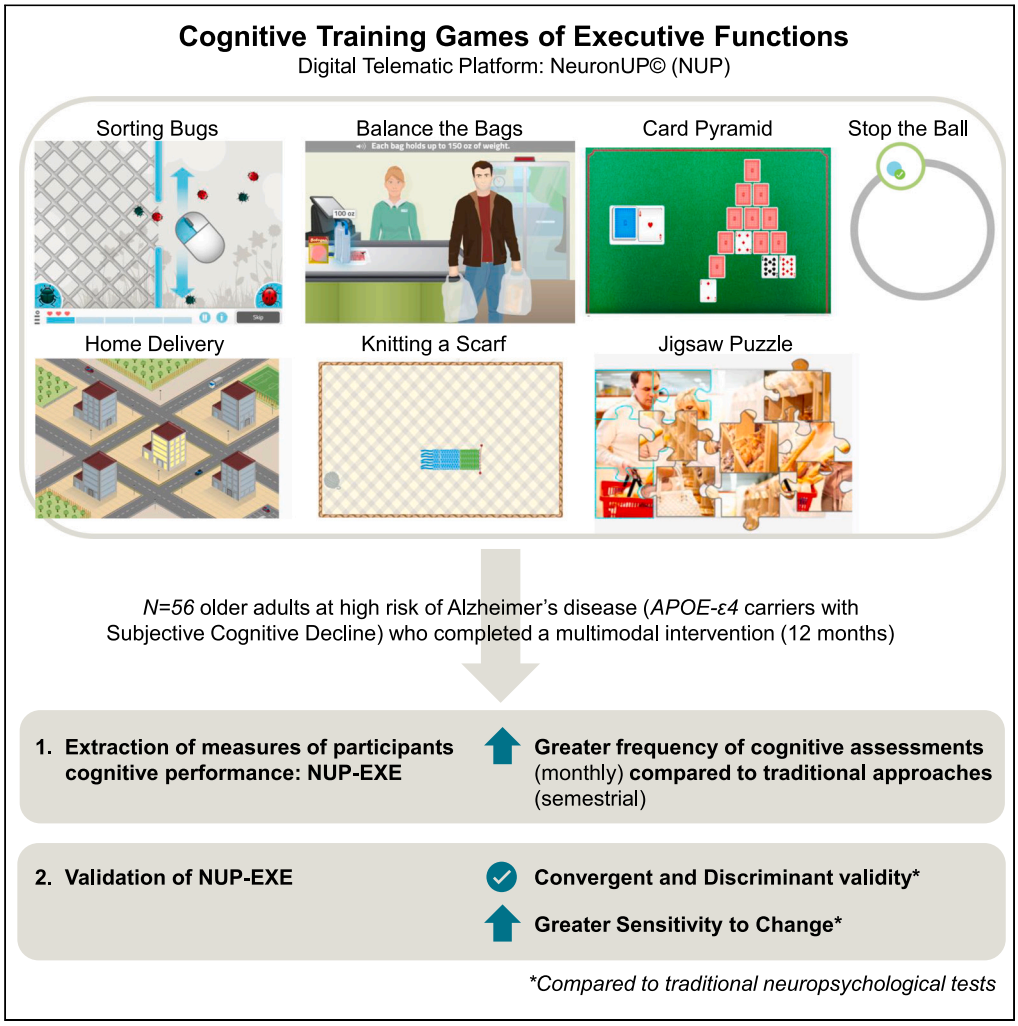
iScience

CellPress
OPEN ACCESS

Article

Intensive assessment of executive functions derived from performance in cognitive training games



Natalia Soldevila-Domenech, Ilario De Toma, Laura Forcano, ..., Antonio Verdejo-Garcia, Rafael de la Torre, PENSA Study Group

rtorre@imim.es

Highlights

Practice effects limit the intensive evaluation of cognition

Cognitive training games (CTG) could minimize practice effects

We tested the psychometric properties of CTG for intensively measuring cognition

Cognitive response profiles to a preventive intervention were evaluated with CTG

Soldevila-Domenech et al., iScience 26, 106886
June 16, 2023
https://doi.org/10.1016/j.isci.2023.106886

iScience

CellPress
OPEN ACCESS

Article

Intensive assessment of executive functions derived from performance in cognitive training games

Natalia Soldevila-Domenech,[1,2] Ilario De Toma,[1] Laura Forcano,[1,3] Patrícia Diaz-Pellicer,[1,2] Aida Cuenca-Royo,[1] Beatriz Fagundo,[1] Thais Lorenzo,[1,2] Maria Gomis-Gonzalez,[1] Gonzalo Sánchez-Benavides,[1,4,5] Karine Fauria,[1,4,5] Carolina Sastre,[6] Íñigo Fernandez De Piérola,[6] José Luis Molinuevo,[1,4] Antonio Verdejo-Garcia,[7] Rafael de la Torre,[1,2,3,9,*] and PENSA Study Group[8]

SUMMARY

Traditional neuropsychological tests accurately describe the current cognitive state but fall short to characterize cognitive change over multiple short time periods. We present an innovative approach to remote monitoring of executive functions on a monthly basis, which leverages the performance indicators from self-administered computerized cognitive training games (NUP-EXE). We evaluated the measurement properties of NUP-EXE in N = 56 individuals (59% women, 60–80 years) at increased risk of Alzheimer's disease (*APOE*-ϵ4 carriers with subjective cognitive decline) who completed a 12-month multimodal intervention for preventing cognitive decline. NUP-EXE presented good psychometric properties and greater sensitivity to change than traditional tests. Improvements in NUP-EXE correlated with improvements in functionality and were affected by participants' age and gender. This novel data collection methodology is expected to allow a more accurate characterization of an individual's response to a cognitive decline preventive intervention and to inform development of outcome measures for a new generation of intervention trials.

INTRODUCTION

Traditional neuropsychological assessment methods have high specificity and sensitivity for assessing the current cognitive state[1–3] but suffer from several limitations due to (i) the time and expenses associated with testing[4]; (ii) the fact that testing sessions can be perceived as intrusive and unnatural and thus trigger stress and hinder attention and motivation[5]; and (iii) the difficulty to reliably administer repeated assessments due to ubiquitous practice effects.[6,7] As a result, cognition is usually evaluated in few occasions over the course of longitudinal studies or clinical trials, typically at baseline, mid-term, and at the end of the study. Obtained results are aggregated statistical data assuming that (i) all participants respond in the same direction (i.e., all are responders or non-responders) and (ii) changes over time follow the same pattern (i.e., thus missing if peak changes happened at a time different from predefined assessments and there is no further improvement despite a continued intervention). An innovative approach to overcome the limitations of traditional assessments would be the incorporation of computer-based algorithms to adjust difficulty to capacity and thus prevent practice effects,[8] which would enable a more frequent sampling of the cognitive performance.

Dementia is the fourth leading cause of death among 70-year-old people,[9,10] being Alzheimer's disease (AD) the most common cause (60–70% of cases).[10] AD has a broad spectrum of clinical manifestations that span from clinically asymptomatic to severely impaired.[11] Therefore, like other neurodegenerative and neurodevelopmental disorders,[12] AD should not only be viewed with discrete and defined clinical stages but also as a multifaceted process moving along a seamless continuum.[13] Prevention of AD has emerged as the best therapeutic opportunity given the long preclinical phase and the known modifiable risk factors that can be addressed in preventive interventions, including diet, physical activity, and cognitive training.[14] Individuals with cognitive performance within normal values experiencing subjective cognitive decline (SCD) are at increased risk of AD[15,16] and are considered the ideal target population where to

[1]Neurosciences Research Programme, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

[2]Department of Medicine and Life Sciences, Pompeu Fabra University, Barcelona, Spain

[3]CIBER de Fisiopatología de la Obesidad y la Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain

[4]BarcelonaBeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain

[5]Centro de Investigación Biomédica en Red de Fragilidad y Envejecimiento Saludable (CIBERFES), Madrid, Spain

[6]NeuronUP SL, Logroño, La Rioja, Spain

[7]School of Psychological Sciences and Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia

[8]Consortia authorship

[9]Lead contact

*Correspondence: rtorre@imim.es

https://doi.org/10.1016/j.isci.2023.106886

perform preventive interventions.[17] However, the sensitivity of cognitive tests to detect subtle cognitive decrements underlying neuropathological burden in early preclinical stages of AD is questionable.[18,19] Moreover, the heterogeneity of AD from a genetic, pathophysiological, and clinical viewpoint suggests that preventive interventions are unlikely to succeed using a "one size fits all" approach.[20–23] To fully release the potential of these preventive interventions, we need to understand the inter-individual variability in treatment response (e.g., how and when cognitive and related functional changes occur) and then tailor interventions to individual needs.

New digital cognitive assessment tools are being developed for detecting early cognitive impairment in preclinical stages of AD.[24,25] These tools use computers, tablets, smartphones, and novel approaches (e.g., speech analysis, eye tracking, and virtual reality), and some of them are remotely administered without supervision.[24] However, they are mainly focused on detecting subtle cognitive changes to distinguish between cognitively normal subjects and those presenting biomarker-confirmed preclinical AD.[24,25] Therefore, in the context of multimodal interventions in lifestyle factors of SCD subjects, there is a lack of tools for longitudinally monitoring cognitive changes during the course of a preventive intervention.

An innovative approach for remote and repeatable cognitive testing would be the extraction of data performance on some cognitive tasks from web-based cognitive training games. Cognitive training refers to the repeated practice on standardized cognitive tasks designed to train specific cognitive abilities. It has been used with success in the improvement of cognitive functioning in old adults.[26–30] Cognitive training games have some characteristics that make them good candidates for the continuous evaluation of cognition. Because they are computerized, they allow an easy administration and data collection, with better standardization and increased accuracy of timing presentation of stimuli and response latencies. Cognitive training games can also introduce novel elements and alternate sequences to minimize learning effects, and they can adapt the testing difficulty to the baseline performance of each individual. Their remote and unsupervised administration increases trial efficiency and reduces the intrusiveness of testing sessions since they are performed in participant's familiar environment. In addition, cognitive training games can simulate real-world situations, increasing the ecological validity of assessment. Training sessions are typically of short duration (20–30 min), and the explored cognitive functions can alternate with sessions, which reduce participant's fatigue usually observed after long testing sessions. Finally, perhaps the most relevant advantage is the intensive longitudinal data collection (e.g., weekly or monthly), which gives the possibility of defining accurate cognitive trajectories to identify which individuals respond and not respond to interventions, as well as, to understand when cognitive changes occur within the time course of the study. However, this approach also faces challenges such as the adequate validity of the cognitive measures, the maintenance of participant engagement in the long run, or the lack of access or technical skills for using digital devices.[24]

This study aims to evaluate the acceptability, reliability, validity, and sensitivity to change of an innovative approach to the assessment of executive functions that leverages the performance indicators from self-administered computerized cognitive training games in individuals *APOE*-ε4 carriers meeting criteria of SCD following a preventive intervention for cognitive decline. We hypothesized that improvements in executive functions derived from performance in cognitive training games will be influenced by the age, gender,[31,32] years of education, and cognitive reserve[33,34] of participants and will correlate with improvements in daily living activities and quality of life.[35,36] We also postulated that, if the performance in cognitive training tasks is an accurate measure of cognitive performance, their metrics should enable the evaluation of the inter-individual variability in treatment response after preventive interventions for AD.

RESULTS

A total of 8 NeuronUP (NUP) games that targeted executive functions were analyzed (Table 1).

Sample characteristics

The ratio of women/men was 33/23, the mean ± SD age was 67.0 ± 4.7 years, and most participants (64.3%) had university or higher educational level (Table S1). Regarding SCD characteristics, 71.1% had an informant who corroborated the perception of cognitive decline, 50.0% had consulted a physician for the SCD, and almost all had memory complaints (98.2%), followed by impairment of language (71.4% in total, 84.8% in women and 52.2% in men) and concentration (62.5%). Moreover, over 80% presented 5 or more SCD-plus criteria, including (i) memory complaints rather than other domains of cognition, (ii) onset of

Table 1. Description of NUP cognitive training games of executive functions used in the cognitive training intervention and scoring system created for the evaluation of executive functions

Game	Pre-specified cognitive domains[a]	Description	Game specifications when increasing phase number[b]	Min-Max phase[c]	Scoring[d]
Sorting Bugs	Planning Hemineglect Processing Speed Selective Attention Sustained Attention	Consists of reorganizing the moving elements (bugs) by placing each type on the side where they belong	There are more bugs, they move faster, and the tool for reorganizing bugs is smaller	[1–9]	$Score = mode\,(phase\ n^o) + A - B$ $A = (0.2 \times n^o\ of\ passed\ exercises)$ $B = (0.005 \times mean\,(seconds))$
Balance the Bags	Working Memory Flexibility Planning Reasoning	Based on the weight of different products; the user has to put them into bags distributing their weight equally. There are two types of errors: when the user balances the bags wrongly and when bags are broken (if a bag is broken three times, the exercise is failed). The performance in each screen is qualified under 5 categories: perfect, regular, bad, very bad, and null.	There are more products and bags. Weights are also more complicated numbers	[1–9]	$Score = mode\,(phase\ n^o) + A - B$ $A = (0.2 \times n^o\ of\ passed\ exercises)$ $B = (0.030 \times n^o\ exercises\ qualified\ ase\ regular) + (0.063 \times n^o\ exercises\ qualified\ as\ bad) + (0.094 \times n^o\ exercises\ qualified\ as\ very\ bad) + (0.125 \times n^o\ exercises\ qualified\ as\ null)$
Home Delivery	Working Memory Episodic Memory	Consists of remembering the order in which the buildings light up to later reproduce it in reverse order	The number of illuminated buildings increases (the sequence is longer)	[1–9]	$Score = mode\,(phase\ n^o) + A - B$ $A = (0.2 \times n^o\ of\ passed\ exercises)$ $B = (0.005 \times mean\,(seconds))$
Card Pyramid	Flexibility Alternating Attention Hemineglect Planning Selective Attention	Consists of arranging the cards that appear in either increasing or decreasing numerical order	The number of cards increases, so there are more valid options and the complexity to organize cards increases	[1–6]	$Score = mode\,(phase\ n^o) + A - B$ $A = (0.2 \times n^o\ of\ passed\ exercises)$ $B = (0.005 \times mean\,(seconds))$
Knitting a Scarf	Planning Processing Speed Spatial Relation Sustained Attention	Consists of catching all the balls of yards that appear without hitting anything.	The scarf moves faster and faster, and its length increases	[1–12]	$Score = mode\,(phase\ n^o) + A$ $A = (0.2 \times n^o\ of\ passed\ exercises)$

(*Continued on next page*)

Table 1. *Continued*

Game	Pre-specified cognitive domains[a]	Description	Game specifications when increasing phase number[b]	Min-Max phase[c]	Scoring[d]
Jigsaw Puzzle	Visuoconstructive Praxis Planning Spatial relation	Consists of connecting the pieces until they make a picture	The number of puzzle pieces increases, the number of clues decreases, and pieces can appear rotated	[1–9]	*Score* = *mode* (*phase nº*) + *A* − *B* *A* = (0.2 × *nº of passed exercises*) *B* = (0.005 × *mean* (*seconds*))
Déjà vu	Working Memory Sustained Attention	Consists of looking closely at all the elements and finding those appearing more than once	The complexity and the number of elements to memorize increase. Elements can be in motion	[1–12]	*Score* = *mode* (*phase nº*) + *A* − *B* *A* = (0.2 × *nº of passed exercises*) *B* = (0.1 × *nº of errors*)
Stop the Ball	Spatial Relation Inhibition Planning Processing Speed	Consists of calculating the exact moment when the element should pass through a specific place	The ball moves faster and faster	[1–12]	*Score* = *mode* (*phase nº*) + *A* − *B* *A* = (0.2 × *nº of passed exercises*) *B* = (0.005 × *sum* (*seconds*))

[a]The first cognitive domain is the main one (in bold) and on the basis of which the game is leveled. The rest of cognitive domains are listed in alphabetical order.
[b]See details in Supplementary Methods Tables S1–S8.
[c]The phase of a game indicates the difficulty level. Higher phase number indicates increased difficulty level. All games were programmed to start at phase 3 except Balance the Bags and Card Pyramid that started at phase 2.
[d]*Mode* = most frequently played difficulty level each month. *Phase nº* = difficulty level of the game. *A* = maximum +1 point. *B* = maximum −0.5 points. Baseline scores are located at month 3.

