. A study by Mårtensson and Lövdén (2011) investigated young conscript interpreters (M_{age} = 20) before and after their first three months of studies at an interpreter academy, with daily language classes from 08:00 to bedtime. While face-name associative memory was enhanced in the experimental group compared to students from other university classes, working memory, strategy-sensitive episodic memory and fluid intelligence were not. Hence, for individuals with high cognitive performance (independent of their age), a significant improvement of performance may be more difficult to start with and is unlikely to be achieved through regular L2 training, let alone a recreational course for beginners. As such, our findings corroborate earlier studies that looked at mean trends before and after 3 to 8 months of L2 training in older adults and did not find increased cognitive performance (Berggren et al. 2020; Bubbico et al. 2019; Ramos et al. 2016; Valis et al. 2019; Ware et al. 2017). What is novel about the present findings, however, is that we could not find any evidence either of cognitive development being different between L2 and non-L2 experimental groups at any given time point *during* the training, thus suggesting that there are also no temporally bound group*time interactions.

We argue that these findings are unlikely to be caused by flaws in the experimental design. While we did observe a selection bias in that group LANG outperformed group PASV from the beginning, (1) our models would still have detected differences in the steepness of the ensuing trajectories given the absence of a ceiling effect in any of the cognitive measures, and (2) this bias was included as a predictor in the second model and did not show group differences for the majority of participants either. Since participants were matched in terms of

background variables and socio-affect, this selection bias appears to be coincidental, and behavioral differences would likely have been exacerbated by randomly assigning participants to a 8-month training they are not motivated for (see dropout rate of 54% in Wong et al. 2019).

Hence, we concur with Ramos et al. (2016) and Berggren et al. (2020) in that, at least for executive functions, memory functions, attentional resources, processing speed and intelligence, the transfer effects of L2 learning are probably negligible or unreliable. At the same time, on a cross-sectional level, our results also showed that overall cognitive baseline performance correlated with the prior degree of multilingualism such that participants with higher knowledge of other languages had better cognitive abilities at study onset. At first glance, this might look like evidence supporting the idea of a bilingual advantage (for a discussion see e.g. Lehtonen et al. 2018; Ware et al. 2020; Monnier et al. 2021), suggesting that lifelong experience with more than one language indeed boosts cognitive resilience, whereas an entry-level L2 course aimed at older learners might be less effective in doing so. However, we did not collect a sufficient number of background variables to ascertain the unique contribution of prior multilingualism to cognitive performance of our participants at baseline.

Finally, while the number of our repeated measures was adequate to detect variability across time points, the between-subject sample size was likely too small to reach generalizability from sample to population –which, however, was also not a priority in the present study. Rather, we would argue with Van Geert (2011) that "a truly general theory of development processes is one that can be 'individualized' – it can generate theory-based descriptions of individual trajectories in a nontrivial sense" (276). In that sense, while our findings may only have an indirect bearing on the larger population of older adult L2 learners, they also do not lend support in favor of Antoniou et al.'s theory (2017) of cognitive transfer effects of L2 learning. It is possible that a larger between-subject sample size would show significant differences in overall cognitive development, particularly if effect sizes are small. Again, however, it is worth asking if effect sizes too small to detect in samples of 90 participants per group (i.e. $d < 0.2$; see Berggren et al. 2020) would be impactful enough to allow us to promote L2 learning as a cognitive training intervention in old adulthood.

CONCLUSION

The results of the present dense-longitudinal intervention study suggest that transfer effects of a 8-months entry-level L2 course for beginners on cognitive capacities in older adults are either negligible or limited to learners with low cognitive baselines. Group differences in cognitive improvement between entry-level L2 learners, a strategy game group and a social passive control group could neither be found on the overall level nor during any period of the 30-week training, thus suggesting that temporal benefits are also not present. At the same time, baseline cognitive performance was predicted by prior multilingualism, so that we cannot rule out that lifelong experience with other languages may have a positive effect on overall cognitive abilities. While L2 learning in old adulthood may not be suitable as a way of staving off age-related cognitive declines in individuals with high cognitive abilities, it is still invaluable as a personal challenge, a way of encouraging communication, a motivation to travel and enhance mobility, and a means of finding integration within a multilingual society.

