Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spontaneous Speech as a Scalable Digital Biomarker for Monitoring Cognitive Change in Older Adults

Jonathan Heitz
University of Zurich
Rudolf Debelak
University of Zurich
Nicolas Langer
n.langer@psychologie.uzh.ch
University of Zurich

Article

Keywords:

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10104180/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spontaneous Speech as a Scalable Digital Biomarker for Monitoring Cognitive Change in Older Adults

Jonathan Heitz[1,2], Rudolf Debelak[3,4], Nicolas Langer[1,5*]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland.
[2]Language & Medicine Competence Centre, University of Zurich, Zurich, Switzerland.
[3]Institute of Education, University of Zurich, Zurich, Switzerland.
[4]Institute of Computer and Communication Sciences, EPFL, Lausanne, Switzerland.
[5]Neuroscience Center Zurich (ZNZ), Zurich, Switzerland.

*Corresponding author(s). E-mail(s): n.langer@psychologie.uzh.ch;
Contributing authors: jonathan.heitz@uzh.ch; rudolf.debelak@ife.uzh.ch;

Abstract

Detecting early cognitive decline at scale requires monitoring tools that are economical, repeatable, and sensitive to small within-person change. Brief cognitive assessments are the standard screening tools, but they are themselves noisy measurements of true cognition. Spontaneous speech is a promising digital biomarker, yet its sensitivity to within-person cognitive change has not been directly benchmarked against cognitive assessment. We extended the Minimal Detectable Change (MDC) framework to machine-learning predictions of latent cognition and applied it to two independent longitudinal cohorts of older adults. Speech-based predictions from three spontaneous speech tasks were compared with cognitive composite scores across language, memory, executive function, and processing speed. The average speech-based prediction across three tasks matched the reference MDC in the most relevant cognitive domains, demonstrating sensitivity to within-person change comparable to brief cognitive assessment. A simulation analysis further identified prediction reliability and regression performance as independently improvable targets. These findings support spontaneous speech as a scalable digital biomarker for monitoring clinically relevant cognitive change in aging.

1 Introduction

The global population is aging rapidly, and the number of people living with dementia is projected to rise from approximately 57 million cases in 2019 to 153 million by 2050 [1]. Earlier identification of cognitive decline is becoming increasingly important because recently approved disease-modifying therapies [2–4] and multidomain lifestyle interventions [5–7] are most likely to be effective when initiated early in the disease course. Yet dementia is substantially underdiagnosed in primary care, with up to 60% of cases remaining unrecognized [8, 9]. Cognitive decline is therefore often detected only after functional impairment or subjective cognitive complaints have become apparent, years after the underlying neuropathology process may have begun [10]. These challenges create an urgent need for scalable tools that enable the early detection of cognitive decline by monitoring cognition repeatedly, economically, and sensitively in large populations of mostly healthy older adults.

Standardized neuropsychological assessment is the current clinical reference standard for estimating cognition, a construct that cannot be observed directly. In practice, comprehensive neuropsychological batteries are time-consuming, require trained examiners, and are infeasible for routine longitudinal monitoring at scale. As a consequence, clinical care and large-scale research rely predominantly on brief cognitive screening instruments such as the Mini-Mental State Examination (MMSE) or the Montreal Cognitive Assessment (MoCA), which are more feasible, but are more limited by a reduced depth and domain coverage, and a more coarse score resolution. More fundamentally, observed scores on any cognitive assessment are imperfect estimates of an individual's true cognition, the latent cognitive abilities they are intended to measure. Performance on a given day reflects not only underlying ability but also measurement noise resulting from transient influences such as focus, fatigue, mood, medication, motivation, time of day, or acute stress [11]. This measurement noise is often large enough that small changes in true cognition cannot be distinguished from random fluctuations between assessments. The Minimal Detectable Change (MDC) formalizes this limit: it defines the smallest within-person change that exceeds measurement error with a specified level of confidence. For the MoCA, the 95%-MDC is approximately 4–5 points on a 30-point scale [12, 13], while the standard clinical cutoff for cognitive impairment is 26. In other words, a change in score is only interpretable as true cognitive decline once it reaches the magnitude of a clinically significant impairment, leaving early, more gradual decline effectively invisible to repeated testing. Frequent re-administration is further constrained by practice effects and a limited availability of validated alternate forms [14]. The instruments most commonly used to monitor cognition in practice are therefore poorly suited to detecting cognitive decline in its earliest stages.

Spontaneous speech has emerged as a promising digital biomarker for cognition. Decades of work have shown that subtle changes in lexical, syntactic, semantic, and acoustic features of speech accompany the earliest stages of Alzheimer's disease (AD) and mild cognitive impairment [15–18], and recent machine learning models can predict continuous cognitive scores from short speech samples with moderate accuracy [19–21]. Compared to cognitive testing, speech can be collected remotely, repeatedly, and at low cost, with minimal participant burden, and reduced susceptibility to practice

effects, making it well suited for longitudinal monitoring. Speech-based predictors of cognition are based on supervised machine learning models trained on cognitive test scores as target labels. These labels are themselves noisy estimates of true cognition. Crucially, since a well-trained supervised regression model learns only the systematic component of its targets and not the random noise [22], speech-based predictions may in principle be less noisy than the cognitive scores they were trained on, and more accurately capture true latent cognition.

Despite this potential, two gaps limit the clinical interpretation and translational value of current speech-based digital biomarkers for cognition. First, previous work has focused predominantly on classifying clinical groups (e.g., dementia vs. cognitively healthy controls), or on cohorts spanning a wide range of cognitive function [23, 24]. In such settings, models can exploit large between-person differences, whereas early monitoring requires sensitivity to much more subtle within-person changes. Only a small number of studies have examined speech-based predictions in more homogeneous, mostly healthy older populations [20, 21, 25, 26], even though these are the populations most relevant for early detection and longitudinal monitoring. Second, the standard evaluation metrics in this literature, such as cross-sectional regression performance or classification accuracy, measure association with noisy targets rather than sensitivity to within-person change in true cognition. These two quantities can diverge substantially, so a candidate biomarker with modest cross-sectional R^2 may still be sensitive to within-person change, and vice versa. Very few studies have examined reliability or longitudinal change, and they have done so separately: prior work has reported the test-retest reliability of speech measures [27, 28], or compared their group-level longitudinal change against clinical endpoints [29], but neither has quantified the smallest within-person change in true cognition detectable from speech, the quantity that determines monitoring sensitivity. As a result, it remains unknown whether current speech-based prediction is already useful for cognitive monitoring, and what improvements would be required to make it clinically viable.

In this study, we addressed these gaps by directly benchmarking the within-person sensitivity of speech-based predictions against that of cognitive assessment, to our knowledge the first such comparison. We extended the classical MDC framework to machine-learning-based predictions of latent cognition, allowing us to estimate the smallest within-person change in true cognition that can be detected from spontaneous speech. We applied this framework to a primary cohort of 305 older adults assessed twice approximately 21 months apart, and replicated the analysis in an independent cohort of 70 participants assessed twice two months apart. The short retest interval in the replication cohort was intentional: because the framework estimates sensitivity from measurement noise, true cognitive change can confound MDC estimation. We additionally used simulation to map how the achievable speech-based MDC depends on two directly observable properties of a regression model: its cross-sectional R^2 and the test-retest reliability of its predictions. Each can be improved through distinct mechanisms, providing two complementary routes towards better speech-based cognitive monitoring.

Table 1: Baseline demographic and cognitive characteristics of the original $n = 1003$ participants who returned for a follow-up assessment (Cohort A, $n = 305$) and those who did not (Non-returners, $n = 698$).