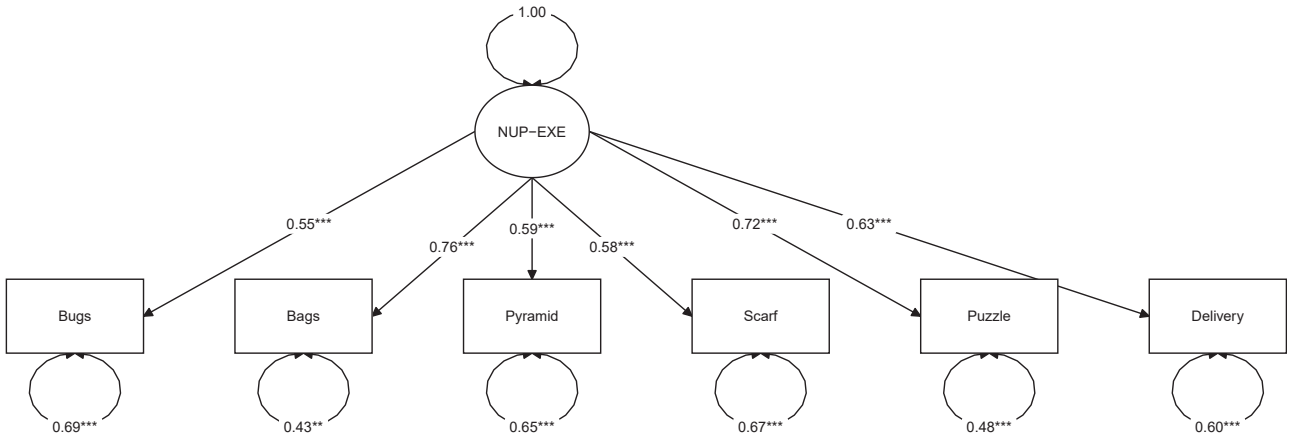


Figure 1. Factorial validity of the measurement model of NUP-Executive functions (NUP-EXE) in the calibration sample (N = 56)
Model fit statistics were $\chi^2(9) = 10$, $p = 0.383$; CFI = 0.99; RMSEA = 0.035; and SRMR = 0.059. Values represent standardized estimates. All values are statistically significant (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

symptoms within the last 5 years, (iii) concern about symptoms, (iv) perception of lower performance compared to same age group, and/or (v) confirmation of symptoms by an informant.

Adherence to NUP

The mean adherence to the cognitive training intervention was 73.8% (95% confidence interval [CI] 71.5, 76.0). Moreover, almost all participants (94.6%) completed at least half of the training. The first month of intervention presented the lowest adherence (44.2%, 95% CI 35.7, 52.7), though in the second month it increased to 73.2% (95% CI 66.1, 80.3) and then reached the maximum of 82.0% (95% CI 75.7, 88.3) in month 5, which means that, on average, participants performed 10 of the 12 sessions that were scheduled (Table S2). Finally, the adherence to NUP remained between 74.0 and 82.0% until month 11, though it slightly decreased to 68.0% in month 12 (Figure S1).

Participants' feedback

Most participants (N = 48, 85.7%) completed an online survey about the acceptability of the cognitive training intervention (Table S3). Almost all reported they had enough resources or knowledge for using NUP, and over 80% stated that the digital platform was easy to use and had learned easily how to use it. However, for 39.6% of participants, game instructions were not clear or understandable enough. In addition, more than 70% reported their motivation to use NUP remained high during the study and that the intervention had an adequate periodicity and duration.

Factorial validity and reliability of NUP

After exploring the univariate pattern of correlations among NUP games of executive functions in the calibration and validation samples, the game 'Stop the ball' was excluded as it did not correlate with the remaining scores. Then, the latent structure of executive functions derived from NUP scores (NUP-EXE) was examined and confirmed by confirmatory factor analysis (CFA). Accordingly, the game 'Déjà vu' was removed as it was not significantly associated with NUP-EXE. Therefore, NUP-EXE was composed by the scores derived from the following six games: 'Sorting bugs', 'Balance the bags', 'Home delivery', 'Card pyramid', 'Knitting a scarf', and 'Jigsaw puzzle'.

In the calibration sample, the hypothesized model provided good fit to the data ($\chi^2(9) = 10$, $p = 0.383$; comparative fit index [CFI] = 0.99; robust mean square error of approximation [RMSEA] = 0.035; and standardized root-mean-square residual [SRMR] = 0.059), and the factor loadings of each game were significant and above 0.5 (Figure 1). The average variance extracted (AVE) was moderate (AVE = 0.41). Construct reliability and replicability were high when measured with the *H* index ($H = 0.82$) and the Cronbach's α ($\alpha = 0.81$, 95% CI 0.68, 0.87) (Table S4).

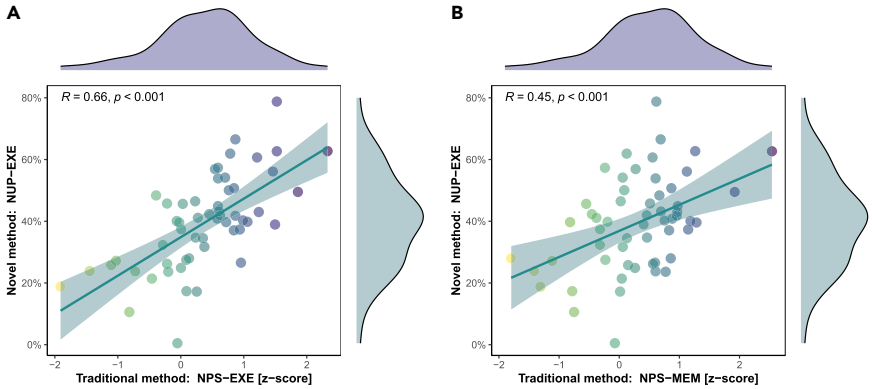


Figure 2. Pearson's correlation between NUP-EXE scores and (A) a composite score from traditional tests of executive functions (NPS-EXE) or (B) a composite score from traditional tests of memory (NPS-MEM) at 6 months (*N=56*)

Each panel includes the correlation in the center and the distribution of each variable in the top and left sides. NPS-EXE refers to a composite score created by averaging the standardized scores from the Five Digits Test (FDT) flexibility score, the Stroop Word-Color score, the Digit Span backwards score, the WAIS-Digit Symbol substitution test direct score, and the Visual Puzzle test direct score. NPS-MEM refers to a composite score created by averaging the standardized scores from the Free and Cued Selective Reminding Test (FCSRT) immediate free recall and delayed free recall and the WMS Logical Memory immediate recall and recognition. The gradient of colors from yellow to purple is based on the gradient of scores (from lower to higher) in traditional neuropsychological tests.

Subsequently, the factorial validity and reliability of NUP-EXE were tested and confirmed in the validation samples that included 6-month and 12-month data (Figure S2). Moreover, reliability values slightly improved after 6 months (AVE = 0.58; H = 0.90; and α = 0.88, 95% CI 0.82, 0.93) and after 12 months (AVE = 0.50; H = 0.86; and α = 0.86, 95% CI 0.77, 0.90).

Longitudinal measurement invariance

The proposed NUP-EXE model was configural and metric invariant, demonstrating the same factor structure and equality of the factor loading matrix in the calibration and validation samples (Table S5). Individual trajectories of NUP-EXE over 10 measurement occasions (month 3 to month 12) are represented in Figure S3.

Convergent and discriminant validity

First, the factorial validity, reliability, and measurement invariance of the traditional composites of executive functions (NPS-EXE) and memory (NPS-MEM) were tested. Using baseline data, the correlated 2-factor model of memory and executive functions provided good fit to the data ($\chi^2(24) = 29$, p = 0.210; CFI = 0.965; RMSEA = 0.063; SRMR = 0.078) (Figure S4). NPS-MEM and NPS-EXE were moderately correlated (β_{STD} = 0.66, p < 0.001). Reliability values were H = 0.80 for NPS-EXE and H = 0.77 for NPS-MEM. The AVE was 0.38 for NPS-EXE and 0.43 for NPS-MEM and was equivalent to the squared multiple correlation coefficient (SMC = 0.43), showing that each factor explained the same variance in its respective indicators than with the other factor of the model. Moreover, configural and metric invariance models for NPS-EXE and NPS-MEM over three measurement occasions (baseline, 6 months, and 12 months) provided reasonable fit to the data (Table S6).

Then, convergent and discriminant validity of NUP-EXE was examined by analyzing the univariate pattern of correlations with similar measures (e.g., NPS-EXE) and less-related measures (e.g., NPS-MEM) (Figure 2). As shown in Table 2, the correlation between NUP-EXE and NPS-EXE was stronger than that between NUP-EXE and NPS-MEM, particularly after 6 months (r = 0.66 vs. r = 0.45) and 12 months (r = 0.57 vs. r = 0.37). NUP-EXE correlated strongly with measures of global cognition, including the Alzheimer Disease Cooperative Study Preclinical Alzheimer Cognitive Composite (ADCS-PACC) (r = 0.65), the ADCS-PACC-plus-exe (r = 0.63), and the MoCA (r = 0.47), and moderately with the mini-mental state examination (MMSE) (r = 0.42). Finally, NUP-EXE negatively correlated with the age of participants (r = −0.55), and weak correlations were also observed between NUP-EXE and World Health Organization quality of life (WHOQOL) measures, including overall health (r = 0.37), overall quality of life (r = 0.34), and the physical domain (r = 0.34) (Table S7).

Table 2. Cross-sectional correlations (Pearson r) of NUP-EXE score with traditional neuropsychological measures at baseline and after 6 and 12 months

Traditional measures		Novel measure of executive functions: NUP-EXE		
		Baseline	6 months	12 months
Domain	Score	*r* (P value)	*r* (P value)	*r* (P value)
Executive functions	Executive functions composite (NPS-EXE)[a]	**0.40** (0.002)	**0.66** (<0.001)	**0.57** (<0.001)
	FDT Flexibility	0.10 (0.467)	**0.40** (0.002)	**0.41** (0.002)
	Stroop WC	0.31 (0.019)	**0.34** (0.010)	**0.37** (0.005)
	Digit span backwards	**0.34** (0.010)	**0.46** (<0.001)	**0.35** (0.008)
	Digit symbol	**0.38** (0.004)	**0.51** (<0.001)	**0.53** (<0.001)
	Visual puzzle	0.30 (0.026)	**0.59** (<0.001)	**0.44** (<0.001)
	FDT Inhibition	0.10 (0.478)	**0.52** (<0.001)	**0.35** (0.009)
	Stroop Interference	0.24 (0.071)	0.05 (0.715)	0.03 (0.822)
Memory	Memory composite (NPS-MEM)[b]	**0.44** (<0.001)	**0.45** (<0.001)	**0.37** (0.005)
	FCSRT IFR	**0.34** (0.009)	0.40 (0.002)	0.15 (0.256)
	FCSRT DFR	**0.42** (0.001)	0.27 (0.042)	0.22 (0.104)
	FCSRT Total Recall	**0.39** (0.003)	0.17 (0.220)	−0.05 (0.723)
	LM IR	0.28 (0.036)	**0.47** (<0.001)	**0.42** (0.001)
	LM DR	0.18 (0.194)	**0.46** (<0.001)	**0.34** (0.010)
	LM recognition	0.27 (0.045)	0.20 (0.130)	**0.35** (0.008)
Global cognition	ADCS-PACC[c]	**0.51** (<0.001)	**0.65** (<0.001)	**0.46** (<0.001)
	ADCS-PACC-plus-exe[d]	**0.47** (<0.001)	**0.63** (<0.001)	**0.46** (<0.001)
	MMSE total score	0.29 (0.033)	**0.42** (0.001)	0.18 (0.194)
	MMSE orientation	−0.04 (0.797)	0.19 (0.171)	0.07 (0.600)
	MMSE attention	0.30 (0.023)	**0.34** (0.010)	0.15 (0.284)
	MMSE memory	0.06 (0.662)	0.31 (0.019)	−0.03 (0.808)
	MMSE language	0.24 (0.079)	0.17 (0.21)	0.21 (0.117)
	MoCA total score	0.27 (0.042)	**0.47** (<0.001)	0.14 (0.298)
	MoCA visuospatial/executive	0.23 (0.087)	**0.40** (0.002)	0.32 (0.018)
	MoCA delayed memory	0.23 (0.085)	0.3 (0.022)	0.07 (0.633)
	MoCA attention	0.32 (0.015)	0.25 (0.065)	0.02 (0.875)
	MoCA language	0.09 (0.510)	0.12 (0.368)	0.07 (0.603)
Other cognitive measures	Boston Naming Test	**0.33** (0.013)	0.30 (0.027)	0.10 (0.454)
	Animal fluency test	0.30 (0.026)	**0.36** (0.007)	0.18 (0.189)

Bold values denote statistical significance at $p < 0.010$.