REFERENCES

Antoniou, M., Gunasekera, G. M., & Wong, P. C. M. (2013). Foreign language training as cognitive therapy for age-related cognitive decline: A hypothesis for future research. *Neuroscience and Biobehavioral Reviews*, *37*(10), 2689–2698. https://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*. https://doi.org/10.3389/fpsyg.2017.02217

Bak, T. H., Long, M. R., Vega-Mendoza, M., & Sorace, A. (2016). Novelty, Challenge, and Practice: The Impact of Intensive Language Learning on Attentional Functions. *PLOS ONE*, *11*(4), e0153485. https://doi.org/10.1371/journal.pone.0153485

Bamidis, P. D., Vivas, A. B., Styliadis, C., Frantzidis, C., Klados, M., Schlee, W., Siountas, A., & Papageorgiou, S. G. (2014). A review of physical and cognitive interventions in aging. *Neuroscience & Biobehavioral Reviews*, *44*, 206–220. https://doi.org/10.1016/j.neubiorev.2014.03.019

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychology and Aging*, *23*(4), 765–777. https://doi.org/10.1037/a0013494

Berggren, R., Nilsson, J., Brehmer, Y., Schmiedek, F., & Lövdén, M. (2020). Foreign language learning in older age does not improve memory or intelligence: Evidence from a randomized controlled study. *Psychology and Aging*, *35*(2), 212–219. https://doi.org/10.1037/pag0000439

Binder, J. C., Martin, M., Zöllig, J., Röcke, C., Mérillat, S., Eschen, A., Jäncke, L., & Shing, Y. L. (2016). Multi-domain training enhances attentional control. *Psychology and Aging*, *31*(4), 390–408. https://doi.org/10.1037/pag0000081

Bubbico, G., Chiacchiaretta, P., Parenti, M., di Marco, M., Panara, V., Sepede, G., Ferretti, A., & Perrucci, M. G. (2019). Effects of Second Language Learning on the Plastic Aging Brain: Functional Connectivity, Cognitive Decline, and Reorganization. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00423

Chiew, K. S., & Braver, T. S. (2013). Temporal Dynamics of Motivation-Cognitive Control Interactions Revealed by High-Resolution Pupillometry. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00015

Christensen, H. (2001). What Cognitive Changes can be Expected with Normal Ageing? *Australian & New Zealand Journal of Psychiatry*, *35*(6), 768–775. https://doi.org/10.1046/j.1440-1614.2001.00966.x

Fagan, E. S., & Pihlstrøm, L. (2017). Genetic risk factors for cognitive decline in Parkinson's disease: A review of the literature. *European Journal of Neurology*, *24*(4), 561-e20. https://doi.org/10.1111/ene.13258

Hayat, S. A., Luben, R., Dalzell, N., Moore, S., Anuj, S., Matthews, F. E., Wareham, N., Brayne, C., & Khaw, K.-T. (2016). Cross Sectional Associations between Socio-Demographic Factors and Cognitive Performance in an Older British Population: The European Investigation of Cancer in Norfolk (EPIC-Norfolk) Study. *PLOS ONE*, *11*(12), e0166779. https://doi.org/10.1371/journal.pone.0166779

Issa, B. I., & Morgan-Short, K. (2018). Effects of external and internal attentional manipulation on second language grammar development. *Studies in Second Language Acquisition*, 1–29. https://doi.org/10.1017/S027226311800013X

Kliesch, M., & Pfenninger, S. E. (2021). Cognitive and Socioaffective Predictors of L2 Microdevelopment in Late Adulthood: A Longitudinal Intervention Study. *The Modern Language Journal*, *105*(1), 237–266. https://doi.org/10.1111/modl.12696