	Non-returners	Cohort A (Baseline)	p-value
# Participants	698	305	
Age	65.49 ± 4.84	65.42 ± 4.72	0.83
Gender			0.12
Female	439 (62.9%)	176 (57.7%)	
Male	259 (37.1%)	129 (42.3%)	
Country			< 0.01
USA	381 (54.6%)	120 (39.3%)	
UK	316 (45.3%)	185 (60.7%)	
Education			0.29
High Education	431 (61.7%)	170 (55.7%)	
Low Education	264 (37.8%)	134 (43.9%)	
Socioeconomic Status	5.57 ± 1.67	5.60 ± 1.55	0.78
Mean Transcript Length (# words)	172.61 ± 126.56	175.96 ± 69.30	0.66
Mean Audio Length (s)	83.24 ± 37.65	82.01 ± 26.73	0.61
Mean Cognitive Score	-0.04 ± 0.79	0.09 ± 0.75	0.02
Language	-0.06 ± 1.01	0.14 ± 0.96	< 0.01
Executive Function	-0.05 ± 1.01	0.11 ± 0.96	0.03
Memory	0.00 ± 1.00	-0.01 ± 1.00	0.86
Speed	-0.04 ± 1.02	0.10 ± 0.94	0.04

2 Results

2.1 Study sample

Our primary analyses are based on *Cohort A*, 305 older English-speaking adults from the USA and UK who completed a longitudinal protocol of spontaneous speech tasks and a standardized cognitive test battery twice, approximately 21 months apart. Cohort A is the returning subset (30.4%) of the $n = 1003$ participants assessed at baseline and described in our previous work [21]. We refer to participants who did not return for the follow-up assessment as *non-returners* ($n = 698$). We compared the baseline characteristics of Cohort A against the non-returners (Table 1). The two groups were comparable on age, gender, education, and socioeconomic status. They differed in country of residence (UK participants returned more often than US participants; $p < 0.01$) and in three of the four composite cognitive scores: returners scored higher than non-returners on *language* ($p < 0.01$), *executive function* ($p = 0.03$), and *speed* ($p = 0.04$), with no significant difference on *memory* ($p = 0.86$). An independent replication cohort, *Cohort B* ($n = 70$), assessed twice over approximately two months, is introduced in Section 2.5.

2.2 Cognitive performance

Composite cognitive scores: Composite cognitive scores were derived from individual test scores using the confirmatory factor analysis (CFA) model established at baseline [21]. To place scores on a common metric across assessments, factor loadings and intercepts were fixed to the baseline estimates (Supplementary Table S14) when computing follow-up scores. We assessed longitudinal measurement invariance between the baseline and Cohort A follow-up assessments. The configural model fitted the data well (robust CFI = 0.929, robust TLI = 0.912, RMSEA = 0.054, SRMR = 0.064). Constraining loadings to equality across occasions (weak invariance) did not degrade fit (scaled $\Delta\chi^2(11) = 5.96$, $p = 0.88$; ΔCFI = +0.002, ΔRMSEA = −0.002, ΔSRMR = 0.001), supporting invariant factor loadings and justifying the use of baseline loadings to derive longitudinally comparable composite scores. Additionally constraining intercepts (strong invariance) yielded a significant chi-square difference (scaled $\Delta\chi^2(10) = 45.14$, $p < 0.001$) but only marginal changes in the descriptive fit indices (ΔCFI = −0.009, ΔRMSEA = +0.002, ΔSRMR = 0.000), within conventional cutoffs, indicating approximate strong invariance.

Stable cognition between baseline and follow-up in Cohort A: Within Cohort A, there was no statistical evidence of group-level cognitive change between baseline and follow-up in any of the four cognitive domains (all p-values ≥ 0.17; Table 2). The proportions of participants showing improved versus declined scores were approximately balanced in each domain, indicating that between-assessment differences were not systematically directional. Despite the absence of systematic group-level change, substantial within-person differences were observed across domains (SD = 0.65–0.84). A symmetric distribution of change scores around zero could reflect either measurement noise around stable cognition, or a combination of true cognitive decline and a counterbalancing practice effect from prior exposure to the protocol. We tested for the latter by comparing follow-up scores in Cohort A against baseline scores of Cohort B participants, who completed the same protocol for the first time (see Section 2.5), using a linear model with a cohort indicator and demographic covariates. The estimated cohort coefficients were small and non-significant for all four cognitive domains (*language*: −0.10, $p = 0.46$; *executive function*: −0.06, $p = 0.68$; *memory*: +0.06, $p = 0.66$; *speed*: −0.08, $p = 0.54$), providing no evidence of a practice effect and masked decline. We therefore treated latent cognition as approximately stable between baseline and follow-up in Cohort A. Although the relatively long retest interval (21 months) would ordinarily make this assumption less secure in older adults, the observed stability is consistent with selective attrition (Table 1), because participants who experienced meaningful cognitive decline after their baseline may have been less likely to return for follow-up.

2.3 Regression performance, test-retest reliability, and MDC estimation

We trained a support vector regression (SVR) model to predict each of the four composite cognitive scores from speech features and demographic variables. One prediction

Table 2: Summary of composite cognitive scores for Cohort A. Scores were standardized to zero mean and unit variance based on the baseline distribution [21]. We report mean and standard deviation of scores in each assessment and their difference.

	Baseline	Follow-up	% Improvement / % Decline	Difference	p-val
Language	0.14 ± 0.96	0.20 ± 0.95	52.5% / 47.5%	0.06 ± 0.75	0.17
Memory	-0.01 ± 1.00	0.03 ± 1.11	53.2% / 46.8%	0.04 ± 0.84	0.39
Speed	0.10 ± 0.94	0.11 ± 0.95	48.5% / 51.5%	0.02 ± 0.65	0.64
Executive Function	0.10 ± 0.97	0.09 ± 0.99	50.5% / 49.5%	-0.01 ± 0.75	0.74

Table 3: Bootstrap test results assessing whether the speech-based MDC is significantly larger than the cognitive assessment's MDC on Cohort A. For each task and composite cognitive score, we report the 95% CI of the difference (speech MDC minus cognitive MDC) across 10000 bootstrap replicates, and the proportion of replicates where the speech MDC is larger (less sensitive) than the cognitive MDC (% larger). An asterisk (*) on the CI indicates that the CI excludes zero, i.e. the speech MDC is significantly larger than the cognitive MDC.

	Language		Memory		Executive Function		Speed	
	95% CI	% larger	95% CI	% larger	95% CI	% larger	95% CI	% larger
Journaling	0.69 – 2.27*	100.0%	0.59 – 2.63*	100.0%	1.18 – 3.16*	100.0%	2.23 – 6.23*	100.0%
Cookie Theft	0.37 – 1.73*	100.0%	0.25 – 1.74*	99.8%	1.16 – 3.21*	100.0%	2.40 – 6.75*	100.0%
Picnic Scene	-0.06 – 0.94	95.4%	-0.10 – 1.23	94.1%	0.54 – 1.85*	100.0%	1.44 – 3.56*	100.0%
Avg. Speech Pred.	-0.28 – 0.59	73.6%	-0.35 – 0.72	61.7%	0.19 – 1.33*	99.8%	1.19 – 3.36*	100.0%

was generated per spontaneous speech task (Cookie Theft, Picnic Scene, journaling), and an average prediction per participant and assessment was computed by combining the three speech task-specific predictions. Figure 1 summarizes the cross-sectional regression performance (R^2) of these predictions, the test-retest reliability (ICC) across the two assessments and the MDC estimates for the speech predictions and for the composite cognitive scores, each discussed in turn below.

Cross-sectional regression performance.

Regression performance varied across composite scores (Figure 1 bottom row; full R^2 and MSE values reported in Supplementary Table S1). Predictions of the *language* composite reached the highest R^2 on every task, with the average prediction achieving $R^2 = 0.25$ [0.16, 0.32]. *Executive function* predictions followed with $R^2 = 0.18$ [0.10, 0.24] for the average prediction, and *memory* with $R^2 = 0.14$ [0.05, 0.21]. *Speed* predictions performed worst, with the average prediction reaching $R^2 = 0.09$ [0.03, 0.15]. Across all four composites, journaling was the weakest task-specific predictor, while the average prediction across the three tasks consistently matched or outperformed the best individual task.

Test-retest reliability.

The test-retest reliability of the four composite cognitive scores between baseline and follow-up ranged from ICC = 0.68 [0.61, 0.74] for *memory* to ICC = 0.76 [0.71, 0.81] for *speed* (Figure 1 middle row; Supplementary Table S2). For all four composites, the