[a]The executive functions composite from the traditional tests refers to a composite score created by averaging the standardized scores from the Five Digits Test (FDT) flexibility score, the Stroop Word-Color score, the Digit Span backwards score, the WAIS-Digit Symbol substitution test direct score, and the Visual Puzzle test direct score.

[b]The memory composite from the traditional tests refers to a composite score created by averaging the standardized scores from the Free and Cued Selective Reminding Test (FCSRT) immediate free recall and delayed free recall and the WMS Logical Memory immediate recall and recognition.

[c]The ADCS-PACC was composed by the FCSRT total immediate recall, the WMS logical memory total delayed recall, WAIS-Digit Symbol Substitution Test direct score, and the MMSE total score.

[d]The ADCS-PACC-plus-exe score added the Stroop Interference and the FDT flexibility score to the original ADCS-PACC composite. MoCA = Montreal Cognitive Assessment. MMSE = Mini-Mental State Examination.

Sensitivity to change of NUP

NUP-EXE was sensitive to change across intervals of two months, presenting moderate-to-large effect size differences from baseline to 12 months (Table S8). However, change in NUP-EXE after 6 or 12 months

Table 3. Correlation (Pearson r) between change in NUP-EXE after 6 and 12 months, baseline sociodemographic characteristics, and the respective change in quality of life and functionality measures

Change in traditional measures		Change in NUP-EXE	
		6-month change	12-month change
Domain	Score	*r* (P value)	*r* (P value)
Baseline sociodemographic characteristics	Age (years)	**−0.39** (0.003)	−0.18 (0.178)
	Cognitive Reserve score	0.14 (0.460)	0.19 (0.380)
	Education (years)	0.11 (0.435)	0.12 (0.364)
Change in quality of life	WHOQOL Environmental	**0.29** (0.029)	0.01 (0.922)
	WHOQOL Physical	0.20 (0.137)	0.02 (0.906)
	WHOQOL Psychological	**0.34** (0.010)	0.19 (0.184)
	WHOQOL Social	0.23 (0.090)	0.07 (0.610)
	WHOQOL Overall quality of life	0.12 (0.398)	0.13 (0.345)
	WHOQOL Overall health	0.20 (0.194)	0.07 (0.653)
Change in functionality	ABAS-II General Adaptive Composite	**0.32** (0.036)	0.09 (0.652)
	ABAS-II Conceptual Index	0.06 (0.709)	−0.19 (0.323)
	ABAS-II Communication	−0.04 (0.819)	−0.20 (0.281)
	ABAS-II Functional academics	0.07 (0.639)	−0.15 (0.423)
	ABAS-II Self-direction	0.26 (0.093)	0.15 (0.423)
	ABAS-II Social Index	**0.30** (0.047)	**0.40** (0.029)
	ABAS-II Leisure	0.22 (0.161)	0.28 (0.129)
	ABAS-II Social	0.27 (0.081)	0.18 (0.346)
	ABAS-II Practical Index	**0.35** (0.021)	0.19 (0.312)
	ABAS-II Community use	0.16 (0.299)	−0.07 (0.708)
	ABAS-II Home Living	0.12 (0.441)	0.24 (0.193)
	ABAS-II Health and safety	**0.40** (0.007)	0.11 (0.573)
	ABAS-II Health care	0.26 (0.092)	0.08 (0.684)

WHOQOL = World Health Organization Quality of Life brief generic questionnaire. ABAS = Adaptive Behavior Assessment System. Bold values denote statistical significance at $p < 0.05$.

did not significantly correlate with the respective change in most traditional neuropsychological measures, except for the Visual Puzzle Test ($r = 0.28$) and the Logical Memory (LM) immediate recall (IR), delayed recall (DR), and recognition scores ($r = 0.30$) after 6 months (Table S9). On the other hand, as shown in Table 3, changes from baseline to 6 months in NUP-EXE negatively correlated with the age of participants ($r = -0.39$) and positively correlated with 6-month changes in functionality and quality of life, including the Adaptive Behavior Assessment System (ABAS) general adaptive composite ($r = 0.32$), the ABAS social index ($r = 0.30$), the ABAS practical index ($r = 0.35$), the ABAS health and safety domain ($r = 0.40$), the WHOQOL psychological domain ($r = 0.34$), and the WHOQOL environmental domain ($r = 0.29$).

Comparison of the sensitivity to change between NUP and traditional neuropsychological tests

NUP-EXE increased linearly from month 3 to month 11, from 30.7% to 50.0%, and then remained at 48.8% points after 12 months (Table S10). The effect size of NUP-EXE changes (Cohen's d = 0.65 after 6 months and d = 1.22 after 12 months) was larger than the effect size of cognitive changes measured with traditional neuropsychological tests (Figure 3). Accordingly, the NPS-EXE increased after 6 months (+0.37 *Z* score units, d = 0.49) and then slightly decreased after 12 months (+0.27 *Z* score units compared to baseline, d = 0.36). On the other hand, the ADCS-PACC only increased after 12 months (+0.34 *Z* score units, d = 0.60), whereas the ADCS-PACC-plus-exe showed improvements after 6 months (+0.17 *Z* score units, d = 0.30) and after 12 months (+0.30 *Z* score units, d = 0.56). Finally, the MMSE remained stable at 29 points

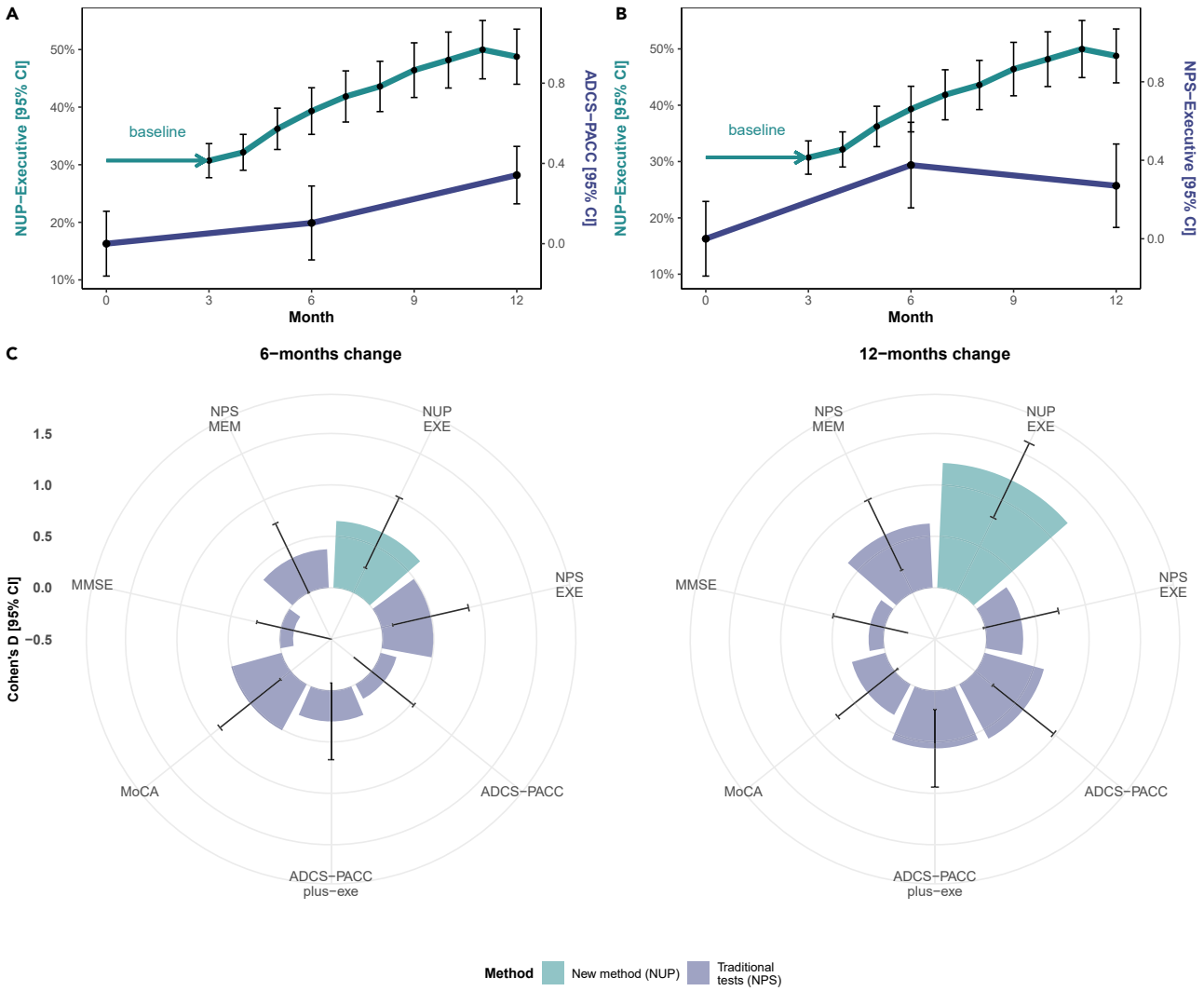


Figure 3. Comparison between longitudinal mean trajectories (panels A and B) and change sensitivity (panel C) between NUP-EXE score and 'traditional' neuropsychological tests (N = 56)

In panel C, bars represent Cohen's d effect size estimates. The ADCS-PACC included the FCSRT total immediate recall, the WMS Logical Memory total delayed recall, WAIS-Digit Symbol Substitution Test direct score, and the MMSE total score. The ADCS-PACC-plus-exe score added the Stroop Interference and the FDT flexibility score to the original ADCS-PACC composite. NPS-MEM refers to a composite score created by averaging the standardized scores from the FCSRT immediate free recall and delayed free recall and the WMS Logical Memory immediate recall and recognition. NPS-EXE refers to a composite score created by averaging the standardized scores from the FDT flexibility score, the Stroop Word-Color score, the Digit Span backwards score, and the WAIS-Digit Symbol Substitution Test direct score. MoCA = Montreal Cognitive Assessment. MMSE = Mini-Mental State Examination. Error bars represent 95% confidence intervals (95% CI).

during all the intervention, and the MoCA increased 1 point after 6 months (d = 0.51) and 0.79 points after 12 months (d = 0.34).

Latent class trajectories of NUP-EXE

We classified participants into three different subgroups based on NUP-EXE trajectory (Figure 4A). A linear random effects model with a quadratic term and 3 latent classes was selected because it presented the best information criterion indices (Table S11). The discrimination index of the selected model was high (entropy of 0.87), and the mean likelihood of class membership was 96% for class 1, 94% for class 2, and 97% for class 3. Most participants belonged to class 2 (64.3%) and presented a mean increase in NUP-EXE of 21%, increasing linearly from baseline (27.3%; 95% CI 24.7, 30.0%) to 12 months (48.6%; 95% CI 43.9, 53.4%). Class 3 subjects represented the 28.6% of participants and

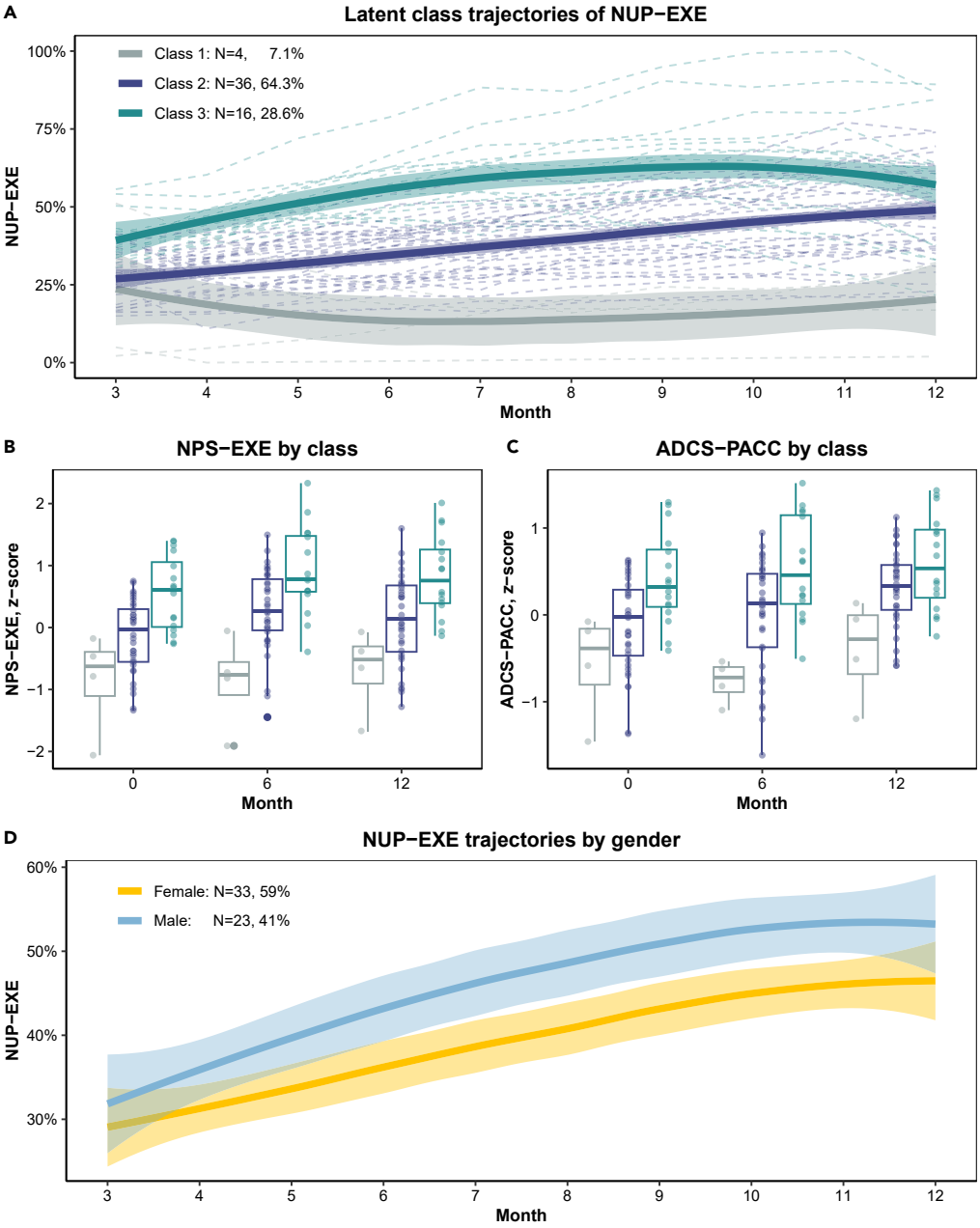


Figure 4. Representation of latent class trajectories derived from NUP-EXE (A), the class-specific scores in NPS-EXE (B) and ADCS-PACC (C), and the overall mean trajectory of NUP-EXE by gender (D)

NPS-EXE refers to a composite score created by averaging the standardized scores from the FDT flexibility score, the Stroop Word-Color score, the Digit Span backwards score, and the WAIS-Digit Symbol Substitution Test direct score. The ADCS-PACC included the FCSRT total immediate recall, the WMS Logical Memory total delayed recall, WAIS-Digit Symbol Substitution Test direct score, and the MMSE total score. Trajectories in pannels (A) and (D) represent the mean value at each time point and its 95% confidence interval (95% CI).

were characterized by higher NUP-EXE scores during all the follow-ups and a quadratic trajectory. Specifically, their baseline NUP-EXE score was 40.0% (95% CI 35.9, 44.2%), which increased to 60% (95% CI 53, 66%) after 7 months and then remained rather stable until the end of follow-up. Finally, a minor group of participants (N = 4, 7%) belonged to class 1 and presented a stable and lower NUP-EXE trajectory during the 12 months. The three latent subgroups also differed in the NPS-EXE and ADCS-PACC scores (Figures 4B and 4C).