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*(4), 861–883. https://doi.org/10.3758/s13423-013-0565-2

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2005). Social Participation Attenuates Decline in Perceptual Speed in Old and Very Old Age. *Psychology and Aging*, *20*(3), 423–434. https://doi.org/10.1037/0882-7974.20.3.423

Mansens, D., Deeg, D. J. H., & Comijs, H. C. (2018). The association between singing and/or playing a musical instrument and cognitive functions in older adults. *Aging & Mental Health*, *22*(8), 970–977. https://doi.org/10.1080/13607863.2017.1328481

Mårtensson, J., & Lövdén, M. (2011). Do Intensive Studies of a Foreign Language Improve Associative Memory Performance? *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00012

Martin, M., & Hofer, S. M. (2004). Intraindividual Variability, Change, and Aging: Conceptual and Analytical Issues. *Gerontology*, *50*(1), 7–11. https://doi.org/10.1159/000074382

Mora, F. (2013). Successful brain aging: Plasticity, environmental enrichment, and lifestyle. *Dialogues Clin Neurosci*, *15*(1), 45–52.

Nasreddine, Z. S., Phillips, N. A., BÃ©dirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment: MOCA: A BRIEF SCREENING TOOL FOR MCI. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Neupert, S. D., & Allaire, J. C. (2012). I think I can, I think I can: Examining the within-person coupling of control beliefs and cognition in older adults. *Psychology and Aging*, *27*(3), 742–749. https://doi.org/10.1037/a0026447

Neupert, S. D., Almeida, D. M., Mroczek, D. K., & Spiro, A. (2006). Daily stressors and memory failures in a naturalistic setting: Findings from the va normative aging study. *Psychology and Aging*, *21*(2), 424–429. https://doi.org/10.1037/0882-7974.21.2.424

Park, D. C., & Reuter-Lorenz, P. (2009). The adaptive brain: Aging and neurocognitive scaffolding. *Annual Review of Psychology*, *60*(November 2008), 173–196. https://doi.org/10.1146/annurev.psych.59.103006.093656

Pfenninger, S. E., & Singleton, D. (2019). A critical review of research relating to the learning, use and effects of additional and multiple languages in later life. *Language Teaching*, *52*(4), 419–449. https://doi.org/10.1017/S0261444819000235

Ramos, S., Fernández García, Y., Antón, E., Casaponsa, A., & Duñabeitia, J. A. (2016). Does learning a language in the elderly enhance switching ability? *Journal of Neurolinguistics*. https://doi.org/10.1016/j.jneuroling.2016.09.001

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society : JINS*, *16*(5), 754–760. https://doi.org/10.1017/S1355617710000706

Strauss, E., Macdonald, S. W. S., Hunter, M., Moll, A., & Hultsch, D. F. (2002). Intraindividual variability in cognitive performance in three groups of older adults: Cross-domain links to physical status and self-perceived affect and beliefs. *Journal of the International Neuropsychological Society*, *8*(7), 893–906. https://doi.org/10.1017/S1355617702870035

Tigka, E., Kazis, D., Tsolaki, M., Bamidis, P., Papadimitriou, M., & Kassapi, E. (2019). *FL learning could contribute to the enhancement of cognitive functions in MCI older adults*. *8*(2), 24.

Valis, M., Slaninova, G., Prazak, P., Poulova, P., Kacetl, J., & Klimova, B. (2019). Impact of Learning a Foreign Language on the Enhancement of Cognitive Functions Among Healthy Older Population. *Journal of Psycholinguistic Research*, *48*(6), 1311–1318. https://doi.org/10.1007/s10936-019-09659-6

Van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using GAMM, R package version 2.3 [computer program]*.