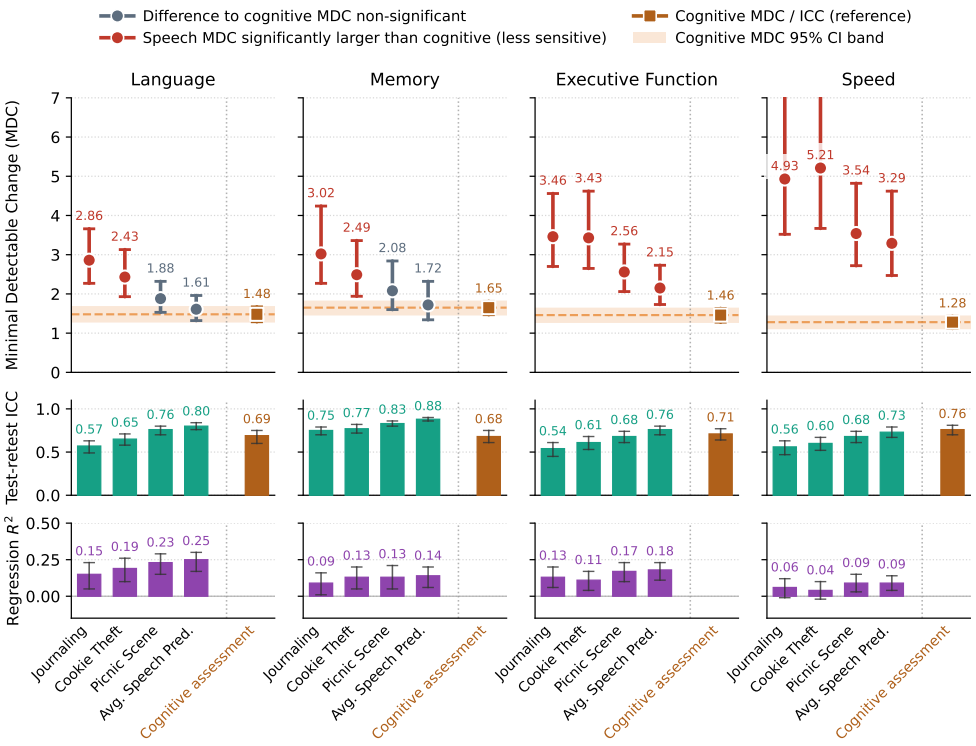


Fig. 1: Speech-based predictions and cognitive composite scores in Cohort A ($n = 294$). For each cognitive domain, three properties are reported, with error bars denoting 95% bootstrap CIs. **Top row (MDC):** Minimal Detectable Change for the three speech tasks (Journaling, Cookie Theft, Picnic Scene), the average prediction across speech tasks, and the reference MDC of the composite cognitive score. **Middle row (ICC):** test-retest reliability of speech-based predictions and the composite cognitive score between baseline and follow-up. **Bottom row (R^2):** cross-sectional regression performance averaged across the two assessments.

average speech prediction across the three tasks was more reliable than any individual task prediction. The average prediction was also more reliable than the cognitive composite itself for three of the four composites (*language*: 0.80 [0.76, 0.84] vs. 0.69; *memory*: 0.89 [0.86, 0.91] vs. 0.68; *executive function*: 0.77 [0.71, 0.81] vs. 0.71), and approximately matched it for *speed* (0.74 [0.68, 0.78] vs. 0.76). This indicates that, on three of the four composites, the average speech-based prediction was more stable across assessments than the cognitive composite it was trained on. Notably, predictions of the *memory* composite reached the highest reliability across all tasks despite only moderate cross-sectional R^2, illustrating that prediction reliability and regression performance are distinct properties, a dissociation we examine systematically in Section 2.6.

Minimal detectable change.

The Minimal Detectable Change (MDC) defines the smallest within-person change that can be distinguished from measurement noise at a given confidence level (here, 95%); lower values indicate higher sensitivity. We estimated the MDC of each cognitive composite score from its test-retest reliability and variance, and derived the MDC of each speech-based prediction by extending the same framework to treat predictions as attenuated, speaker-biased measurements of latent cognition (details in Section 4.5). Both MDCs are expressed on the latent-cognition scale and are therefore directly comparable. Results are summarized in Figure 1 (top row), with bootstrap tests reported in Table 3. The cognitive MDC ranged from 1.28 for *speed* to 1.65 for *memory*, defining the cognitive-assessment reference for within-person change detection in each composite. Across all four cognitive domains, the average speech prediction across the three tasks outperformed any individual task prediction, and the more constrained picture description tasks (Cookie Theft, Picnic Scene) outperformed the less constrained journaling task.

For the *language* and *memory* composites, the average speech prediction across the three speech tasks (*language*: MDC = 1.61 [1.32, 1.96]; *memory*: MDC = 1.72 [1.34, 2.32]) was only slightly larger than the cognitive reference, and these differences were not statistically significant (bootstrap 95% CIs of the difference: [−0.28, 0.59] and [−0.35, 0.72] respectively). Individual task predictions yielded larger MDCs and were therefore less sensitive than the average prediction. For Picnic Scene, the difference to the cognitive reference was not statistically significant on either composite (*language*: MDC = 1.88 [1.53, 2.32], CI of the difference: [−0.06, 0.94]; *memory*: MDC = 2.08 [1.60, 2.84], CI of the difference: [−0.10, 1.2]), whereas Cookie Theft and journaling yielded significantly larger MDCs than the cognitive reference on both composites.

For the *executive function* and *speed* composites, all speech-based MDCs were significantly larger than the cognitive reference (Table 3), with the average prediction reaching MDC = 2.15 [1.73, 2.73] for *executive function* and 3.29 [2.47, 4.62] for *speed*. Speech-based predictions are therefore less sensitive than cognitive assessment to within-person change on these two composites.

Overall, the speech-based average prediction matches the cognitive assessment on *language* and *memory*, but not on *executive function* and *speed*.

2.4 Validation of the speech measurement model

The MDC estimation in the previous section relies on a measurement model in which the speech-based prediction is treated as an imperfect measurement of latent cognition (details in Section 4.5). The model decomposes the prediction into a stable speaker-specific bias β that captures non-cognitive characteristics of speech (such as habitual speaking style, accent, or vocabulary preferences), an attenuation factor λ that captures how strongly variation in latent cognition is expressed in the prediction, and a residual prediction noise term.

We tested this measurement model using structural equation modeling (SEM), separately for each speech task and each composite cognitive score in Cohort A. Absolute fit was adequate to excellent across all twelve task-composite combinations, with CFI and TLI mostly above 0.95 and SRMR at most 0.06 (Table 4). The lowest TLI value

Table 4: Absolute fit indices of the speech measurement model for each speech task and composite score for Cohort A. CFI and TLI above 0.95 and SRMR below 0.08 indicate adequate fit. The likelihood-ratio tests comparing fit to the five alternative models are reported in the supplementary material (Table S3).

	Language			Memory			Executive Function			Speed		
	TLI	CFI	SRMR	TLI	CFI	SRMR	TLI	CFI	SRMR	TLI	CFI	SRMR
Cookie Theft	1.00	1.00	0.05	0.97	0.98	0.06	1.01	1.01	0.02	1.01	1.00	0.02
Picnic Scene	0.96	0.97	0.02	0.97	0.98	0.06	1.00	1.00	0.02	1.00	1.00	0.03
Journaling	0.92	0.95	0.05	0.98	0.98	0.06	0.96	0.97	0.04	0.97	0.98	0.02

(0.92, Journaling on *language*) remained close to the conventional threshold, and the corresponding CFI (0.95) and SRMR (0.05) met it. The baseline measurement model therefore captured the covariance structure of the speech-based predictions well across composites and tasks.

We further compared the baseline model against five alternatives (Supplementary Table S3). Across all twelve task-composite combinations, the three nested restrictions of the model were strongly rejected in its favor: constraining $\lambda = 1$, dropping β, and dropping both jointly (all $\Delta\chi^2 \geq 64.5$, all $p < 0.001$). This confirmed that both components are necessary across all four composites, i.e., that the regression model attenuates predictions ($\lambda < 1$) and that the speaker-specific bias β captures substantial variance. Two further alternatives relaxed the stationarity assumption underlying the MDC derivation by allowing residual variances of cognitive scores or speech predictions to differ between assessments. Across the 24 stationarity tests (12 task-composite combinations $\times$ 2 relaxations), only one yielded a marginally significant improvement (Cookie Theft on *language*, $\Delta\chi^2 = 4.57$, $p = 0.033$). This is within the range expected under multiple testing, and we therefore retained the stationary baseline model for the MDC analyses.

2.5 Replication in Cohort B

To assess the robustness of the Cohort A findings, we replicated the main analyses in Cohort B, an independent sample of 70 older adults who completed the same protocol twice over a substantially shorter retest interval of approximately two months (mean 2.25 months, SD 0.19). The short interval makes meaningful change in true cognition unlikely, providing a setting in which the stationarity assumption underlying the MDC framework is more strongly justified than in Cohort A.

Returners and non-returners in Cohort B were comparable on all demographic variables and on all four composite cognitive scores at baseline (all $p \geq 0.11$; Supplementary Table S6). Attrition over the short retest interval was therefore not systematically related to cognition or demographics, supporting the use of Cohort B as an independent replication sample.

Composite cognitive scores showed small mean improvements between baseline and follow-up in three of the four domains (*language*: +0.17, $p = 0.05$; *speed*: +0.17, $p = 0.01$; *executive function*: +0.20, $p < 0.01$; *memory*: −0.10, $p = 0.18$; Supplementary

Table S7). Given the short retest interval, these improvements are most plausibly attributable to practice effects rather than true cognitive change. Note that presence of practice effects is restricted to the measured cognitive scores (Y). It has no effect on the latent true cognition C and on the speech-based predictions $\hat{Y}$, which are modeled as attenuated measurements of C.