Moreover, the three subgroups differed in their monthly adherence to the cognitive training intervention ($p < 0.001$), which was 35% in class 1 (meaning ~1 session/week instead of 3 sessions/week as scheduled) and 81.9% and 75.7% in class 2 and 3, respectively (Table S12). Moreover, there was a greater proportion of males in class 3 (62.5%) than in class 1 (30.6%) or 2 (50.0%). Class 3 subjects were the youngest ones (mean ± SD age of 64.6 ± 4.1 years), followed by class 2 subjects (mean of 67.4 ± 4.6) and class 1 subjects (mean of 72.5 ± 3.11 years). A gradient from class 3 to 1 was also observed in global cognition evaluated with the MMSE (from 29.8 ± 0.5 points to 28.8 ± 1.4 and. 28.5 ± 0.6, $p = 0.022$) and the MoCA (from 27.3 ± 2.4 points to 26.8 ± 2.5 and 23.8 ± 1 points, $p = 0.034$). Class 1 subjects, compared to class 2 and 3 subjects, also presented lower functionality evaluated with the ABAS general adaptive composite score (84.4 vs. 104.0 vs. 95.9 points, respectively; $p = 0.027$) and the ABAS conceptual and social indexes ($p = 0.029$), as well as in the specific domains of communication, self-direction, and social life.

Gender-specific trajectories of NUP-EXE

Although no gender differences in NUP-EXE scores were detected at baseline, males presented greater improvements in executive functions than females (Figure 4D). Accordingly, after 7 months, the average NUP-EXE improvement in males was 14.8% whereas in females it was 8.6% (Cohen's $d = -0.52$, $p = 0.062$). The effect size of gender differences at each time point was moderate during all the follow-up (Cohen's d between −0.40 and −0.50) (Table S13).

DISCUSSION

We provide evidence of acceptability, reliability, validity, and sensitivity to change of the executive functions measure NUP-EXE extracted from the performance in six cognitive training games delivered via the digital platform NeuronUP©. Our findings support the feasibility of remote and unsupervised monitoring of executive functions on a monthly basis, enabling digital assessments of cognitive change and contributing to parsing heterogeneity in treatment effectiveness.[24,37,38] The study population consisted of older adults at risk of AD (*APOE*-ε4 carriers meeting SCD-plus criteria) who were participating in a lifestyle-based multimodal intervention for preventing cognitive decline (PENSA Study, Prevention of cognitive decline in subjective cognitive decline APOE-ε4 carriers after epigallocatechin gallate and a multimodal intervention).[39] Given the similar prodromal cognitive stages that characterize other neurodegenerative disorders and the burgeoning interest on the efficacy of lifestyle-based prevention interventions to counter cognitive decline, our findings have potential to generalize to other disorders and inform development of outcome measures for a new generation of intervention trials. Results show that NUP-EXE is a reliable and valid measure of executive functions and it is adequate to intensively assessing changes in this cognitive domain since learning effects were minimized by automatically adjusting difficulty to capacity and by obtaining reliable measures of the optimal performance of each subject after repeated testing. We also show that NUP-EXE was able to capture cognitive changes with greater sensitivity than 'traditional' tests. Moreover, improvements in NUP-EXE correlated with improvements in functionality and quality of life and were influenced by the age and gender of participants. In addition, the adherence to the cognitive training intervention was high, and participants positively valued the utility of the digital platform. Finally, we demonstrate the potential of this novel approach to enable previously elusive characterization of cognitive trajectories over time.

One critical drawback of traditional cognitive assessments is learning effects after repeated administrations.[6,7] In the present study, the difficulty of the cognitive training task increased, decreased, or remained stable according to the performance of each participant. Moreover, games were conceived using principles from the cognitive science literature that may reduce practice and ceiling effects, such as reduction of memorization of responses, alternative forms, or distractors. All these technical characteristics of the NUP platform were leveraged for addressing practice effects. Accordingly, a massed practice strategy was applied[7] so that the initial months of games' exposure were used to personalize baseline performance and calibrate difficulty. Moreover, the performance in NUP was evaluated using a conservative scoring system as monthly scores were mainly based on the most repeated difficulty level achieved for each individual in each game. Therefore, a significant improvement can only be detected when the individual spends most of the session at a higher difficulty level than in the previous month.

NUP-EXE presented good psychometric properties evaluated with CFA. The one-domain structure showed good fit with the data and high factor loadings in the calibration and validation samples. Moreover, NUP-EXE presented measurement invariance over time, suggesting psychometric stability of executive

functions, as well as replicability of findings. Although games were very heterogeneous in the form of presentation and most combined the simultaneous training of several specific domains, internal consistency was remarkable. These results were comparable to the reliability coefficients derived from traditional neuropsychological tests in our sample and are similar to the ones observed with classic neuropsychological batteries.[40–43] Given that the standardized factor loadings were very close for the NUP-EXE factor, for the sake of simplicity the NUP-EXE composite score was created by summing scores instead of computing factor scores.[44]

Convergent and discriminant validity was demonstrated by showing that NUP-EXE had higher correlations with the traditional executive functions composite (NPS-EXE) than with the traditional memory composite (NPS-MEM). NUP-EXE also exhibited moderately high associations with the gold-standard cognitive measure for preclinical AD, the ADCS-PACC.[1] These results are comparable or even superior to the observed in other studies using digital assessments.[25,45,46] Moreover, as hypothesized, there was a gradient in NUP-EXE scores according to the age of participants. However, correlations between NUP-EXE and the cognitive reserve of participants were weak.

NUP-EXE presented noteworthy internal change sensitivity as it was able to detect the rising trajectory of cognitive changes until the ninth measurement occasion at 11 months. The magnitude of changes detected with NUP was significantly larger than that detected via traditional tests. This improved sensitivity of NUP-EXE is important given the consistently reported low sensitivity to change of cognitive measures in preclinical stages of AD, particularly when administered at short time intervals.[47,48] However, NUP presented poor external change sensitivity as change after 6 or 12 months did not correlate with the respective change in most of the traditional measures of executive functions. Cognitive training games are usually 'task specific' so that the training of a specific domain may improve the performance in that domain, but the transfer capacity to other domains may be limited.[49,50] In this study, the NPS-EXE composite included measures of inhibition, resistance to interference and processing speed that were not specifically trained by any NUP game. By contrast, 6-month change in NUP-EXE correlated with change in the visual puzzle test, which was trained with the NUP 'Jigsaw puzzle' game. On the other hand, change in NUP-EXE positively correlated with change in functionality and quality of life and was influenced by the age and gender of participants, which is consistent with results from other studies[32,35,36] and may support the potential of NUP to evaluate the clinical relevance of preventive interventions for cognitive decline.[51]

Finally, thanks to the resolution of longitudinal data of cognitive performance with NUP, we could examine inter-individual differences in cognitive change over time. Three different trajectories defining linear, quadratic, and stagnating cognitive changes were described by analyzing NUP-EXE, emphasizing the heterogeneity in response to the intervention. Most participants followed the class 2 linear trajectory or the class 3 quadratic trajectory, both characterized by an improvement in executive functions at 12 months but worse and better performance levels over time, respectively. However, a minority of participants followed the class 1 trajectory that was characterized by remaining stable over time. Class 1 participants presented substantially lower adherence to the intervention and were on average older than the participants of other two classes, which could justify the lower cognitive scores in NUP-EXE and in the traditional tests. Conversely, class 2 and 3 had both higher basal cognitive performance and adherence to the cognitive intervention but differed in the male gender distribution, which was higher in class 3. Although more research is needed to consolidate these findings, they already illustrate the potential of this approach for identifying people who have less probability of responding to a preventive intervention for cognitive decline and offer to them a more personalized and intensive intervention. Moreover, the integration of NUP data with ecological momentary assessments (e.g., diet compliance, mental health)[39,52] and activity tracker data may help advance understanding of the dynamics among cognition, lifestyle, and mental health over time, as well as assess the impact of the COVID-19 pandemic on study outcomes.[53]

Limitations of the study

This study must be interpreted in light of some limitations. First, it is a secondary analysis of a clinical trial that was not specifically designed to evaluate NUP. An optimum design should have included a well-matched control group performing cognitive training only and an active control group including healthy older adults without SCD or individuals with mild cognitive impairment. This would have allowed the comparison of NUP outcomes between a multimodal intervention that would theoretically produce greater cognitive improvement and a single-domain intervention that would produce lower cognitive

improvements.[54] Moreover, data from a separate study population could have been useful for gaining more insights about the specificity of findings. Actually, the personalized design of the PENSA Study and the intensity of the follow-up of participants, as well as the inclusion of well-educated individuals that were aware of their increased risk of AD, may partly explain the high adherence to the cognitive training intervention compared to other studies, though can affect the generalizability of findings. Second, the observed correlations between NUP-EXE and traditional measures were larger in the validation samples than in the calibration sample. This may be probably due to the most contiguous assessments in the validation samples (measures obtained at the same time) compared with the calibration sample. Specifically, in the calibration sample traditional tests were obtained about 1–3 months prior to starting the intervention, whereas baseline NUP scores were obtained in the third month of intervention, so there was a difference of at least 4 months between measurement occasions. Third, the lack of consistent external change sensitivity of NUP-EXE represents a limitation of this measure and could be due to the lack of generalizability of the acquired skills across other cognitive domains, differences in test-retest reliability between measures, the small range for improvement in traditional tests by individuals with cognitive performance within the normal range, and the higher number of cognitive assessments by NUP that lead to different smooth curves of change. Fourth, the scoring system to evaluate participants' performance in NUP games assumed that the difficulty of each activity increased linearly with increasing phase number, but a non-linear gradient in difficulty could exist. Finally, the sample size was limited to N = 56 individuals, but they were completely followed during 12 months of multimodal intervention and were extensively characterized in terms of cognition and functionality.

Conclusions

In summary, findings from this study suggest that NUP is, in principle, well positioned to address some of the recently highlighted limitations of traditional neuropsychological assessment methods in SCD subjects,[1] allowing a more frequent sampling of the cognitive performance on a monthly basis with high sensitivity to detect subtle cognitive changes. The combination of this novel data collection methodology based on the performance in cognitive training games, together with gold-standard lab-based neuropsychological assessments, is expected to allow a more accurate characterization of an individual's response to a cognitive decline preventive intervention.

CONSORTIA

PENSA Study Group: From IMIM: Rafael de la Torre, Neus Pizarro, Laura Forcano, Albert Puig-Pijoan, Natalia Soldevila-Domenech, Anna Boronat, Thais Lorenzo, Aida Cuenca-Royo, Iris Piera, Ana Aldea, Patricia Diaz-Pellicer, Ilario De Toma, Mara Dierssen, Maria Gomis-González, Esther Mur Gimeno, Sergi Martínez, Joana Crivillé, Julian Mateus. From BBRC: José Luis Molinuevo, Gonzalo Sánchez-Benavides, Juan Domingo Gispert, Karine Fauria, Carolina Minguillón, Oriol Grau-Rivera, Iva Knezevic, Sofia Menezes-Cabral, Anna Soteras, José Maria González-de-Echávarri.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Study design and population
  - Ethics statement
  - PENSA study multimodal intervention
  - Recruitment strategy
  - Inclusion and exclusion criteria
  - Timing
  - Data exclusions
  - Non-participation
- METHOD DETAILS
  - Cognitive training

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106886.

ACKNOWLEDGMENTS

The authors thank the participants for their enthusiastic collaboration and the PENSA Study Group personnel for their invaluable support. The PENSA Study is primarily supported by the Alzheimer Association (18PTC-R-592192 The PART THE CLOUD to RESCUE Brain Cell Degeneration in Alzheimer's disease Program) and secondarily by Instituto de Salud Carlos III (ISCIII PI17/00223). This work is supported by the Spanish Society of Epidemiology ('Miguel Carrasco Awards') and the Departament d'Economia i Coneixement from the Generalitat de Catalunya (2017 SGR 138). Natalia Soldevila-Domenech is supported by a predoctoral grant (FI_B2021/00104) from the Agency for Management of University and Research Grants (AGAUR) of the Generalitat de Catalunya. Thais Lorenzo is supported by a predoctoral fellowship PFIS-ISCIII (FI18/00041). CIBER de Fragilidad y Envejecimiento Saludable (CIBERFES) and Fisiopatología de la Obesidad y Nutrición (CIBEROBN) are initiatives of the Instituto de Salud Carlos III, Madrid, Spain, and funded by the European Regional Development Fund.