Ware, C., Damnee, S., Djabelkhir, L., Cristancho, V., Wu, Y.-H., Benovici, J., Pino, M., & Rigaud, A.-S. (2017). Maintaining Cognitive Functioning in Healthy Seniors with a Technology-Based Foreign Language Program: A Pilot Feasibility Study. *Frontiers in Aging Neuroscience*, *9*. https://doi.org/10.3389/fnagi.2017.00042

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wilson, R. S., Bennett, D. A., Bienias, J. L., Aggarwal, N. T., De Leon, C. M., Morris, M. C., Schneider, J. A., & Evans, D. A. (2002). Cognitive activity and incident AD in a population-based sample of older persons. *Neurology*, *59*(12), 1910–1914. https://doi.org/10.1212/01.WNL.0000036905.59156.A1

Wong, P. C. M., Ou, J., Pang, C. W. Y., Zhang, L., Tse, C. S., Lam, L. C. W., & Antoniou, M. (2019a). Language Training Leads to Global Cognitive Improvement in Older Adults: A Preliminary Study. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2411–2424. https://doi.org/10.1044/2019_JSLHR-L-18-0321

Wong, P. C. M., Ou, J., Pang, C., Zhang, L., Tse, C., Lam, L., & Antoniou, M. (2019b). *Foreign language learning as potential treatment for mild cognitive impairment*. *25*(5), 3.

Zuelsdorff, M. L., Koscik, R. L., Okonkwo, O. C., Peppard, P. E., Hermann, B. P., Sager, M. A., Johnson, S. C., & Engelman, C. D. (2019). Social support and verbal interaction are differentially associated with cognitive function in midlife and older age. *Aging, Neuropsychology, and Cognition*, *26*(2), 144–160. https://doi.org/10.1080/13825585.2017.1414769

Supplementary Material

Figure SF1

Distribution of z-scores within each cognitive variable. Violin plots show the probability density of the data at different values, while box plots indicate the median and the respective quartiles. The rhombus at the center represents the mean, and blue dots are outliers, defined as values above 1.5 IQR from the median.

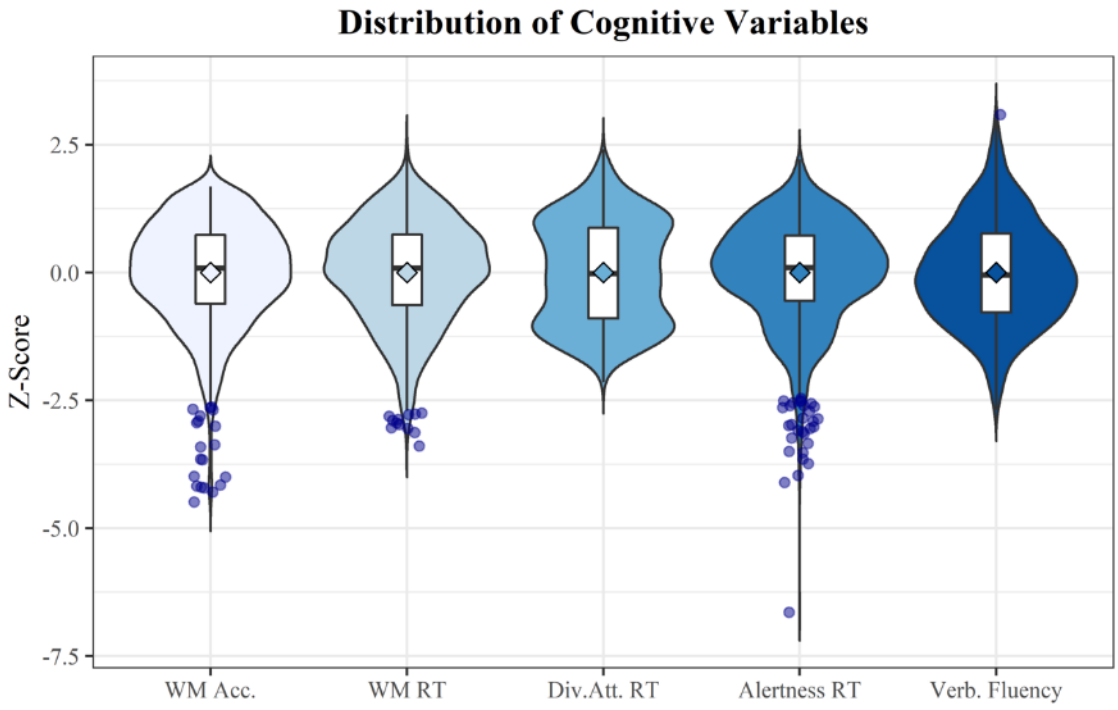


Figure SF2

Smooths over time per individual for each cognitive measure. RTs were log transformed and inverted to make interpretation consistent across all measures (higher = better).