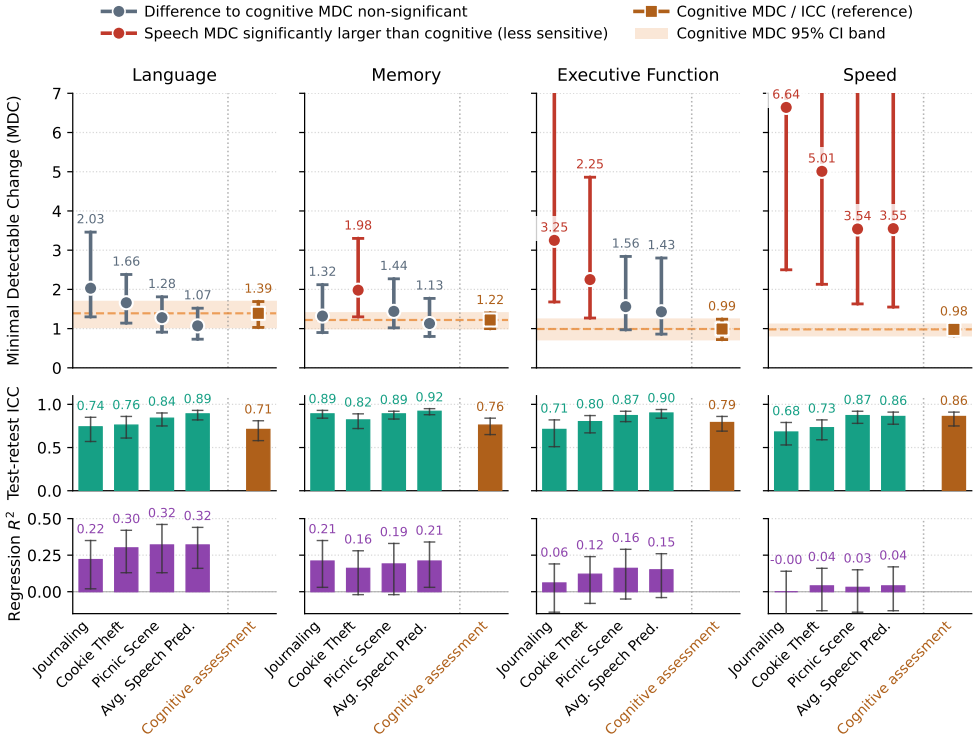


Fig. 2: Replication results in Cohort B ($n = 70$): Speech-based predictions and cognitive assessment reference of the four composite cognitive scores. For each composite, three properties are reported, with error bars showing 95% bootstrap CIs. **Top row (MDC):** Minimal Detectable Change for the three speech tasks (Journaling, Cookie Theft, Picnic Scene), the average prediction across tasks, and the MDC of the composite cognitive score. **Middle row (ICC):** test-retest reliability of speech-based predictions and the composite cognitive score between baseline and follow-up. **Bottom row (R^2):** cross-sectional regression performance averaged across the two assessments.

Out-of-sample regression performance was comparable to Cohort A, with the average prediction reaching $R^2 = 0.32$ [0.14, 0.45] for *language*, 0.21 [−0.01, 0.36] for *memory*, 0.15 [−0.09, 0.30] for *executive function*, and 0.04 [−0.15, 0.19] for *speed* (Figure 2, Supplementary Table S8). Test-retest reliability was higher in Cohort B than in Cohort A across all measures (Figure 2, Supplementary Table S9), consistent

with the shorter retest interval: cognitive composite reliabilities ranged from ICC = 0.72 [0.58, 0.82] for *language* to ICC = 0.87 [0.79, 0.92] for *speed*, and the average speech prediction reached ICC $\geq$ 0.86 on every composite.

The speech measurement model fit well across all twelve task-composite combinations (CFI/TLI $\geq$ 0.94, SRMR $\leq$ 0.09; Supplementary Table S10). As in Cohort A, the three nested restrictions ($\lambda = 1$, dropping β, and both jointly) were strongly rejected in favor of the full model on all task-composite combinations (all $\Delta\chi^2 \geq 14.4$, all $p < 0.001$), while only one of the 24 stationarity relaxations yielded a significant improvement (Picnic Scene on *memory*, $\Delta\chi^2 = 4.71$, $p = 0.03$; Supplementary Table S11). This is within the range expected under multiple testing, supporting the measurement-model assumptions underlying the speech-based MDC in the independent replication cohort.

The MDC results were consistent with the findings from Cohort A (Figure 2, Supplementary Table S12). Point estimates of the difference between speech-based and cognitive assessment MDC tended to be smaller in Cohort B, but confidence intervals were wider due to the smaller sample size, yielding the same qualitative picture. For *language* and *memory*, the average speech prediction reached MDCs comparable to the cognitive reference (bootstrap 95% CIs of the difference: [−0.74, 0.24] and [−0.46, 0.58] respectively, both including zero). The average prediction was smaller than the cognitive MDC on 87.1% and 62.5% of bootstrap samples respectively, although neither difference reached significance. For *executive function*, the average prediction was not significantly worse than the cognitive reference in Cohort B (CI of the difference: [−0.29, 2.05]), an improvement over Cohort A that should be interpreted in light of the wider CI. For *speed*, all speech-based MDCs remained significantly larger than the cognitive reference (all CIs excluding zero), confirming the Cohort A finding that speech-based predictions are not competitive with cognitive assessment on this composite. These replication results support the conclusion that the average speech-based prediction reaches within-person sensitivity comparable to cognitive assessment for the *language* and *memory* composites.

2.6 Simulation results: What determines the speech-based MDC?

To determine which model properties drive the speech-based MDC, we simulated longitudinal cognitive scores and speech features across a grid of two parameters: the proportion of stable non-cognitive variance in speech features (ρ) and the variance of feature measurement noise (σ^2_η). The variance and measurement noise of the cognitive scores were fixed to match the empirical Cohort A *language* data ($\mathrm{Var}(Y) = 0.91$, test-retest reliability $r_Y = 0.7$), so that the simulation isolated the speech-side determinants of the MDC.

The simulation, summarized in Figure 3, revealed that a speech-based $\mathrm{MDC}_{\mathrm{latent}}$ outperformed the cognitive MDC (blue color) when either the prediction reliability $\hat{R}_{\mathrm{pred}}$ was very high despite modest R^2, or regression performance was strong despite only moderate reliability. For instance, at $R^2 = 0.3$, the speech-based biomarker had to achieve $\hat{R}_{\mathrm{pred}} \geq 0.82$ to outperform the cognitive assessment. Not all combinations of R^2 and $\hat{R}_{\mathrm{pred}}$ were attainable. In particular, high regression performance with low