AUTHOR CONTRIBUTIONS

R.d.l.T., N.S.-D., and I.T. conceptualized the study design. L.F., A.C.-R., T.L., P.D.-P. M.G-G., C.S., and Í.P. collected the data. N.S.-D. and I.T. analyzed the data. N.S-D. prepared the figures and tables. R.d.l.T., N.S.-D., I.T., A.C.-R., and L.F. wrote the initial manuscript draft. A.V.-G., G.S.-B., J.L.M., and K.F. contributed to the critical interpretation of data. All authors revised the manuscript and approved the submitted version.

DECLARATION OF INTERESTS

The authors declare the following competing interests: Carolina Sastre and Íñigo Fernandez De Piérola are employees of the company NeuronUP©. This company did not fund the study, had no role in its design, and did not participate during its execution or in the decision to submit results. Dr. José Luis Molinuevo is currently a full-time employee of H. Lundbeck A/S and previously has served as a consultant or at advisory boards for the following for-profit companies or has given lectures in symposia sponsored by the following for-profit companies: Roche Diagnostics, Genentech, Novartis, Lundbeck, Oryzon, Biogen, Lilly, Janssen, Green Valley, MSD, Eisai, Alector, BioCross, GE Healthcare, and ProMIS Neurosciences.

Received: October 24, 2022
Revised: February 26, 2023
Accepted: May 11, 2023
Published: May 16, 2023

REFERENCES

1. Donohue, M.C., Sperling, R.A., Salmon, D.P., Rentz, D.M., Raman, R., Thomas, R.G., Weiner, M., and Aisen, P.S.; Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Cooperative Study (2014). The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol. *71*, 961–970. https://doi.org/10.1001/jamaneurol.2014.803.

2. Folstein, M.F., Folstein, S.E., and McHugh, P.R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. *12*, 189–198. https://doi.org/10.1016/0022-3956(75)90026-6.

3. Ojeda, N., Del Pino, R., Ibarretxe-Bilbao, N., Schretlen, D.J., and Pena, J. (2016). Montreal cognitive assessment test: normalization and standardization for Spanish population. Rev. Neurol. *63*, 488–496.

4. Miller, J.B., and Barr, W.B. (2017). The technology crisis in neuropsychology. Arch. Clin. Neuropsychol. *32*, 541–554. https://doi.org/10.1093/arclin/acx050.

5. Chaytor, N., and Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. Neuropsychol. Rev. *13*, 181–197. https://doi.org/10.1023/B:NERV.0000009483.91468.fb.

6. Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., and Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. BMC Neurosci. *11*, 118. https://doi.org/10.1186/1471-2202-11-118/FIGURES/4.

7. Goldberg, T.E., Harvey, P.D., Wesnes, K.A., Snyder, P.J., and Schneider, L.S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimers Dement. *1*, 103–111. https://doi.org/10.1016/j.dadm.2014.11.003.

8. Pavel, M., Jimison, H., Hayes, T., Kaye, J., Dishman, E., Wild, K., and Williams, D. (2008). Continuous, unobtrusive monitoring for the assessment of cognitive function. Handb. Cogn. Aging Interdiscip. Perspect. 524–541.

9. GBD 2019 Collaborators (2021). Global mortality from dementia: application of a new method and results from the global burden of disease study 2019. Alzheimers Dement. *7*, e12200. https://doi.org/10.1002/trc2.12200.

10. World Health Organization. Dementia Key Facts. https://www.who.int/news-room/fact-sheets/detail/dementia.

11. Molinuevo, J.L., Minguillon, C., Rami, L., and Gispert, J.D. (2018). The rationale behind the new Alzheimer's disease conceptualization: lessons learned during the last decades. J. Alzheimer's Dis. *62*, 1067–1077. https://doi.org/10.3233/JAD-170698.

12. Hickman, R.A., O'Shea, S.A., Mehler, M.F., and Chung, W.K. (2022). Neurogenetic disorders across the lifespan: from aberrant development to degeneration. Nat. Rev. Neurol. *18*, 117–124. https://doi.org/10.1038/s41582-021-00595-5.

13. Aisen, P.S., Cummings, J., Jack, C.R., Morris, J.C., Sperling, R., Frölich, L., Jones, R.W., Dowsett, S.A., Matthews, B.R., Raskin, J., et al. (2017). On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimer's Res. Ther. *9*, 60. https://doi.org/10.1186/s13195-017-0283-5.

14. Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet *396*, 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6.

15. Mitchell, A.J., Beaumont, H., Ferguson, D., Yadegarfar, M., and Stubbs, B. (2014). Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. Acta Psychiatr. Scand. *130*, 439–451. https://doi.org/10.1111/acps.12336.

16. Jessen, F., Amariglio, R.E., Buckley, R.F., van der Flier, W.M., Han, Y., Molinuevo, J.L., Rabin, L., Rentz, D.M., Rodriguez-Gomez, O., Saykin, A.J., et al. (2020). The characterisation of subjective cognitive decline. Lancet Neurol. *19*, 271–278. https://doi.org/10.1016/S1474-4422(19)30368-0.

17. Molinuevo, J.L., Rabin, L.A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K.A., Ewers, M., Hampel, H., Klöppel, S., Rami, L., et al. (2017). Implementation of subjective cognitive decline criteria in research studies. Alzheimers Dement. *13*, 296–311. https://doi.org/10.1016/j.jalz.2016.09.012.

18. Elkana, O., Eisikovits, O.R., Oren, N., Betzale, V., Giladi, N., and Ash, E.L. (2016). Sensitivity of neuropsychological tests to identify cognitive decline in highly educated elderly individuals: 12 months follow up. J. Alzheimer's Dis. *49*, 607–616. https://doi.org/10.3233/JAD-150562.

19. Snyder, P.J., Kahle-Wrobleski, K., Brannan, S., Miller, D.S., Schindler, R.J., Desanti, S., Ryan, J.M., Morrison, G., Grundman, M., Chandler, J., et al. (2014). Assessing cognition and function in Alzheimer's disease clinical trials: do we have the right tools? Alzheimers Dement. *10*, 853–860. https://doi.org/10.1016/j.jalz.2014.07.158.

20. Jutten, R.J., Sikkes, S.A.M., Van der Flier, W.M., Scheltens, P., Visser, P.J., and Tijms, B.M.; Alzheimer's Disease Neuroimaging Initiative (2021). Finding treatment effects in Alzheimer trials in the face of disease progression heterogeneity. Neurology *96*, e2673–e2684. https://doi.org/10.1212/WNL.0000000000012022.

21. Dujardin, S., Commins, C., Lathuiliere, A., Beerepoot, P., Fernandes, A.R., Kamath, T.V., De Los Santos, M.B., Klickstein, N., Corjuc, D.L., Corjuc, B.T., et al. (2020). Tau molecular diversity contributes to clinical heterogeneity in Alzheimer's disease. Nat. Med. *26*, 1256–1263. https://doi.org/10.1038/s41591-020-0938-9.Tau.

22. Van der Flier, W.M. (2016). Clinical heterogeneity in familial Alzheimer's disease. Lancet Neurol. *15*, 1296–1298. https://doi.org/10.1016/S1474-4422(16)30275-7.

23. Galvin, J.E. (2017). Prevention of Alzheimer's disease: lessons learned and applied. J. Am. Geriatr. Soc. *65*, 2128–2133. https://doi.org/10.1111/jgs.14997.

24. Öhman, F., Hassenstab, J., Berron, D., Schöll, M., and Papp, K.V. (2021). Current advances in digital cognitive assessment for preclinical Alzheimer's disease. Alzheimers Dement. *13*, 1–19. https://doi.org/10.1002/dad2.12217.

25. Papp, K.V., Samaroo, A., Chou, H.C., Buckley, R., Schneider, O.R., Hsieh, S., Soberanes, D., Quiroz, Y., Properzi, M., Schultz, A., et al. (2021). Unsupervised mobile cognitive testing for use in preclinical Alzheimer's disease. Alzheimers Dement. *13*, e12243. https://doi.org/10.1002/dad2.12243.

26. Shah, T.M., Weinborn, M., Verdile, G., Sohrabi, H.R., and Martins, R.N. (2017). Enhancing cognitive functioning in healthly older adults: a systematic review of the clinical significance of commercially available computerized cognitive training in preventing cognitive decline. Neuropsychol. Rev. *27*, 62–80. https://doi.org/10.1007/s11065-016-9338-9.

27. Mendoza Laiz, N., Del Valle Díaz, S., Rioja Collado, N., Gomez-Pilar, J., and Hornero, R. (2018). Potential benefits of a cognitive training program in mild cognitive impairment (MCI). Restor. Neurol. Neurosci. *36*, 207–213. https://doi.org/10.3233/RNN-170754.

28. Hill, N.T.M., Mowszowski, L., Naismith, S.L., Chadwick, V.L., Valenzuela, M., and Lampit, A. (2017). Computerized cognitive training in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis. Am. J. Psychiatry *174*, 329–340. https://doi.org/10.1176/appi.ajp.2016.16030360.

29. Shao, Y.K., Mang, J., Li, P.L., Wang, J., Deng, T., and Xu, Z.X. (2015). Computer-based cognitive programs for improvement of memory, processing speed and executive function during age-related cognitive decline: a meta-analysis. PLoS One *10*, e0130831. https://doi.org/10.1371/journal.pone.0130831.

30. Kueider, A.M., Parisi, J.M., Gross, A.L., and Rebok, G.W. (2012). Computerized cognitive training with older adults: a systematic review. PLoS One *7*, e40588. https://doi.org/10.1371/journal.pone.0040588.

31. Jockwitz, C., Wiersch, L., Stumme, J., and Caspers, S. (2021). Cognitive profiles in older males and females. Sci. Rep. *11*, 6524. https://doi.org/10.1038/s41598-021-84134-8.

32. Soldevila-Domenech, N., Forcano, L., Vintró-Alcaraz, C., Cuenca-Royo, A., Pintó, X., Jiménez-Murcia, S., García-Gavilán, J.F., Nishi, S.K., Babio, N., Gomis-González, M., et al. (2021). Interplay between cognition and weight reduction in individuals following a Mediterranean Diet: three-year follow-up of the PREDIMED-Plus trial. Clin. Nutr. *40*, 5221–5237. https://doi.org/10.1016/j.clnu.2021.07.020.

33. Stern, Y. (2013). Cognitive reserve in ageing. Lancet Neurol. *11*, 1006–1012. https://doi.org/10.1016/S1474-4422(12)70191-6.Cognitive.

34. Rami, L., Valls-Pedret, C., Bartrés-Faz, D., Caprile, C., Solé-Padullés, C., Castellví, M., Olives, J., Bosch, B., and Molinuevo, J.L. (2011). Cognitive reserve questionnaire. Scores obtained in a healthy elderly population and in one with Alzheimer's disease. Rev. Neurol. *52*, 195–201. https://doi.org/10.33588/rn.5204.2010478.

35. Marshall, G.A., Rentz, D.M., Frey, M.T., Locascio, J.J., Johnson, K.A., and Sperling, R.A.; Alzheimer's Disease Neuroimaging Initiative (2011). Executive function and instrumental activities of daily living in mild cognitive impairment and Alzheimer's

disease. Alzheimers Dement. *7*, 300–308. https://doi.org/10.1016/j.jalz.2010.04.005.

36. Tarantino, V., Burgio, F., Toffano, R., Rigon, E., Meneghello, F., Weis, L., and Vallesi, A. (2021). Efficacy of a training on executive functions in potentiating rehabilitation effects in stroke patients. Brain Sci. *11*, 1002. https://doi.org/10.3390/brainsci11081002.

37. Weuve, J., Proust-Lima, C., Power, M.C., Gross, A.L., Hofer, S.M., Thiébaut, R., Chêne, G., Glymour, M.M., and Dufouil, C.; MELODEM Initiative (2015). Guidelines for reporting methodological challenges and evaluating potential bias in dementia research. Alzheimers Dement. *11*, 1098–1109. https://doi.org/10.1016/j.jalz.2015.06.1885.

38. Weintraub, S., Carrillo, M.C., Farias, S.T., Goldberg, T.E., Hendrix, J.A., Jaeger, J., Knopman, D.S., Langbaum, J.B., Park, D.C., Ropacki, M.T., et al. (2018). Measuring cognition and function in the preclinical stage of Alzheimer's disease. Alzheimers Dement. *4*, 64–75. https://doi.org/10.1016/j.trci.2018.01.003.

39. Forcano, L., Fauria, K., Soldevila-Domenech, N., Minguillón, C., Lorenzo, T., Cuenca-Royo, A., Menezes-Cabral, S., Pizarro, N., Boronat, A., Molinuevo, J.L., et al. (2021). Prevention of cognitive decline in Subjective Cognitive Decline APOE-e4 carriers after EGCG and a multimodal intervention (PENSA): study design. Alzheimers Dement. *7*, e12155. https://doi.org/10.1002/trc2.12155.

40. Smith, P.J., Need, A.C., Cirulli, E.T., Chiba-Falek, O., and Attix, D.K. (2013). A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with "traditional" neuropsychological testing instruments. J. Clin. Exp. Neuropsychol. *35*, 319–328. https://doi.org/10.1080/13803395.2013.771618.

41. Heaton, R.K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., Beaumont, J., Casaletto, K.B., Conway, K., Slotkin, J., et al. (2014). Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. J. Int. Neuropsychol. Soc. *20*, 588–598. https://doi.org/10.1017/S1355617714000241.

42. Karin, A., Hannesdottir, K., Jaeger, J., Annas, P., Segerdahl, M., Karlsson, P., Sjögren, N., von Rosen, T., and Miller, F. (2014). Psychometric evaluation of ADAS-Cog and NTB for measuring drug response. Acta Neurol. Scand. *129*, 114–122. https://doi.org/10.1111/ane.12153.

43. Amariglio, R.E., Donohue, M.C., Marshall, G.A., Rentz, D.M., Salmon, D.P., Ferris, S.H., Karantzoulis, S., Aisen, P.S., and Sperling, R.A.; Alzheimer's Disease Cooperative Study (2015). Tracking early decline in cognitive function in older individuals at risk for Alzheimer's disease dementia: the Alzheimer's Disease Cooperative Study Cognitive Function Instrument. JAMA Neurol. *72*, 446–454. https://doi.org/10.1001/jamaneurol.2014.3375.

44. McNeish, D., and Wolf, M.G. (2020). Thinking twice about sum scores. Behav. Res. Methods *52*, 2287–2305. https://doi.org/10.3758/s13428-020-01398-0.