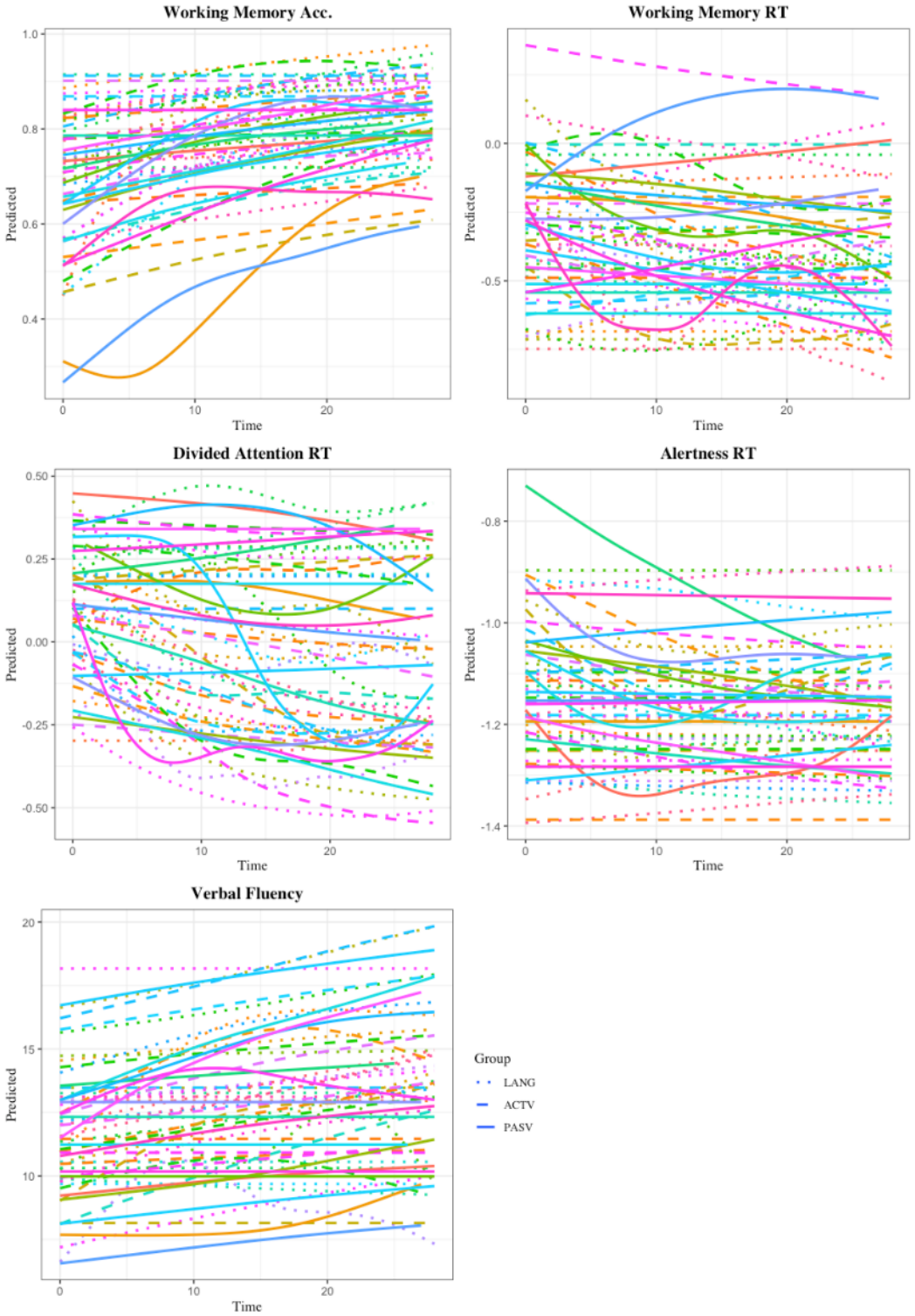


Figure SF3

Frequency distribution of normalized baseline levels per task and subject.

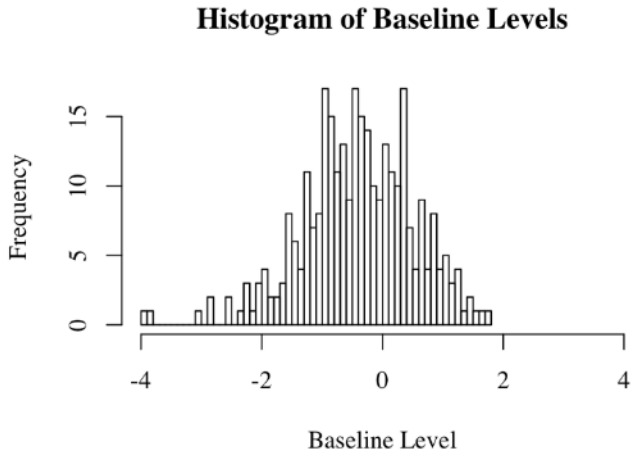


Table ST1

Model Fit of the GAMM Predicting Intercept and Development of Each Cognitive Tasks Based on Training Type

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.17	0.11	1.59	0.11	
TasksrtWM	0.01	0.08	0.12	0.91	
TasksrtDivAtt	-0.12	0.08	-1.56	0.12	
TasksrtAlert	-0.11	0.08	-1.37	0.17	
TaskssRWT	-0.09	0.08	-1.09	0.28	
isACTVissWM01	-0.23	0.18	-1.27	0.21	
isACTVisrtWM01	-0.43	0.18	-2.42	0.02	*
isACTVisrtDivAtt01	0.07	0.18	0.37	0.71	
isACTVisrtAlert01	-0.08	0.18	-0.47	0.64	
isACTVissRWT01	-0.12	0.18	-0.67	0.50	
isPASVissWM01	-0.55	0.18	-3.00	0.00	**
isPASVisrtWM01	-0.29	0.18	-1.61	0.11	
isPASVisrtDivAtt01	-0.39	0.18	-2.17	0.03	*
isPASVisrtAlert01	-0.23	0.18	-1.27	0.20	
isPASVissRWT01	-0.29	0.18	-1.60	0.11	
	edf	Ref.df	F	p-value	