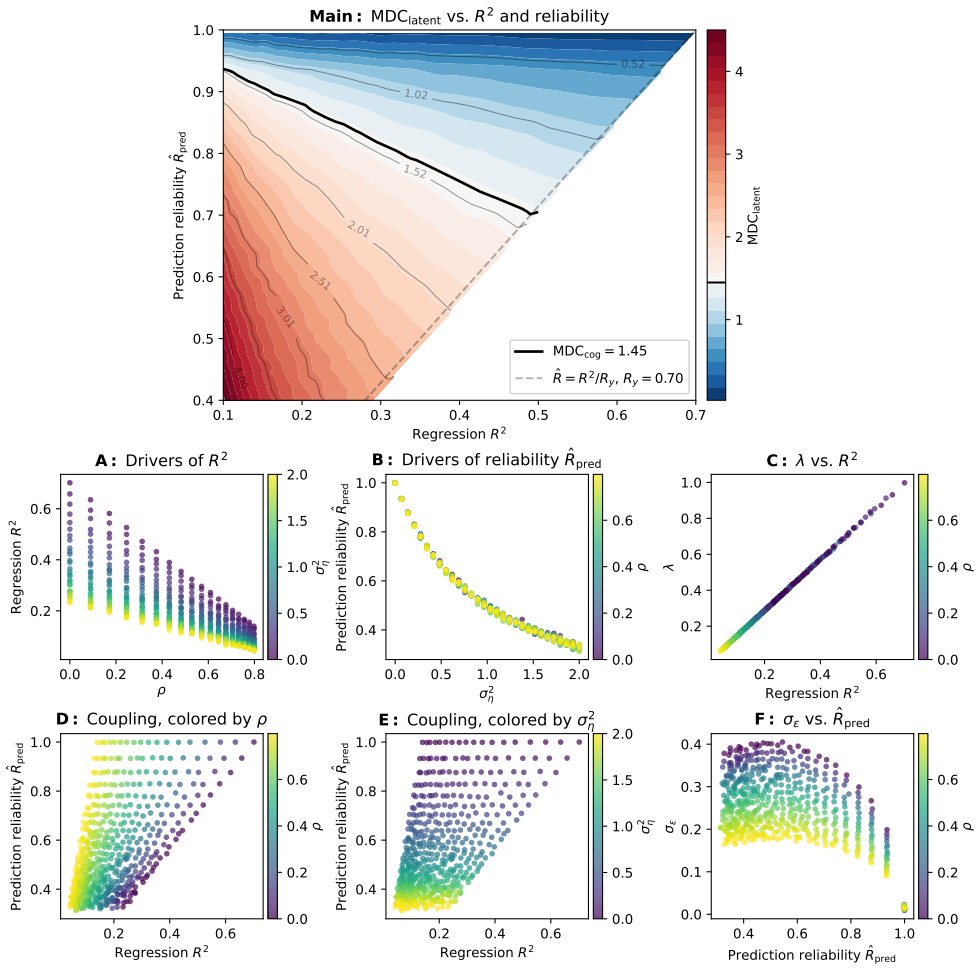


Fig. 3: Simulation results across a grid of proportions of non-cognitive variance in the simulated speech features (ρ) and feature noise values (σ^2_η). **Main panel:** $\text{MDC}_{\text{latent}}$ as a function of the two observable quantities of a machine learning regression model: regression performance (R^2) and speech prediction test-retest reliability $\hat{R}_{\text{pred}}$. The solid black contour marks the fixed cognitive $\text{MDC}_{\text{cog}} = 1.45$, and blue colors indicate smaller (=better) speech MDC, while red represents larger (=worse) speech MDC compared to the cognitive MDC. The dashed gray line is the identifiability bound $\hat{R}_{\text{pred}} = R^2/r_Y$ with $r_Y = 0.7$. Visualization is limited to $R^2 \geq 0.1$ and $\hat{R}_{\text{pred}} \geq 0.4$ to focus on areas with a relevant speech MDC. **Bottom panels:** decomposition of R^2 and reliability into their generative drivers (ρ, σ^2_η) [A, B], the coupling between the two observables colored by ρ and σ^2_η [D, E], and the mapping from observables to the two ingredients of the speech MDC in Equation 11 [C, F]. Visualizations are limited to $\sigma^2_\eta \leq 2$ and $\rho \leq 0.8$, representing the most informative range of values.

prediction reliability was theoretically incompatible with the model: if predictions closely captured cognition, and cognition was stable across assessments, the predictions were also reliable across assessments. More precisely, the achievable combination was bounded from below by $\hat{R}_{\text{pred}} = R^2/r_Y$, with r_Y representing the test-retest reliability of Y. Similarly, R^2 was capped at r_Y, because no regression model can predict Y more accurately than Y predicts itself on retest.

The simulations clarified that regression performance and prediction reliability were driven by partly distinct properties of the speech features, defined by the underlying simulation parameters. Regression performance depended on both sources of feature quality: it decreased when speech features contained less cognition-related signal, reflected by higher ρ, and when feature measurements were noisier across assessments, reflected by higher σ_η^2 (Panel A). Prediction reliability showed a different pattern. It was determined primarily by feature measurement noise σ_η^2, but was largely unaffected by the proportion of non-cognitive feature variance ρ (Panel B). This was because non-cognitive speech characteristics were stable within individuals and therefore contributed similarly to predictions at both assessments. They may have reduced cross-sectional cognitive prediction, but they did not necessarily reduce within-person stability. In contrast, feature measurement noise varied across assessments and therefore directly reduced prediction reliability.

The simulations further demonstrated that a biomarker with high reliability but low R^2 pointed to large non-cognitive variance in the features (ρ high), which may in principle be mitigated by better features; a biomarker with low reliability pointed to high feature measurement noise (σ_η^2 high), which might indicate unstable features or short recordings (Panels D and E).

Lastly, the simulations showed that the attenuation factor λ approximated a linear function of R^2 and was insensitive to ρ (Panel C), consistent with the interpretation that λ measures how faithfully a unit of cognitive change propagates into prediction-space change. Conversely, the residual prediction noise σ_ϵ decreased with prediction reliability (Panel F).

3 Discussion

This study evaluated speech-based cognitive prediction not by how closely it reproduced cognitive test scores, but by how sensitively it detected within-person cognitive change. Across two independent longitudinal cohorts, averaged predictions from three spontaneous speech tasks achieved Minimal Detectable Change values comparable to standardized cognitive assessment for *language* and *memory*, despite only moderate cross-sectional regression performance. In contrast, speech-based predictions for *executive function* and *speed* were less sensitive to within-person change. The averaged speech-based predictions also matched or exceeded the test-retest reliability of the cognitive composite scores across cognitive domains and cohorts. These findings show that spontaneous speech can provide monitoring sensitivity comparable to brief cognitive assessment in cognitive domains relevant to early neurodegenerative change, while also identifying where speech-based models remain insufficient.

Prior work predicting cognition from spontaneous speech in non-clinical older populations has reported moderate R^2 values [19–21], which could be interpreted as evidence of limited utility. Our results suggest a different interpretation. Cross-sectional speech-cognition associations are constrained by non-cognitive speaker characteristics, such as habitual speaking style, accent, dialect, education, bilingualism, personality, and cultural background [30–34]. These factors can reduce cross-sectional prediction accuracy because they introduce between-person variance that is not cognitive. However, many of these characteristics are stable within individuals and therefore cancel out, at least partly, when the goal is to detect within-person change. Our measurement model formalized this through the speaker-specific bias β. The high test-retest reliability of speech-based predictions further confirmed that substantial variance in the predictions was stable across assessments rather than random within-person noise.

A second reason why moderate R^2 values may be misleading is that cognitive test scores are themselves noisy targets. Regression performance metrics such as R^2 quantify agreement with observed scores, not with true cognition, the latent construct of interest. If part of the target score reflects random measurement error, a model should not be expected to reproduce that error. Supervised regression is optimized to capture systematic variation in the target labels, while random label noise is attenuated. As a result, speech-based predictions can show a stronger association with true cognition than the noisy labels they were trained on. This does not mean that speech predictions necessarily provide a purer measure of cognition: they may also contain random noise or stable non-cognitive speaker bias. It does mean, however, that cross-sectional accuracy and longitudinal monitoring sensitivity answer different questions and should not be treated as interchangeable.

The metric most directly aligned with longitudinal monitoring is therefore the MDC, rather than R^2. A digital biomarker does not need to perfectly estimate true cognition to be clinically useful for monitoring; it needs to detect when an individual has changed beyond expected measurement noise. The MDC formalizes exactly this requirement. In this sense, the MDC provides a clinically interpretable benchmark for evaluating whether a speech-based biomarker is sensitive enough to support repeated monitoring.

This benchmark also highlights an often overlooked limitation of the clinical reference standard. Brief cognitive assessments are indispensable in clinical practice, but their repeated use is constrained by measurement noise, practice effects, limited alternate forms, and transient influences such as fatigue, motivation, attention, mood, and testing conditions [11]. These constraints limit test-retest reliability. In healthy older adults, the most widely used clinical screening tools exhibit reliabilities ranging from 0.64 to 0.79 (MMSE) [35] and from 0.64 to 0.82 (MoCA) [12, 13]. The higher values often cited for these instruments come from heterogeneous clinical samples and reflect greater between-person variance rather than lower error [36]. Our cognitive composite scores, with test-retest reliabilities of ICC $\approx$ 0.7 – 0.8, were comparable to these standard instruments, indicating that our benchmark is realistic. These reliabilities also show that even established instruments track within-person change imperfectly, so the appropriate bar for a digital biomarker is not the perfect estimation of cognition but the within-person sensitivity of the tools currently used to monitor it.

Speech-based monitoring may sidestep one of these constraints directly. Practice effects improve scores on most cognitive tests through item recall, task familiarity, and deliberate strategy, and can thereby mask true decline [14]. In Cohort B, the cognitive scores showed improvements consistent with practice effects, whereas the speech-based predictions did not (Supplementary Table S5), even though the same pictures were presented at both assessments. This is plausibly because speech reflects largely implicit features of production that are not readily improved through recall, familiarity, or deliberate strategy.

By the standard of within-person sensitivity, speech-based predictions were competitive for *language* and *memory*, which are arguably the two most important cognitive domains for the early detection of AD. Prior work using longitudinal modeling has indicated that memory diverges from healthy aging up to eight years before MCI [37], and correlates with amyloid deposition [38]. Semantic fluency, the test contributing most strongly to our *language* composite score, has been identified as an equally early marker of AD-related neurodegenerative processes and predicts clinical progression in amyloid-positive individuals [39]. Our results therefore suggest that speech-based monitoring may be most useful for domains closely aligned with early cognitive changes in AD.