45. Papp, K.V., Rentz, D.M., Maruff, P., Sun, C.K., Raman, R., Donohue, M.C., Schembri, A., Stark, C., Yassa, M.A., Wessels, A.M., et al. (2021). The computerized cognitive composite (C3) in A4, an Alzheimer's disease secondary prevention trial. J. Prev. Alzheimers Dis. *8*, 59–67. https://doi.org/10.14283/jpad.2020.38.

46. Bott, N., Madero, E.N., Glenn, J., Lange, A., Anderson, J., Newton, D., Brennan, A., Buffalo, E.A., Rentz, D., and Zola, S. (2018). Device-embedded cameras for eye tracking–based cognitive assessment: validation with paper-pencil and computerized cognitive composites. J. Med. Internet Res. *20*, e11143. https://doi.org/10.2196/11143.

47. Brewster, P., Rush, J., Ozen, L., Jacobs, D.M., Kaye, J., Nygaard, H.B., Feldman, H.H., and Hofer, S.M. (2020). Intensive measurement of cognition to support early detection of cognitive change in individuals at risk of dementia. Alzheimers Dement. *16*, 1–2. https://doi.org/10.1002/alz.042599.

48. Brewster, P.W.H., Rush, J., Ozen, L., Vendittelli, R., and Hofer, S.M. (2021). Feasibility and psychometric integrity of mobile phone-based intensive measurement of cognition in older adults. Exp. Aging Res. *47*, 303–321. https://doi.org/10.1080/0361073X.2021.1894072.

49. Gobet, F., and Sala, G. (2023). Cognitive training: a field in search of a phenomenon. Perspect. Psychol. Sci. *18*, 125–141. https://doi.org/10.1177/17456916221091830.

50. Sagi, D., and Tanne, D. (1994). Perceptual learning: learning to see. Curr. Opin. Neurobiol. *4*, 195–199. https://doi.org/10.1016/0959-4388(94)90072-8.

51. Borland, E., Edgar, C., Stomrud, E., Cullen, N., Hansson, O., and Palmqvist, S. (2022). Clinically relevant changes for cognitive outcomes in preclinical and prodromal cognitive stages: implications for clinical Alzheimer trials. Neurology *99*, E1142–E1153. https://doi.org/10.1212/WNL.0000000000200817.

52. Boronat, A., Clivillé, J., Soldevila-Domenech, N., Forcano, L., Pizarro, N., Fitó, M., Schröder, H., Fauria, K., and De La Torre, R. (2021). Mobile device-assisted dietary ecological momentary assessments for the evaluation of the adherence to the mediterranean diet in a continuous manner. J. Vis. Exp. https://doi.org/10.3791/62161.

53. Röhr, S., Arai, H., Mangialasche, F., Matsumoto, N., Peltonen, M., Raman, R., Riedel-Heller, S.G., Sakurai, T., Snyder, H.M., Sugimoto, T., et al. (2021). Impact of the COVID-19 pandemic on statistical design and analysis plans for multidomain intervention clinical trials: experience from World-Wide FINGERS. Alzheimer's Dement. Transl. Res. Clin. Interv. *7*, 1–10. https://doi.org/10.1002/trc2.12143.

54. Stephen, R., Barbera, M., Peters, R., Ee, N., Zheng, L., Lehtisalo, J., Kulmala, J., Håkansson, K., Chowdhary, N., Dua, T., et al. (2021). Development of the first WHO guidelines for risk reduction of cognitive decline and dementia: lessons learned and future directions. Front. Neurol. *12*, 763573. https://doi.org/10.3389/fneur.2021.763573.

55. Kivipelto, M., Mangialasche, F., Snyder, H.M., Allegri, R., Andrieu, S., Arai, H., Baker, L., Belleville, S., Brodaty, H., Brucki, S.M., et al. (2020). World-Wide FINGERS Network: a global approach to risk reduction and prevention of dementia. Alzheimers Dement. *16*, 1078–1094. https://doi.org/10.1002/alz.12123.

56. Liu, C.C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat. Rev. Neurol. *9*, 106–118. https://doi.org/10.1038/NRNEUROL.2012.263.

57. Neu, S.C., Pa, J., Kukull, W., Beekly, D., Kuzma, A., Gangadharan, P., Wang, L.S., Romero, K., Arneric, S.P., Redolfi, A., et al. (2017). Apolipoprotein E genotype and sex risk factors for Alzheimer disease: a meta-analysis. JAMA Neurol. *74*, 1178–1189. https://doi.org/10.1001/JAMANEUROL.2017.2188.

58. EFSA Panel on Food Additives and Nutrient Sources added to Food ANS, Younes, M., Aggett, P., Aguilar, F., Crebelli, R., Dusemund, B., Filipič, M., Frutos, M.J., Galtier, P., Gott, D., Gundert-Remy, U., et al. (2018). Scientific opinion on the safety of green tea catechins. EFSA J. *16*, e05239. https://doi.org/10.2903/j.efsa.2018.5239.

59. de la Torre, R., de Sola, S., Hernandez, G., Farré, M., Pujol, J., Rodriguez, J., Espadaler, J.M., Langohr, K., Cuenca-Royo, A., Principe, A., et al. (2016). Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD): a double-blind, randomised, placebo-controlled, phase 2 trial. Lancet Neurol. *15*, 801–810. https://doi.org/10.1016/S1474-4422(16)30034-5.

60. Rami, L., Mollica, M.A., García-Sanchez, C., Saldaña, J., Sanchez, B., Sala, I., Valls-Pedret, C., Castellví, M., Olives, J., and Molinuevo, J.L. (2014). The subjective cognitive decline questionnaire (SCD-Q): a validation study. J. Alzheimer's Dis. *41*, 453–466. https://doi.org/10.3233/JAD-132027.

61. Moeller, J. (2015). A word on standardization in longitudinal studies: don't. Front. Psychol. *6*, 1389. https://doi.org/10.3389/fpsyg.2015.01389.

62. Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., and Elvevåg, B. (2020). Applying speech technologies to assess verbal memory in patients with serious mental illness. npj Digit. Med. *3*, 33. https://doi.org/10.1038/s41746-020-0241-7.

63. Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. (2003). User acceptance of information technology: toward a unified view. MIS Q. Manag. Inf. Syst. *27*, 425–478. https://doi.org/10.2307/30036540.

64. Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale, Revised (Psychological Corporation).

65. Wechsler, D. (2008). Wechsler Adult Intelligence Scale—Fourth Edition Administration and Scoring Manual (Pearson).

66. Jardim De Paula, J., Teixeira De Ávila, R., De Souza Costa, D., Nunes De Moraes, E., Bicalho, M.A., Nicolato, R., Corrêa, H., Sedó, M., and Fernandes Malloy-Diniz, L. (2011). Assessing processing speed and executive functions in low educated older adults: the use of the five digits test in patients with Alzheimer's disease, mild cognitive impairment and major depressive disorder. Clin. Neuropsychiatry *8*, 339–346.

67. Stroop, J.R. (1935). Studies of interference in serial verbal reactions. J. Exp. Psychol. *18*, 643–662. https://doi.org/10.1037/h0054651.

68. Buschke, H. (1984). Cued recall in amnesia. J. Clin. Neuropsychol. *6*, 433–440. https://doi.org/10.1080/01688638408401233.

69. Wechsler, D. (2009). Wechsler Memory Scale IV (WMS-IV) (Psychological Corporation).

70. Dunn, J.C., Almeida, O.P., Barclay, L., Waterreus, A., and Flicker, L. (2002). Latent semantic analysis: a new method to measure prose recall. J. Clin. Exp. Neuropsychol. *24*, 26–35. https://doi.org/10.1076/jcen.24.1.26.965.

71. Peña-Casanova, J., Quiñones-Ubeda, S., Gramunt-Fombuena, N., Quintana-Aparicio, M., Aguilar, M., Badenes, D., Cerulla, N., Molinuevo, J.L., Ruiz, E., Robles, A., et al. (2009). Spanish multicenter normative studies (NEURONORMA project): norms for verbal fluency tests. Arch. Clin. Neuropsychol. *24*, 395–411. https://doi.org/10.1093/arclin/acp042.

72. Casals-Coll, M., Sánchez-Benavides, G., Meza-Cavazos, S., Manero, R.M., Aguilar, M., Badenes, D., Molinuevo, J.L., Robles, A., Barquero, M.S., Antúnez, C., et al. (2014). Spanish multicenter normative studies (NEURONORMA project): normative data and equivalence of four BNT short-form versions. Arch. Clin. Neuropsychol. *29*, 60–74. https://doi.org/10.1093/arclin/act085.

73. Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., and Chertkow, H. (2005). The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. *53*, 695–699. https://doi.org/10.1111/J.1532-5415.2005.53221.X.

74. Oakland, T., and Harrison, P. (2008). ABAS-II. Clinical Use and Interpretation (New York, NY: Springer). https://doi.org/10.1016/B978-0-12-373586-7.X0001-X.

75. Skevington, S.M., Lotfy, M., and O'Connell, K.A.; WHOQOL Group (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial a Report from the WHOQOL Group. Qual. Life Res. *13*, 299–310. https://doi.org/10.1023/B:QURE.0000018486.91360.00.

76. Kline, R. (2015). Principles and Practice of Structural Equation Modeling 4th Editio (Guilford Press).

77. Nunnally, J., and Bernstein, I. (1994). Psychometric Theory (MGraw-Hill).

78. Mackinnon, S., Curtis, R., and O'Connor, R. (2022). Tutorial in longitudinal measurement invariance and cross-lagged panel models using lavaan. Meta-Psychology *6*. https://doi.org/10.15626/mp.2020.2595.

79. Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., and de Vet, H.C.W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. *63*, 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.

80. Husted, J.A., Cook, R.J., Farewell, V.T., and Gladman, D.D. (2000). Methods for assessing responsiveness: a critical review and recommendations. J. Clin. Epidemiol. *53*, 459–468. https://doi.org/10.1016/S0895-4356(99)00206-1.

81. Cohen, J. (1992). A power primer. Psychol. Bull. *112*, 155–159.

82. Sawilowsky, S.S. (2009). New efect size rules of thumb. J. Mod. Appl. Stat. Methods *8*, 597–599.

83. Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Arch. Clin. Neuropsychol. *27*, 248–261. https://doi.org/10.1093/arclin/acr120.

84. McSweeny, A.J., Naugle, R.I., Chelune, G.J., and Lüders, H. (1993). "T Scores for Change": an illustration of a regression approach to depicting change in clinical neuropsychology. Clin. Neuropsychol. *7*, 300–312. https://doi.org/10.1080/13854049308401901.

85. Grimm, K.J., Ram, N., and Estabrook, R. (2017). Chapter 7. Growth mixture modeling. In Growth modeling : structural equation and multilevel modeling approaches, D.A. Kenny and T.D. Little, eds. (Guilford Press).

86. Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likehood in multivariate Gaussian mixture models. Comput. Stat. Data Anal. *41*, 561–575. https://doi.org/10.1016/S0167-9473(02)00163-9.

87. Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. J. Stat. Softw. *48*, 1–36. https://doi.org/10.18637/jss.v048.i02.

88. Pinheiro, J., Bates, D., DebRoy, S., and D, S.; R Core Team (2020). nlme: Linear and Nonlinear Mixed Effects Models. https://cran.r-project.org/package=nlme.

89. Moritz, S., Gatscha, S., and Wang, E. (2021). Package "imputeTS". Time Series Missing Value Imputation 10.32614/RJ-2017-009.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data and code	This paper; Mendeley Data	Mendeley Data: https://data.mendeley.com/datasets/xzr9b4skm7/1
Software and algorithms		
R version 4.2.1	The R Foundation for Statistical Computing	https://www.r-project.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Rafael de la Torre (rtorre@imim.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw de-identified data have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the key resources table.

All original code has been deposited at Mendeley and is publicly available as of the date of publication. The DOI is listed in the key resources table.

Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study design and population

Prospective cohort study design including a subset of N=56 participants (23 male and 33 female between 60-80 years old) from the PENSA Study[39] who simultaneously engaged in a multimodal intervention on diet, physical activity and cognitive training. The PENSA Study is a clinical trial included in the World Wide FINGERs network[55] that aims to assess the efficacy of a personalized multimodal intervention in slowing down cognitive decline and improving brain connectivity.[39] The PENSA Study will finish on 2023. At the time of the present study (August 2022), N=56 participants had already finished the 12 months of multimodal intervention. The 'calibration sample' was composed by the baseline data and the 'validation samples' were composed by the six-months and twelve-months data.

All PENSA Study participants have unimpaired cognition at baseline according to a standardized neuropsychological evaluation[1] and have increased risk of AD. Accordingly, they are *APOE*-ε4 carriers[56,57] and they meet subjective cognitive decline (SCD) criteria (based on a positive answer to the question 'Have you experienced a decrease in your cognitive ability compared to a few years ago?'),[16] and fulfil at least two additional SCD "plus" criteria.[17]

Ethics statement

All participants gave written informed consent. The study protocol (2018/8179) was approved by the local institutional review board (Parc de Salut Mar Clinical Research Ethics Committee CEIm-PSMAR) and adheres to standards of the WAMA Declaration of Helsinki (Brazil, October 2013). The PENSA Study is registered in ClinicalTrials.gov (NCT03978052).

PENSA study multimodal intervention

The multimodal intervention of the PENSA Study lasts 12 months and includes dietary counselling, physical activity and cognitive training and stimulation, and can be supplemented with epigallocatechin gallate (EGCG, 400 to 600 mg/day prior to meals) or placebo.[39] EGCG is a green tea flavanol with antioxidant properties that has shown to be safe in the doses proposed[58] and, combined with cognitive training, has shown to improve the executive functioning performance and the adaptive functionality.[59] The allocation to the 'multimodal intervention + EGCG' or the 'multimodal intervention + placebo' groups is randomized (balanced by sex) and it is double-blinded until the end of the study.