s(Time):TaskssWM	3.18	4.06	12.08	< 0.001	***
s(Time):TasksrtWM	1.00	1.00	6.44	0.01	*
s(Time):TasksrtDivAtt	2.42	3.08	6.75	< 0.001	***
s(Time):TasksrtAlert	1.00	1.00	0.60	0.44	

s(Time):TaskssRWT	2.03	2.57	6.22	< 0.01	**
s(Time):isACTVissWM01	1.00	1.00	0.12	0.73	
s(Time):isACTVisrtWM01	1.00	1.00	0.01	0.91	
s(Time):isACTVisrtDivAtt01	1.00	1.00	0.02	0.89	
s(Time):isACTVisrtAlert01	1.82	2.30	0.76	0.47	
s(Time):isACTVissRWT01	1.00	1.00	0.08	0.77	
s(Time):isPASVissWM01	1.00	1.00	1.15	0.28	
s(Time):isPASVisrtWM01	1.00	1.00	0.00	0.97	
s(Time):isPASVisrtDivAtt01	1.00	1.00	0.14	0.71	
s(Time):isPASVisrtAlert01	2.11	2.68	1.15	0.34	
s(Time):isPASVissRWT01	1.69	2.11	0.56	0.57	
s(Time,subject)	54.11	547.00	1.47	< 0.001	***

Note: Terms not marked “isACTV0” and "isPASV0" refer to the reference level for group LANG. All other terms constitute ordered difference smooths that capture the difference between trajectories LANG-PASV and LANG-ACTV.