The simulation analysis extends the empirical findings by identifying separable determinants of speech-based monitoring sensitivity: regression performance (R^2) and prediction reliability jointly determine the achievable MDC, and either axis can be improved largely independently. A biomarker with low R^2 suggests that the speech features or model architecture fail to capture sufficient cognition-related variance, which may be addressed through richer feature representations or more expressive algorithms, including deep learning, transformer-based architectures, and advanced speech embeddings [40, 41]. A biomarker with low reliability points to unstable features or short recordings, which can be addressed through longer samples, more robust feature extraction, or averaging predictions across tasks [42, 43]. The simulation therefore provides a practical development map for improving speech-based digital biomarkers.

Our empirical results support this interpretation. Averaging predictions across the three speech tasks substantially improved prediction reliability in our data. This was expected, because the cognitive signal is shared across tasks while measurement noise is largely task-specific, so averaging across tasks canceled much of the noise. The more constrained picture description tasks (Picnic Scene, Cookie Theft) outperformed the less constrained journaling task on every metric, likely because the journaling prompt introduced greater variability in topic, content, emotional salience, and recent life events. In practical terms, several short, standardized speech tasks may provide a simple and low-burden way to improve monitoring sensitivity. This is particularly relevant for remote or app-based deployment, where repeated sampling is feasible and inexpensive.

Despite its central role, test-retest reliability is rarely reported in speech-based cognition studies. While some studies have found varying but often poor reliabilities of individual acoustic or linguistic features [43–45], we are aware of only a single study reporting the reliability of a speech-based composite score or machine learning prediction [46]. Our results indicate that this should become standard practice. For

digital biomarkers intended for monitoring, we recommend that future work report at least the test-retest reliability of speech-based predictions alongside R^2. In the absence of longitudinal data, the within-session ICC between predictions from comparable speech tasks can serve as a useful proxy. For example, the within-session ICC between Cookie Theft and Picnic Scene predictions for the *language* domain (ICC = 0.74 in Cohort A, 0.83 in Cohort B; Supplementary Table S4) aligned with the across-session test-retest ICC in our data.

The validity of the MDC comparison rests on the assumption that speech predictions and cognitive scores are noisy measurements of the same latent construct, and that the speech predictions include both an attenuated cognitive signal λ and stable speaker-specific bias β. Our SEM analysis showed our data to be consistent with all three assumptions: the baseline measurement model fit well across tasks and cohorts, nested alternatives dropping attenuation ($\lambda = 1$) or speaker bias (β) were rejected, and relaxing stationarity did not improve fit.

Our findings do not imply that speech assessment can diagnose cognitive impairment or replace comprehensive neuropsychological evaluation. Rather, spontaneous speech may serve as a scalable, low-cost, and repeatable monitoring instrument, where its sensitivity to within-person change is competitive with brief cognitive screening. In a population-level setting, individuals could provide short speech samples remotely and repeatedly; a within-person change exceeding the speech-based MDC would flag the need for more specific cognitive, neurological, or clinical follow-up. This use case is well aligned with digital medicine: the goal is not to replace specialist assessment, but to increase the frequency, scalability, and sensitivity of monitoring in non-clinical populations or between clinical visits.

Our results should be interpreted in light of several limitations. First, while we estimated sensitivity to detect cognitive decline, our data does not contain longitudinal data of participants with true cognitive decline. Validation in longitudinal clinical cohorts with documented MCI or AD progression is therefore the critical next step. Second, the sample was relatively homogeneous: participants were predominantly native English-speaking, well educated, independently living older adults with relatively high baseline cognition. Generalizability to more diverse linguistic, cultural, educational, and clinical populations remains to be established. Third, our measurement model rests on stationarity and independence assumptions that are difficult to test. Our SEM analysis revealed no obvious misspecification, but larger sample sizes are needed to strengthen or refine our measurement model. Fourth, we interpreted a within-person change in speech predictions exceeding the MDC as a change in true cognition C. However, it could also reflect changes in the non-cognitive characteristics summarized in β. Follow-up clinical assessment would be required to determine which.

Several directions for future work emerge from this study. First, the field needs large longitudinal datasets that combine speech samples with cognitive scores in cohorts spanning the full range from healthy aging to dementia. Such datasets would allow direct validation of whether the speech-based MDC framework reflects sensitivity to clinically meaningful decline. Second, while our speech-based predictions match brief cognitive assessment in within-person sensitivity, both remain insensitive to small cognitive changes. Lowering the MDC further is therefore necessary, and future work could

explore combining speech analysis with quick, remote, and easily repeatable cognitive probes, to capture more cognitive variance per session while preserving the scalability of speech-based monitoring. Third, dense longitudinal sampling should move beyond pairwise change thresholds towards trajectory modeling, which can use repeated observations to detect slower and smaller deviations over time. Finally, validation across other languages, linguistic backgrounds, and demographic groups will be essential before such tools can be deployed in clinical practice.

In conclusion, the key question for speech-based cognitive monitoring is not whether predictions perfectly reproduce cognitive test scores, but how reliably they detect within-person change. By that standard, speech-based predictions are already a viable monitoring tool in healthy older adults; the next step is to test whether they remain so in clinical decline.

4 Methods

4.1 Study design and participants

Data collection for this study extended a previous cross-sectional dataset described in detail in Heitz et al. [21]. 1003 older English-speaking adults from the USA and UK were recruited via the online crowdsourcing platform Prolific between April and July 2024. Inclusion criteria required participants to be aged 60 or older, native English speakers, residents of the USA or UK, and to have access to a computer with internet and a microphone. The protocol consisted of four parts. First, a demographic survey collected age, gender, country, education (binarized along ISCED levels 0–4 vs. 5–8 [47]), and subjective socioeconomic status on the 10-point MacArthur Scale [48]. Second, three spontaneous speech tasks were recorded via the participant's microphone: two picture description tasks (Cookie Theft from the Boston Diagnostic Aphasia Examination [49], Picnic Scene from the Western Aphasia Battery [50]) and a journaling task in which participants described their past week, each with a one-minute minimum and no upper time limit. Third, participants completed the Amsterdam Cognition Scan (ACS), a validated online neuropsychological battery covering memory, executive function, and processing speed [51]. Fourth, participants completed three standardized language tasks: semantic fluency (animals, one minute), phonemic fluency (words starting with "F", one minute), and a shortened Boston Naming Test [52] with 20 items.

Approximately 21 months later, between January and March 2026, all original participants were re-invited via Prolific to complete the same protocol a second time. $n = 305$ returned and formed the primary longitudinal cohort of this study, which we refer to as *Cohort A* (30.4% of the original sample, mean retest interval 20.9 months, SD 0.95). Of these, $n = 294$ completed all spontaneous speech and cognitive tasks at both assessments, and longitudinal analyses were restricted to this complete-case subset. Within Cohort A, we refer to the original 2024 assessment as the *baseline assessment* and the 2026 reassessment as the *follow-up assessment*. An independent longitudinal cohort, *Cohort B*, with a shorter retest interval, is introduced in Section 4.8 and served for replication of the main analyses.

The follow-up assessment included all components of the baseline assessment. The only modification to the original protocol was the use of an alternate word list (constructed to match the original words across 13 criteria [53]) for the verbal learning test within the ACS, to minimize practice effects on verbal learning and recall.

The study was approved by the ethics committee of the Faculty of Arts and Social Sciences at the University of Zurich (Number: 23.08.03), and all participants provided informed consent for the use of their data.

To assess potential attrition biases, we compared baseline characteristics of participants who returned for the follow-up assessment (Cohort A) against those who did not return (*non-returners*), on demographic variables and the four composite cognitive scores (*language*, *executive function*, *memory*, and *speed*), using Student's t-tests [54] for continuous variables and Fisher exact tests [55] for categorical variables.

4.2 Data preprocessing

Data preprocessing followed the procedures described in our previous study [21]: Spontaneous speech recordings shorter than 45 seconds were manually checked and, if deemed insufficient, removed. Recorded audio was transcribed using Google Speech's *chirp* ASR model [56]. ACS scores were screened for outliers using demographic-adjusted linear regression: observations with z-scores > 4 were marked missing and imputed using a multivariate Round-Robin Bayesian ridge regression based on the remaining cognitive scores from the same participant [57]. Scores for the three standardized language tasks were generated by an automatic scoring algorithm operating on the ASR transcripts [21].