Recruitment strategy

N=34 participants (60.7%) learned about the study through the mass media. Accordingly, a press conference of the PENSA Study took place on December 2019 and a press release was covered by a number of newspapers, radios and TV channels. The launchment of the PENSA Study was also announced in the Barcelonaβeta Brain Research Center (BBRC) and the Hospital del Mar information channels (e.g. web, Twitter, Instagram). Advertisements were also placed on the underground (public transport), Twitter and Facebook, and infographic materials of the study (roll-up and information brochures) were also distributed to associations of older adults, cultural institutions and pharmacies. A call centre provided information of the PENSA Study during all the recruitment period. On the other hand, N=2 participants (3.6%) were derived from the Neurology Service of Hospital del Mar (Barcelona, Spain). Finally, N= 20 participants (35.7%) were already included in registers of participants from the BBRC and were invited to participate in the PENSA Study by a phone call from the BBRC. Therefore, there is a risk of self-selection bias because these participants had prior research experience.

All individuals interested in participating in the PENSA Study were invited to fulfil the online form of the web https://pensaalzheimer.org/. Registered individuals were filtered according to their age and subjective cognitive decline (SCD) status. *A priori* eligible participants were invited to perform a short face-to-face pre-screening visit for the collection of buccal swab for *APOE* genotyping. *APOE*-ε4 carriers underwent to a second screening visit that included neuroimaging tests, neurological examination and neuropsychological assessment.[1,60] Only participants with cognitive performance within normal values and no abnormalities in the neuroimaging tests were eligible.[39]

Inclusion and exclusion criteria

Inclusion criteria included male or female subjects aged 60 to 80 years with SCD (based on a positive answer to the question Have you experienced a decrease in your cognitive ability [e.g., memory, concentration, planning, orientation, or language) compared to a few years ago]?) and *APOE*-ε4 carriers (either hetero or homozygotes), fulfilling at least two additional SCD "plus" criteria (memory complaints rather than other domains of cognition, onset of symptoms within the last 5 years, concern about symptoms, perception of lower performance compared to same age group and/or confirmation of symptoms by an informant).[16,17] Exclusion criteria included (i) history of neurological or psychiatric conditions according to Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria, (ii) clinically significant abnormalities in laboratory test, (iii) any contraindication for brain MRI, (iv) presence of mild to moderate leukoaraiosis (scoring <3 on Fazekas scale (Fazekas, 1987), and/or less than three lacunar infarcts not localized on strategic territory [e.g., bilateral thalamic]), (v) primary or recurrent malignant disease treated within the last 2 years, (vi) evidence of medical conditions/medications that may interfere with study assessments, (vii) body mass index < 18.5 or ≥35 kg/m^2, or (viii) current intake of vitamins or products containing EGCG supplements for at least 3 months previous to the screening visit.

Timing

The 56 participants entered the study progressively. After randomization, individuals were pooled in groups of 9 to 14 people allocated to the multimodal intervention group. The group size was based on optimal ratios for behaviour-change group interventions, which allows optimal interactions between participants as well as promotes social change processes. Group 1 (N=11) started the multimodal intervention on 06/07/2020, group 2 (N=9) on 23/11/2020, group 3 (N=13) on 03/05/2021, group 4 (N=13) on 31/05/2021 and group 5 (N=14) on 12/07/2021. The usability survey about NUP was answered between the 10/11/2021 - 23/11/2021. The NeuronUP cognitive training database was downloaded on 28/06/2022 and the electronic case report form (eCRF) database was downloaded on 19/08/2022.

Data exclusions

N=4 participants from group 1 to group 5 had withdrawn from the study and were not included in the present report.

Non-participation

N=1 from group 4 at month 7 due to prohibited medication use (donazepile). N=1 from group 2 at month 7 for personal reasons. N=2 for medical reasons: 1 participant from group 4 at month 7 and 1 participant from group 5 at month 5, both due to breast cancer diagnosis (exclusion criteria).

METHOD DETAILS

Cognitive training

The cognitive training program is delivered through a digital platform, NeuronUP© (NUP) that offers neurorehabilitation materials for cognitive stimulation to professionals. The training plan for the PENSA study is designed by experienced neuropsychologists and includes 36 different NUP activities that cover different cognitive domains relevant to the AD' related cognitive impairment profile. Specifically, the program contains 8 activities of executive functions, 6 of memory, 6 of language, 6 of attention, 7 of visuospatial abilities and 3 of orientation. These 36 activities are monthly distributed in 12 sessions (2-3 sessions/week) of about 30 minutes each. Therefore, each cognitive training session includes 3 activities of about 10 minutes each that exercise different domains, and each activity is performed once/month.

Participants complete the sessions remotely using their own computer or tablet. Before starting the intervention, they receive a face-to-face training session to learn how the NUP platform works and they also perform two short test sessions to familiarise with the functioning of the tasks. Most activities have 9 or 12 different difficulty levels and are called 'games'. However, some activities do not automatically change the difficulty level and are called 'worksheets' or 'generators' (10 out of 36). In the case of games, the starting level is predefined by the investigators at level 2-3 (low-medium difficulty) and, depending on the participant performance, the difficulty of the game increases, decreases or does not change. Accordingly, to move up a phase (increase game's difficulty) it is necessary to correctly complete 5 exercises, whereas to move down a phase (decrease game's difficulty) it is necessary to fail 3 exercises. Therefore, each month the same pattern of games is administered but the difficulty can vary, as each game starts at the maximum difficulty level achieved in the previous month, which reduces practice effects.

Measures

NUP measures of executive functions

Researchers can request a database with raw NUP data, which includes the following variables: participant identifier, date, name of the activity, phase number (difficulty level), duration, and number of correct answers, errors and attempts. Depending on the activity, additional variables are included (e.g. reaction time). A total of 8 NUP games that targeted executive functions were analysed (Table 1 and Video S1). Déjà vu: Several scenarios appear in which various elements are presented. The activity consists of identifying which elements have appeared more than once (Table S13). Sorting Bugs: The screen is divided in two by a bar with a hole and there are elements of two different types that move around the screen. The activity consists of getting all the elements of one type on one side of the screen and all the elements of the other type on the other side of the screen (Table S14). Balance the Bags: A person appears at the supermarket checkout and must put all the products in bags. The activity consists of calculating the weight of each product so that each arm carries the same weight (Table S15). Home Delivery: Several buildings appear and are illuminated in turn. The activity consists of remembering the order in which they have been illuminated and reproducing it in reverse order (Table S16). Stop the Ball: A circle appears on the screen which is traversed by a ball at a constant speed. The activity consists of clicking when the ball passes a specific point on the circle (Table S17). Card Pyramid: Several cards appear on the screen forming a figure. The cards at the ends are faced up, the rest are faced down. In addition, there are one or two decks with one card faced up. The activity consists of placing the cards available in the figure on top of the deck in ascending and descending order. As they are placed, the cards that were faced down next to the one that has just been placed will be turned over, thus becoming available (Table S18). Knitting a Scarf: A piece of cloth with knitting needles appears on the screen and, in another place, a ball of wool. The activity consists of reaching for the ball of wool that appears to make the loop longer. Each time you catch one, another one will appear. It is essential not to bump against the edges or against the scarf itself, which will be longer and longer (Table S19).

Jigsaw puzzle: Several loose pieces appear on the screen. The activity consists of putting them together to form a complete image (Table S20).

In order to control for practice effects, the first two months served to familiarise participants with the games instructions and function and were not considered for the analysis. Performance during the third month was used to estimate the baseline capacity of each participant, since it was assumed that after 3 months participants had had sufficient time to reach their optimal performance level in each game. Therefore, baseline data is located at month 3 instead of month 1. In turn, performance during months 4-12 was used to monitor cognitive change over time. Accordingly, we designed a scoring system to evaluate the monthly performance of each individual on each game. Basically, each NUP score was constituted by an integer (the phase number representing the most frequently played difficulty-level for each subject in each game), and a decimal that, depending on the game specifications, was based on the number correct exercises, the number of failed exercises or the time spent. The use of the most frequently played difficulty-level for each subject on each game was expected to be a conservative criterion for assessing cognitive change. Moreover, the inclusion of a decimal increases the variability of scores within the same difficulty level. The suitability of each score was evaluated by a panel of researchers integrated by 4 neuropsychologists. Scores were finally standardized using the proportion of maximum scaling method[61] ($POMS = \frac{(observed - minimum)}{(maximum - minimum)}$) and values are reported as percentages.

The rationale behind the selection of executive functions games was based on the available panel of traditional neuropsychological measures of executive functions that were also administered in the PENSA Study, so we these two measures of executive functions can be compared. On the contrary, NUP memory games were not analysed because they are focused on visual memory, whereas the available traditional tests of memory in the PENSA Study measure verbal episodic memory, so they are not comparable measures of memory. Verbal episodic memory games are currently more difficult to design since they need to incorporate speech technologies to count units of information recalled, starting by the administration of the task, the transcription of voice to text, and the automating rating of the transcript to simulate expert human ratings.[62]

NUP adherence and acceptability

Monthly adherence to the cognitive training intervention was calculated for each participant by dividing the number of sessions of cognitive training performed each month by the number of sessions programmed each month (a total of 12 sessions). A modified version of the Unified Theory of Acceptance of Use of Technology (UTAUT) questionnaire[63] was administered to participants, who answered to the survey on a voluntary basis.

Traditional neuropsychological tests

Executive functions were assessed with the WAIS Digit Symbol Substitution Test total score,[64] the WAIS Visual Puzzle Test total score,[65] the WAIS Digit Span Test backwards score,[65] the Five Digits Test (FDT) flexibility score,[66] and the word-colour score from the Stroop Colour and Word Test.[67] Memory was assessed with the Free and Cued Selective Reminding Tests (FCSRT) immediate free recall (IFR) and delayed free recall (DFR) scores[68] and the WMS Logical Memory (LM) sub-test immediate recall (IR), delayed recall (DR) and recognition scores.[69] Global cognition was assessed with the Alzheimer Disease Cooperative Study Preclinical Alzheimer Cognitive Composite (ADCS-PACC),[1] which is the primary outcome of the PENSA Study. Moreover, a modified version of this composite including the Interference score from the Stroop Colour and Word Test and the Flexibility score from the FDT, resulting in the ADCS-PACC-plus-exe, is also considered as a primary outcome of the PENSA Study. Additional cognitive measures included the Mini-Mental State Exam (MMSE) total score[2] and the Montreal Cognitive Assessment Test (MoCA) total score,[3] that were also used to evaluate global cognition, and the Semantic Verbal Fluency Test of Animals ('Animals Fluency test') and the Boston Naming Test. Scores were standardized (z-scores) on baseline mean and standard deviation. Composite scores of memory (abbreviated as 'NPS-MEM') and executive functions (abbreviated as 'NPS-EXE') were created by averaging the Z scores of tests.

The FCSRT[68] is a widely used measurement of verbal episodic memory, designed to dissociate the different processes involved in the formation of new memories. It consists of a list of 16 written words that the examinee should memorise. Each word belongs to a different semantic category. The examiner provides category cues to promote deep, controlled information processing. The task includes 6 different phases. (1)

Reading and identification of words. (2) Interference task to prevent subvocal repetition, by performing a serial subtraction task for 20 seconds. (3) Free recall, where the examinee is asked to say as many words as he or she can remember, in any order, with a time limit of 90 seconds. (4) Cued recall, immediately after each free recall trial, where the examinee completes a cued recall task for the items that is unable to recall spontaneously, as the examiner provides a category cue for each word. (5) Selective recall of non-recalled words, only completed in the first two trials, where the examiner provides the items that were not recalled with cues. (6) Delayed free and cued recall, approximately 30 minutes ($\pm$5) later, where the examinee completes another free and cued recall trial. After completing the first three learning trials, the examinee is informed that he or she will be asked to recall the words at a later time. Phases 2–5 are repeated three times during the learning process.

The LM[69] is considered a useful and effective measure of episodic memory, as it addresses three processes involved in memory: encoding, storage and recall. It is sensitive for detecting cognitive decline in early dementia, since prose recall depends upon a range of high-level cognitive functions such as episodic memory, conceptual organization, and schema formation.[70] It consists of three parts: LM I (immediate recall), LM II (delayed recall), and LM Recognition (delayed recognition). In the LM I, subjects are required to immediately recall details of two short passages. In the LM II, subjects are asked to recall the passages after a 20 to 30-minute delay. In the LM Recognition, subjects are asked to answer yes/no questions regarding the passages learned earlier.

The Digit Symbol Substitution Test from the Wechsler Adult Intelligence Scale–Revised[64] is sensitive to both the presence of cognitive dysfunction and change in cognitive function, across a wide range of clinical populations. It consists of a paper-and-pencil cognitive test presented on a single sheet of paper where the examinee is asked to fill in the correct symbols into the spaces below the numbers, by matching them according to a key located on the top of the page. The number of correct symbols within 120 seconds constitutes the total score.

The Mini Mental State Examination (MMSE)[2] is a widely used screening test for cognitive impairment in older adults. It takes between 7 and 10 minutes to complete and contains items that assess orientation, registration, attention and calculation, recall, language, repetition, reading, writing, comprehension of commands, and drawing. Scores range between 0 and 30 points. The Stroop Colour and Word Test[67]: This test consists of three printed sheets with 100 words in each, distributed in 5 columns. Participants are allowed to read each sheet for 45 seconds and the total number of words read is recorded. Errors are discounted for the total of words in each part. Four scores are obtained: W (number of words correctly read in the first sheet), C (number of colours correctly named in the second sheet), WC (number of items correctly named in the third sheet), and interference index (calculated with the following formula: $Interference = WC - \frac{W \times C}{W + C}$).

The Five Digit Test (FDT)[66] is a multilingual, non-reading test that minimizes the effects of education and social class, and allows the testing of some severe clinical cases, who may not be able to read words or name colours. The FDT quickly measures mental processing speed and the ability to direct and switch the attentional control. It is composed of four subtests. In part 1 (Reading) participants are asked to read the digits presented in a series of text boxes, each containing as many repetitions of the digit as it indicates itself. In part 2 (Counting), the boxes contain asterisks and participants are asked to state the number of asterisks in each box. In part 3 (Focusing), the boxes are similar to those in part 1, except that the number identity does not correspond to the amount of digits in the box. Participants are then asked to state the number of digits by ignoring their identity. In part 4 (Switching), an extra clue indicates whether the participant must state the number of digits or their identity (reading or counting). In each section of the test, performance is measured in terms of the time required to complete the task. Inhibition (Focusing minus Reading) and Flexibility (Switching minus Reading) scores are calculated to measure working memory components related to the executive system.