Table ST2

Mean, Median, Standard Deviation, Minimum and Maximum Scores for Individual Tests Over All Measurement Points (Before Standardization)

Variable	Group	N	Mean	Median	SD	Min	Max
Working Memory Acc.	PASV	455	0.72	0.74	0.15	0.16	1.00
	ACTV	489	0.77	0.78	0.14	0.20	1.00
	LANG	805	0.80	0.80	0.12	0.39	1.00
Working Memory RT [log]	PASV	450	-0.36	-0.37	0.25	-0.98	0.33
	ACTV	489	-0.32	-0.36	0.27	-1.02	0.41
	LANG	802	-0.44	-0.46	0.25	-0.98	0.49
Divided Attention Acc.	PASV	452	6.27	6.78	1.61	1.49	8.22
	ACTV	489	6.87	7.01	1.29	1.57	8.22
	LANG	796	6.93	7.03	1.34	0.23	8.22
Divided Attention RT [log]	PASV	452	0.07	0.11	0.26	-0.55	0.53
	ACTV	489	-0.03	-0.05	0.25	-0.58	0.44
	LANG	805	-0.02	-0.03	0.26	-0.62	0.56
Alertness RT [log]	PASV	453	-1.15	-1.17	0.14	-1.39	-0.29
	ACTV	489	-1.17	-1.16	0.12	-1.46	-0.67
	LANG	809	-1.17	-1.19	0.14	-1.42	-0.62
Verbal Fluency	PASV	451	12.11	12.10	3.44	4.10	22.17
	ACTV	488	12.60	12.19	3.10	4.19	21.10
	LANG	806	13.01	13.10	3.05	4.19	22.51
Socio-Affect	PASV	440	78.39	81.02	17.19	16.80	100.00
	ACTV	457	76.05	79.88	20.82	11.59	100.00
	LANG	785	74.26	77.73	18.53	8.72	100.00

Note: Divided Attention accuracy was removed from further analysis due to ceiling effects.

Table ST3

Model Fit of the GAMM Predicting Intercept and Development of Overall Cognition Based on the Interaction Between Time and Baseline Performance per Training Type

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.03	0.05	0.56	0.58	
	edf	Ref.df	F	p-value	
te(Time,initLevel)	16.18	18.69	78.35	< 0.001	***
te(Time,initLevel):isACTV	9.42	9.86	3.79	< 0.001	***
te(Time,initLevel):isPASV	6.92	7.91	2.00	0.04	*
s(Time,subject)	85.60	547.00	1.03	< 0.001	***

Note: Terms not marked “isACTV” and "isPASV" refer to the reference level for group LANG. All other terms constitute binary curve smooths that capture the difference between trajectories LANG-PASV and LANG-ACTV.

Model Code SC1

```
model = bam(score ~
        Group +
        s(Time) +
        s(Time, by = Group) +
        s(Time, subject, bs = "fs", m = 1),
      data = df_overall, family = "scat", discrete = T, nthreads = 7)
```

Model Code SC2

```
model = bam(score ~
        isACTV0 +
        isPASV0 +
        s(Time) +
        s(Time, by = isACTV0) +
        s(Time, by = isPASV0) +
        s(Time, subject, bs = "fs", m = 1),
      data = df_overall, rho = r1,  AR.start = df_overall$start.event,
      family = "scat", discrete = T, nthreads = 7)
```
[5]

[5] Autocorrelation was determined by fitting the same model without the autocorrelation pattern and using the ACF at lag = 1.