4.3 Composite cognitive scores

Four composite scores were derived from individual test scores to represent the cognitive domains *language*, *executive function*, *memory*, and *speed*. The composite score structure and factor loadings were established in our previous study on the baseline assessment data of the 1003 participants through confirmatory factor analysis (CFA) using R's `lavaan` package [21]. This CFA was informed by Mukherjee et al. [58], in which a panel of experts mapped cognitive tests to domains based on theoretical considerations. The model was then iteratively refined based on modification indices and theoretical plausibility. The final model demonstrated robust fit (robust CFI: 0.94, TLI: 0.92, RMSEA: 0.06) and is used here as the reference CFA model for the longitudinal analyses. The complete list of tests and their primary outcome measures is provided in Supplementary Table S13, and the factor loadings onto the four composite scores are given in Supplementary Table S14.

To evaluate whether this measurement model generalizes to the Cohort A follow-up assessments, we tested for longitudinal measurement invariance between the baseline and follow-up occasions. We fitted a longitudinal CFA in which each composite factor was specified at both occasions, with the residuals of matching indicators allowed to correlate across time, and compared a hierarchy of nested models: a configural model (equal structure, freely estimated parameters), a weak (metric) invariance model additionally constraining factor loadings to equality across occasions, and a strong

(scalar) invariance model additionally constraining indicator intercepts. We report robust fit indices for each model (robust CFI, TLI, RMSEA, and SRMR) and, for each step in the hierarchy, the scaled chi-square difference test and the changes in fit indices (ΔCFI, ΔRMSEA, ΔSRMR), evaluated against the cutoffs of Chen [59]. Composite scores for the follow-up observations were then computed with loadings fixed to baseline values, ensuring direct comparability across assessments. As in the original analysis, all composite scores were standardized relative to the original baseline distribution, such that a score of 0 corresponds to the average baseline participant and a one-unit change corresponds to one baseline standard deviation.

We tested for within-person change in each composite cognitive score between the baseline and follow-up assessments using paired t-tests, and report the proportion of Cohort A participants showing improved versus declined scores. Any observed change between baseline and follow-up reflects a combination of true cognitive change, measurement error, and potential practice effects from prior exposure to the protocol. To estimate practice effects, we compared follow-up scores in Cohort A against the baseline scores of Cohort B, who completed the protocol for the first time (see Section 4.8). For each composite score, we fit a linear model with a group indicator (Cohort A follow-up vs. Cohort B baseline) as the predictor of interest and demographic variables as covariates. A positive group coefficient is consistent with a practice effect in Cohort A.

4.4 Speech features and prediction model

Feature extraction: We extracted the same set of linguistic and acoustic features from the ASR transcripts and audio recordings as in our previous study (see Supplementary Table S15 for the full set of features and their definitions). 38 linguistic features covered lexical, syntactic, and semantic properties of the transcripts, including measures of lexical diversity, part-of-speech ratios, constituency parse statistics, readability, and word-level properties such as frequency, age of acquisition, familiarity, ambiguity, and concreteness of nouns. 26 acoustic features comprised hand-crafted measures of pause behavior and speaking rate, as well as a curated subset of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [60]. Extreme feature values were removed and imputed from the remaining features of the same observation, mirroring the procedure used for cognitive scores. Features were then standardized within cross validation, using training-fold statistics.

Prediction model: For each of the four composite cognitive scores, we trained a machine learning regression model to predict the score from speech features and demographic variables (age, gender, education, country, socioeconomic status). The three speech tasks (Cookie Theft, Picnic Scene, journaling) were treated as separate observations, so that each participant contributed up to three observations per assessment, and the target variable of the prediction was the participant's composite score from the corresponding assessment. Models were trained on all baseline observations from the original 2024 collection together with the Cohort A follow-up observations.

We tested three regression model families and tuned hyperparameters to identify an algorithm well-suited to our setting: support vector regression (SVR) [61], Ridge

regression [62], and XGBoost [63]. Because the speech-based MDC described below depends jointly on regression performance and the stability of predictions across assessments, we selected hyperparameters that minimized the estimated MDC rather than optimizing R^2 alone. Specifically, for each candidate configuration we computed R^2 and the prediction test-retest reliability $\hat{R}_{\text{pred}}$, and mapped this pair onto our simulation surface (Section 2.6) to obtain an estimated speech-based MDC. To avoid leakage into the MDC estimates, which are the primary outcome of this study, model and hyperparameter selection was performed on the subset of original 2024 participants who did not return for the follow-up assessment. This kept all longitudinal data unused during model selection. The best-performing algorithm was SVR, which we report in the main results section. Details on the hyperparameter tuning, the range of tested hyperparameters, and the best hyperparameters for all model families are listed in Supplementary Section D.

Regression performance evaluation: Models were evaluated using stratified 10-fold cross-validation. To prevent data leakage, all speech samples from a given participant (across all tasks and assessments) were assigned to the same fold. Folds were stratified by age, country, gender, education, socioeconomic status, and the participant's mean composite cognitive score. Performance metrics were computed on the concatenation of out-of-fold test predictions across all ten folds, yielding one prediction per (participant, assessment, task) sample. We report the coefficient of determination (R^2) and mean squared error (MSE), with 95% bootstrap confidence intervals computed over 10000 resamples. Metrics are reported for each task separately (Cookie Theft, Picnic Scene, journaling), and for the average prediction, defined as the mean of the three task-specific predictions within each participant and assessment. For conciseness, we report the mean regression performance of the baseline and follow-up assessments.

Test-retest reliability: We assessed the test-retest reliability between baseline and follow-up assessments: For each of the four cognitive composite scores, we computed the intraclass correlation coefficient ICC(3,1) [64] between the two assessments. For the speech-based measures, we analogously computed ICC(3,1) between the baseline and follow-up out-of-fold predictions, separately for each of the four composite scores. Prediction reliability was estimated for each task separately and for the average prediction across the three tasks.

4.5 Cognitive assessment and speech predictions as measurements of true cognition

We framed both the observed composite cognitive scores and the speech-based predictions as noisy measurements of the same latent, unobserved construct, an individual's true cognition C. This common framing allowed the MDC of the speech predictions to be derived in direct analogy to the classical MDC of a neuropsychological test, and to be expressed on the same scale across the two modalities.

4.5.1 Measurement model for cognitive assessment

Following classical test theory [65], we modeled an observed composite cognitive score Y as the sum of true cognition C and a random measurement error δ:

$$Y = C + \delta, \qquad \delta \sim \mathcal{N}(0, \sigma_\delta^2), \qquad \delta \perp C \tag{1}$$

For two assessments of the same individual at times t_1 and t_2, the errors δ_1 and δ_2 were assumed to be independent. The test-retest reliability of the assessment was defined as the proportion of observed-score variance attributable to true cognition:

$$r = \frac{\mathrm{Var}(C)}{\mathrm{Var}(Y)} \tag{2}$$

We estimated r using the intraclass correlation coefficient (ICC). The standard error of measurement (SEM) then followed, given the independence of δ and C:

$$\mathrm{SEM} = \sigma_\delta \;=\; \sqrt{(1 - r)\,\mathrm{Var}(Y)} \tag{3}$$

4.5.2 MDC of the cognitive assessment

For monitoring, cognition is measured repeatedly, e.g. at time points t_1 and t_2, and the quantity of interest is the change $\Delta Y = Y_2 - Y_1$. Under the null hypothesis of no true change ($C_1 = C_2$), the MDC is the threshold above which ΔY is unlikely to arise from measurement error alone, at a chosen confidence level with critical value z ($z = 1.96$ for 95% confidence):

$$\mathrm{MDC}_{cog} = z \cdot \sqrt{\mathrm{Var}(\Delta Y)} \tag{4}$$

Under equation 1 and assuming the independence of δ_1 and δ_2, $\mathrm{Var}(\Delta Y) = 2\sigma_\delta^2$, so that

$$\mathrm{MDC}_{cog} = z \cdot \sqrt{2} \cdot \mathrm{SEM} \;=\; z \cdot \sqrt{2} \cdot \mathrm{SD}(Y) \cdot \sqrt{1 - r} \tag{5}$$

which is the standard MDC formulation used in neuropsychological test development.

4.5.3 Speech predictions as attenuated measurements of true cognition

We modelled the speech-based prediction $\hat{Y}$ of our machine learning regression model as an attenuated, speaker-biased measurement of true cognition:

$$\hat{Y} = \lambda\, C + \beta + \epsilon \tag{6}$$

Here, λ is the attenuation factor. It quantifies how strongly variation in true cognition is expressed in the speech-based prediction. Values of $\lambda < 1$ indicate that predictions are compressed toward the population mean, as expected when a supervised regression model predicts a noisy target. Since the speech features determine cognition only incompletely, a trained regressor shrinks its predictions toward the conditional mean, a standard consequence of the bias-variance trade-off [22].