The Semantic Verbal Fluency 'Animals' test[71] entails the generation of words from a given category (animals) within a pre-set time of 60 seconds. The Boston Naming Test[72] (BNT): It is the most widely used test of visual confrontation naming. The reduced 15-item version of the BNT is the one that we used. The subject is asked to name each object correctly within a maximum of 20s. Semantic or phonemic cues are provided when necessary. According to the standard test criteria, the score is calculated from

those items that are correctly named spontaneously plus additional items correctly named after semantic cues. Its administration takes between 3 and 5 minutes.

The Digit span sub-test from Wechsler Adult Intelligence Scale IV[65] requires subjects to repeat series of digits of increasing length. It includes three tasks: Digit Span Forward (DSF), Digit Span Backward (DSB), and the new Digit Span Sequencing (DSS). DSF is a good measure of simple attention and short-term memory. DSB and DSS represent a qualitatively different type of task that relies more upon working memory skills. Its administration takes between 5 and 8 minutes.

The Visual Puzzle Test from Wechsler Adult Intelligence Scale IV[65] is designed to measure nonverbal reasoning and the ability to analyse and synthesise abstract visual stimuli. The test consists of 26 puzzles that are presented complete to the examinee. The examinee is asked to select three pieces to build the presented puzzle. The estimated application time is between 5 and 10 minutes.

The Montreal Cognitive assessment (MoCA)[3] is a screening assessment for detecting cognitive impairment with a high charge of executive function task, found to be useful in the detection of patients with cognitive impairment at higher risk for incident dementia.[73]

The Alzheimer Disease Cooperative Study Preclinical Alzheimer Cognitive Composite (ADCS-PACC)[1] is designed to serve as the primary outcome measure for trials conducted at the asymptomatic phase of Alzheimer's disease, as such it has demonstrated its feasibility for measuring cognition in normal elderly participants with evidence of Alzheimer's disease pathology. The ADCS-PACC is composed by the total immediate recall from the FCSRT,[68] the total delayed recall from the WMS Logical Memory II sub-test,[69] the WAIS Digit Symbol Substitution Test[64] and the MMSE total score.[2] The ADCS-PACC-plus-exe is a modified version of the ADCS-PACC composite including the Interference score from the Stroop Colour and Word Test[67] and the Flexibility score from the FDT.[66]

Functionality and quality of life

Functionality was assessed with The Adaptive Behaviour Assessment System – Second Edition (ABAS-II).[74] The ABAS-II for adults (ages 16 to 89) includes 239 items that assess the individual's competence (in terms of behaviour frequency) in 10 different skill areas: (i) *communication abilities* (i.e. to talk, listen, engage in conversation and provide a response), (ii) *community use* (i.e. use of community resources such as shopping or getting around the neighbourhood), (iii) *functional academics* (i.e. skills related to reading, writing, mathematics and other areas necessary for independent daily functioning), (iv) *home living* (i.e. home care skills such as tidying, cleaning, repairing, and caring for objects), (v) *health and safety* (i.e. skills related to maintaining an adequate state of health, such as respecting safety rules, using medicines and showing caution), (vi) *leisure* (i.e. participation in recreational activities, compliance with the rules of the games, and leisure planning), (vii) *self-care* (i.e. activities related to food, clothing and hygiene), (viii) *self-direction* (i.e. performing tasks, complying with deadlines and time constraints, following instructions and other activities involving responsibility and self-monitoring), (ix) *social interaction* (i.e. interacting socially, getting along with others, making friends and maintaining friendships, showing good manners and communicating one's emotions) and (x) *working/labour skills* (functional skills to perform successfully on the job, including performing assigned tasks and complying with schedules and instructions). Scalar scores for each domain were computed with a mean of 10 and a standard deviation of 3. These scores were used to obtain 3 subscales: Conceptual (including communication abilities, functional academics and self-direction domains), Social (including leisure and social domains) and Practical (including all the other domains), as well as a General Adaptive Composite. These indices have a mean of 100 and a SD of 15. All answers to this questionnaire are reported by an informant. Those items rated as 'guessed' by the informant were scored as zero, in order to avoid subjective judgments concerning functional changes. Higher scores indicate higher adaptive skills and independency in everyday living. Given that most individuals in our sample were already retired, scores in the work skill area were not included in the analyses and were not considered for calculating the General Adaptive Composite.

Quality of life was evaluated with the World Health Organization Quality of Life brief generic questionnaire (WHOQOL-BREF).[75] The WHOQOL-BREF is a cross-culturally sensitive self-reported measure of quality of life in the previous two weeks. It includes 26 items that are grouped in four domains: physical (pain, energy, sleep, mobility, activities, medication and work), psychological (positive feelings, cognitions, self-esteem,

body image, negative feelings and spirituality), social relationships (personal relations, social support and sex life) and environment (safety and security, home environment, finance, health/social care, information, leisure, physical environment and transport). It also contains quality of life and general health items. Each item is scored from 1 to 5. Higher scores indicate higher quality of life.

Sociodemographic characteristics

The following baseline factors were included: gender, age, years of education and cognitive reserve, evaluated with the cognitive reserve questionnaire (CRQ),[34] that comprises eight questions about education, employment, languages, musical education, reading habits and use of intellectual games (e.g. chess, puzzles) whose total score serves as a proxy for cognitive reserve.

QUANTIFICATION AND STATISTICAL ANALYSIS

General overview

First, we assessed the acceptability of NUP according to participants' adherence to the intervention and their responses to a usability survey. Second, for the assessment of the psychometric properties (reliability and validity), we treated NUP scores as items of a scale measuring executive functions (NUP-EXE). Accordingly, we first performed a univariate description of individual scores (Figure S5) and we examined the matrix of correlation between scores in order to identify those items that did not correlate with the other items. We then tested the factor structure of NUP-EXE, analysed its reliability and metric invariance over time, and examined the convergent/discriminant validity and the sensitivity to change. We also evaluated the psychometric properties of traditional measures of executive functions and memory before using them to test the convergent and discriminant validity of NUP-EXE. Moreover, we tested the hypothesised relationships between NUP-EXE and sociodemographic, quality of life and functional factors. Most of these analyses were initially performed in the calibration sample and then attempted to be replicated in the validation samples. This approach was applied to avoid splitting the population into two random sub-samples. Finally, to explore the potential of NUP in the study of the inter-individual variability in treatment response after preventive interventions for AD, we compared the sensitivity to change of NUP measures and traditional cognitive measures and we applied latent class growth modelling techniques to identify subgroups of subjects with different NUP-EXE trajectories. Ultimately, we tested gender differences in NUP-EXE trajectories.

Factorial validity, reliability and measurement invariance

The factorial structure and reliability of NUP-EXE were evaluated with gold-standard structural equation modelling (SEM) techniques.[76] In SEM, all variables were centred (deviations from the mean) in order to focus on covariance structures and do not deal with mean structures. The factorial validity of NUP-EXE was tested using confirmatory factor analysis (CFA). Then, construct reliability was examined by calculating the ratio of explained to unexplained variance across indicators of each construct with the index H $\left(H = 1 / \left[1 + \frac{1}{\sum_{i=1}^{k} \frac{\lambda i^2}{1 - \lambda i^2}}\right]\right)$; where λ is the fully standardized factor loading. H varies from 0 to 1, with higher values indicating that the latent variable is empirically well defined and will be reproducible across studies. The degree of interrelatedness among items of NUP-EXE was also assessed using the Cronbach's alpha (α) coefficient. General reliability estimates thresholds of >0.90 (excellent) and 0.70-0.90 (good/substantial) were used.[77] The accuracy of NUP-EXE was evaluated with the standard error of the measurement (SE_m), which is the variation around a true score for an individual when repeated measures are taken. It is calculated with the following formula: $SE_m = SD \times \sqrt{(1 - r)}$; where SD is standard deviation and r is the Cronbach's α coefficient. The smaller the SE_m, the more accurate are the assessments. Finally, the longitudinal measurement invariance of NUP-EXE was also tested using SEM techniques. Measurement invariance is achieved in a study when participants across all time periods interpret the individual questions and the underlying latent factor in the same way.[78] Specifically, measurement invariance was tested over three measurement occasions: baseline, 6 months and 12 months. The least stringent level of invariance called 'configural invariance', which establishes the same factor structure over time, was compared to 'metric or weak invariance', which constraints factor loadings across time, so items do not become more or less representative of the latent construct at different time points.

SEM models were estimated using maximum likelihood estimation with robust (Huber-White) standard errors and a scaled test statistic that is (asymptotically) equal to the Yuan-Bentler test statistic. Overall model fit was evaluated with 'global' and 'approximate' fit indexes, including the chi-squared (χ^2) test statistic, the comparative fit index (CFI), the robust mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR). Higher probability values of χ^2 (>0.05 cut-off) indicate greater likelihood of the null hypothesis of perfect fit of the model. Moreover, higher values of CFI (>0.90) and lower values of SRMR (<0.08) and RMSEA (<0.05) are indicative of good or better model fit. Nested models were also compared with Bayesian information criterion (BIC), with lower BIC indicating better fit.

Convergent and discriminant validity

Convergent validity occurs when measures of the same trait evaluated with different methods are correlated, whereas discriminant validity occurs when measures of different traits evaluated with different methods are not correlated. Accordingly, we expected NUP-EXE to be more strongly correlated with traditional measures of executive functions (i.e. the executive functions composite abbreviated as 'NPS-EXE') than with traditional measures of memory (i.e. the memory composite abbreviated as 'NPS-MEM'). We also expected correlations between NUP-EXE and traditional measures of global cognition (e.g. the ADCS-PACC, the MMSE or the MoCA). The factorial structure, reliability and measurement invariance of the traditional composites NPS-EXE and NPS-MEM was first tested using the SEM techniques described above. Discriminant validity of NPS-MEM and NPS-EXE was tested by showing that the average variance extracted ($AVE = \frac{\sum_{i=1}^{n} \lambda^2}{n}$) exceeded the amount of variance shared with other factors quantified by the squared factor intercorrelation coefficient (that is, the squared of the fully standardised factor correlation coefficient). Ultimately, Pearson's correlations were used to examine NUP-EXE convergence/discriminant validity in relation to traditional cognitive tests, as well as to test the hypothesised correlations between NUP-EXE and sociodemographic factors, functionality and quality of life.

Sensitivity to change

The responsiveness or sensitivity to change is the degree of an instrument to detect change over time in the construct to be measured.[79] Internal change sensitivity is the ability of a measure to change over a particular prespecified time frame, whereas external change sensitivity is the extent to which changes in a measure over a specified time relate to corresponding changes in a reference measure.[80] On the one hand, internal change sensitivity of NUP-EXE was evaluated in intervals of two months (baseline *vs.* 5 months, 5 *vs.* 7 months, 7 *vs.* 9 months and 9 *vs.* 11 months) using paired t-test and Cohen's d effect size statistics, with cut-off values of 0.2 (small effect), 0.5 (moderate effect), 0.8 (large effect) and 1.2 (very large).[81,82] On the other hand, external change sensitivity was evaluated by testing the correlation between change in NUP scores after 6 and 12 months and the respective change in traditional cognitive tests, as well as in measures of functionality and quality of life. Reliable change indexes from standardized regression-based formulas (RCI_{SRB}) were used to estimate change in traditional cognitive tests.[83,84] First, '*T* scores' were estimated separately using linear regression models with baseline scores (T_0) as predictors, using the formula $T_1' = bT_0 + c$, with T_1' indicating the predicted T_1 score, *b* representing the regression slope and *c* the regression intercept. Then, RCI_{SRB} were calculated as $RCI_{SRB} = (T_1 - T_1')/SEE$, where SEE is the standard error of the estimate of the regression equation. Compared to simple discrepancy scores ($T_1 - T_0$), RCI_{SRB} consider the distribution of baseline scores and provide a more precise estimate of relative change by correcting for practice effects, test-retest reliability and variability in T_1 scores.[83] Finally, the sensitivity to change of NUP measures and traditional cognitive measures at 6- and 12-months was compared using Cohen's d effect size statistics and unadjusted linear mixed effects models.

NUP-EXE trajectories

To describe how change in NUP-EXE proceeded during the course of the PENSA Study multimodal intervention, growth mixture models (GMMs) were used to test whether there was evidence that between-person differences in NUP-EXE were better represented by considering more than one typology or trajectory. GMMs search for classes or groups of individuals, such that individuals thought to be in the same class have similar growth trajectories, whereas individuals thought to be in different classes have sufficiently different growth trajectories.[85] Six different models with different distributions of intercepts and slopes over classes were tested, and each of these six models was assumed to have 1 to 3 latent classes, so a total of 18 models were compared. The optimal starting values for GMMs were estimated from a grid of random initial values from the 1-class model with 10 iterations from 100 random departures.[86] Once model parameters were

estimated, the posterior estimate of the likelihood that each individual belongs to each class was calculated. The model that better represented the observed data was selected according to lower Bayesian Information Criterion (BIC), higher entropy (uncertainty in class assignment, with higher values indicating clearer delineation of classes) and higher posterior probability values of each class. Between-group differences in intervention adherence and baseline characteristics were tested with one-way analysis of variance (ANOVA).

Finally, gender differences in NUP-EXE at each time point and in change over time were tested. Linear models were used to test cross-sectional differences in NUP-EXE between males and females, and linear mixed effects models were used to test gender differences in the rate of change over time. These models were adjusted by age and years of education.

Missing data

Rates of missing data in NUP scores at each time point are reported. Missing data in NUP scores were imputed by applying a linear interpolation. A visualization of missing value replacements is included in Figure S6. There is no missing data in traditional neuropsychological tests or in measures of quality of life. However, given that the ABAS-II was completed by an informant on a voluntary basis and takes about 20-30 minutes to answer its 239 items, the rates of missing data in this questionnaire were N=6 (10.7%) at baseline, N=12 (21.4%) after 6 months and N=25 (44.6%) after 12 months. Missing data in the ABAS-II was assumed to be completely at random so each specific analysis was performed on individuals with complete information on the variables involved.

Software and packages

All the analyses were performed using *R* software version 4.2.1. SEM was conducted using the *R* package 'lavaan'.[87] Mixed effects models were computed with the *R* package 'nlme'.[88] Latent class growth models were computed with the *R* package 'lcmm'. Finally, the univariate imputation of missing values was performed with the *R* package 'imputeTS'.[89]

ADDITIONAL RESOURCES

The PENSA Study is registered in ClinicalTrials.gov (NCT03978052). The protocol of the PENSA Study has been published elsewhere.[39]

Interested researchers can gain access to the cognitive training platform by directly contacting the company NeuronUP© on https://www.neuronup.com/.