Model Code SC3

```
model = bam(score ~
        # smooths for reference level LANG
        s(Time, by = Tasks) +
        Tasks +

        # smooths for ACTV as ordered factors
        s(Time, by = isACTVissWM0) +
        isACTVissWM0 +
        s(Time, by = isACTVisrtWM0) +
        isACTVisrtWM0 +
        s(Time, by = isACTVisrtDivAtt0) +
        isACTVisrtDivAtt0 +
        s(Time, by = isACTVisrtAlert0) +
        isACTVisrtAlert0 +
        s(Time, by = isACTVissRWT0) +
        isACTVissRWT0 +

        # smooths for PASV as ordered factors
        s(Time, by = isPASVissWM0) +
        isPASVissWM0 +
        s(Time, by = isPASVisrtWM0) +
        isPASVisrtWM0 +
        s(Time, by = isPASVisrtDivAtt0) +
        isPASVisrtDivAtt0 +
```

```
    s(Time, by = isPASVisrtAlert0) +
    isPASVisrtAlert0 +
    s(Time, by = isPASVissRWT0) +
    isPASVissRWT0 +

    # random effects for subject and task in LANG
    s(Time, subject, bs = "fs", m = 1),
    data = df_overall, family = "scat", rho = r1,  AR.start = df_overall$start.event,
  discrete = T, nthreads = 7)
```
[6]

[6] Autocorrelation was determined by fitting the same model without the autocorrelation pattern and using the ACF at lag = 1.

Model Code SC4

```
model = bam(score ~
          te(Time, initLevel) +
          te(Time, initLevel, by = isACTV) +
          te(Time, initLevel, by = isPASV) +
          s(Time, subject, bs = "fs", m = 1),
        data = df_overall, rho = r1,  AR.start = df_overall$start.event,
        family = "scat", discrete = T, nthreads = 7)
```
[7]

[7] Autocorrelation was determined by fitting the same model without the autocorrelation pattern and using the ACF at lag = 1.

Model Code SC5

```
# remove outliers
df_overall_reduced = df_overall %>%
  filter(initLevel > -3.5 )


m2.alt = bam(score ~
               isACTV0 +
               isPASV0 +
               s(Time) +
               s(Time, by = isACTV0) +
               s(Time, by = isPASV0) +
               s(Time, subject, bs = "fs", m = 1),
             data = df_overall_reduced, rho = r1,  AR.start = df_overall_reduced$start.event,
             discrete = T, select = T, nthreads = 7) [8]


m4.alt = bam(score ~
                 te(Time, initLevel) +
                 te(Time, initLevel, by = isACTV) +
                 te(Time, initLevel, by = isPASV) +
                 s(Time, subject, bs = "fs", m = 1),
               data = df_overall_reduced, rho = r1,  select = T, AR.start =
df_overall_reduced$start.event,
               discrete = T, nthreads = 7) [4]
```

[8] Autocorrelation was determined by fitting the same model without the autocorrelation pattern and using the ACF at lag = 1.

```
# comparison based on REML

compareML(m2.alt, m4.alt)
```

Model Code SC6

```
data = all_data %>%
  filter(initLevel > -3.5 )


model = bam(score ~
              te(Time, initLevel, by = Tasks) +
              te(Time, initLevel, by = isACTVissWM) +
              te(Time, initLevel, by = isACTVisrtWM) +
              te(Time, initLevel, by = isACTVisrtDivAtt) +
              te(Time, initLevel, by = isACTVisrtAlert) +
              te(Time, initLevel, by = isACTVissRWT) +

              te(Time, initLevel, by = isPASVissWM) +
              te(Time, initLevel, by = isPASVisrtWM) +
              te(Time, initLevel, by = isPASVisrtDivAtt) +
              te(Time, initLevel, by = isPASVisrtAlert) +
              te(Time, initLevel, by = isPASVissRWT) +
              s(Time, subject, bs = "fs", m = 1),
            data = data, rho = r1,  AR.start = data$start.event,
            discrete = T, nthreads = 7)
```