The speaker-specific intercept β captures stable non-cognitive characteristics of a person's speech that can shift the predicted score independently of cognition. These include, for example, habitual speaking rate, voice quality, accent, vocabulary preferences, and discourse style. Because these characteristics persist across sessions, β was treated as time-invariant and thus cancels in within-person change scores.

The residual term ϵ captures prediction noise that is not stable across assessments. For the derivations below, we assumed $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\epsilon \perp C$, $\epsilon \perp \delta$, $\beta \perp C$, and $\beta \perp \delta$. For two repeated assessments of the same person, the errors ϵ_1 and ϵ_2 were assumed to be independent.

We defined λ as the loading of the speech-based prediction onto true cognition:

$$\lambda = \frac{\mathrm{Cov}(\hat{Y}, C)}{\mathrm{Var}(C)} \tag{7}$$

Because C is unobserved, λ cannot be estimated directly. To express λ in terms of observable quantities, we substituted $C = Y - \delta$ and used the independence assumptions stated above, which together imply $\mathrm{Cov}(\hat{Y}, \delta) = 0$. Combined with equation 2, this yielded:

$$\lambda = \frac{\mathrm{Cov}(\hat{Y}, Y - \delta)}{\mathrm{Var}(C)} = \frac{\mathrm{Cov}(\hat{Y}, Y)}{r \cdot \mathrm{Var}(Y)} \tag{8}$$

4.5.4 MDC of the speech predictions

The MDC of the speech-based prediction $\hat{Y}$ is the smallest within-person change that is unlikely to arise from speech measurement noise alone under the null hypothesis of no true cognitive change ($C_1 = C_2$). From equation 6, together with the time-invariance of β and the independence of ϵ_1 and ϵ_2,

$$\mathrm{Var}(\Delta\hat{Y}) = \mathrm{Var}(\epsilon_2 - \epsilon_1) = 2\sigma_\epsilon^2, \tag{9}$$

so that, analogously to equation 4, the MDC in prediction space is

$$\mathrm{MDC}_{\mathrm{speech}}^{\mathrm{pred}} = z \cdot \sqrt{2\sigma_\epsilon^2} \tag{10}$$

The superscript ^{pred} indicates that this threshold is expressed in the *prediction space*, which is compressed by λ relative to both C and Y. To make it directly comparable to the MDC_{cog} (equation 5), we rescaled into units of latent cognition by dividing by λ:

$$\mathrm{MDC}_{\mathrm{speech}}^{\mathrm{latent}} = \frac{z}{\lambda} \cdot \sqrt{2\sigma_\epsilon^2} \tag{11}$$

A within-person change $\Delta\hat{Y}$ exceeding $\mathrm{MDC}_{\mathrm{speech}}^{\mathrm{pred}}$, or equivalently $\Delta\hat{Y}/\lambda$ exceeding $\mathrm{MDC}_{\mathrm{speech}}^{\mathrm{latent}}$, is therefore unlikely to reflect speech measurement noise alone and constitutes evidence of a change in true cognition.

4.6 Validation of the speech measurement model

We validated the speech measurement model on Cohort A. Because cognition did not change meaningfully between baseline and follow-up (Section 2.2), we treated true cognition as constant throughout this analysis.

To test the assumptions of the model jointly, we fitted a structural equation model separately for each speech task using four observed indicators per task ($Y_1, Y_2, \hat{Y}_1, \hat{Y}_2$). The first factor represents true cognition C, with unit loadings on Y_1, Y_2 and equal loadings λ on $\hat{Y}_1, \hat{Y}_2$, encoding a stable true cognition with a time-invariant attenuation factor. The second factor represents the stable speaker-specific bias β, with unit loadings on $\hat{Y}_1, \hat{Y}_2$ only. Residual variances of cognitive scores (σ^2_δ) and speech predictions (σ^2_ϵ) were constrained to be constant over time, and C and β were constrained to be uncorrelated ($\beta \perp C$). Fit was evaluated using CFI, TLI, and SRMR [66–68]; values of CFI and TLI above 0.95 and SRMR below 0.08 indicate adequate fit [68].

To assess robustness, we compared this full model against five alternative specifications. Three nested alternatives tested the model's defining features: *No-attenuation*, which fixed $\lambda = 1$, *No-bias* dropped β, and *No-attenuation/No-bias* imposed both restrictions. Two additional alternatives relaxed stationarity assumptions by allowing assessment-specific residual variances ($\sigma^2_{\epsilon,1} \neq \sigma^2_{\epsilon,2}$ and $\sigma^2_{\delta,1} \neq \sigma^2_{\delta,2}$), an assumption underlying the derivations of $\text{MDC}_{\text{speech}}$ and MDC_{cog} (Equations 5 and 10).

All models were estimated by maximum likelihood using Python's `semopy` framework [69] in the complete Cohort A ($n = 294$), separately for each speech task. We report a likelihood-ratio $\Delta\chi^2$ test for each nested comparison, and we considered the baseline adequate if the three nested alternatives were rejected in its favor, and neither relaxation of stationarity significantly improved model fit.

4.7 MDC estimation for cognitive assessment and speech predictions

We estimated all MDCs in Cohort A, treating true cognition as stable between baseline and follow-up. Under this assumption, between-assessment variation in the observed cognitive scores Y and in the speech-based prediction $\hat{Y}$ reflects measurement error rather than true cognitive change. These relevant variance components can therefore be estimated from between-assessment covariances and within-person differences.

Cognitive MDC: The test-retest reliability of cognitive scores was estimated as $\hat{r} = \text{Cor}(Y_1, Y_2)$ and $\widehat{\text{Var}}(Y)$ as the average of the two assessment-specific sample variances. The cognitive MDC was then computed from equation 5:

$$\widehat{\text{MDC}}_{\text{cog}} = z \cdot \sqrt{2\,\widehat{\text{Var}}(Y)\,(1 - \hat{r})}. \tag{12}$$

Speech MDC: The speech-based MDC requires estimating both the attenuation factor λ from equation 8, and the residual prediction noise σ^2_ϵ. λ was estimated using $\widehat{\text{Var}}(C) = \text{Cov}(Y_1, Y_2)$, which follows from $\delta_1 \perp \delta_2$ under constant C, and averaging

across assessments:

$$\hat{\lambda} = \frac{1}{2}\left(\frac{\mathrm{Cov}(\hat{Y}_1, Y_1)}{\mathrm{Cov}(Y_1, Y_2)} + \frac{\mathrm{Cov}(\hat{Y}_2, Y_2)}{\mathrm{Cov}(Y_1, Y_2)}\right). \tag{13}$$

Residual prediction noise σ_ϵ^2 was estimated from the within-person differences in the speech-based predictions: Under $C_1 = C_2$, and with the speaker-specific bias β cancelling across assessments, we computed $\hat{\sigma}_\epsilon^2 = \frac{1}{2}\widehat{\mathrm{Var}}(\hat{Y}_2 - \hat{Y}_1)$.

The prediction-space and latent-cognition speech MDCs were then computed as:

$$\widehat{\mathrm{MDC}}^{\mathrm{pred}}_{\mathrm{speech}} = z \cdot \sqrt{2\hat{\sigma}_\epsilon^2}, \qquad \widehat{\mathrm{MDC}}^{\mathrm{latent}}_{\mathrm{speech}} = \widehat{\mathrm{MDC}}^{\mathrm{pred}}_{\mathrm{speech}} / |\hat{\lambda}|. \tag{14}$$

All estimates were computed separately for each of the three speech tasks. In addition, we also estimated the MDCs for the average speech prediction across all three tasks. Averaging across tasks is expected to reduce task-specific prediction noise, provided that errors are not perfectly correlated across tasks, and should therefore lower the MDC relative to individual task predictions.

Confidence intervals and significance testing: We obtained 95% percentile confidence intervals from nonparametric bootstrap ($B = 10000$) resamples. Participants were resampled with replacement, retaining both assessments of each selected participant to preserve the paired longitudinal structure. To test whether each speech-based MDC was larger than the cognitive MDC, we computed the paired bootstrap difference $\widehat{\mathrm{MDC}}^{\mathrm{latent}}_{\mathrm{speech}} - \widehat{\mathrm{MDC}}_{\mathrm{cog}}$ on each bootstrap sample. We report the 95% CI of this difference together with the proportion of bootstrap samples in which the speech-based MDC exceeded the cognitive MDC. A CI entirely above zero indicates that the speech-based measure is less sensitive to within-person change than the cognitive assessment.

4.8 Replication in Cohort B

To assess the robustness of our main findings, we replicated the central analyses on an independent longitudinal cohort, *Cohort B*. $n = 103$ new participants were recruited via Prolific to undergo the Cohort A follow-up protocol for the first time, applying identical inclusion criteria as in the original collection. Because of temporary technical problems affecting the US version of the test battery, these new participants were recruited exclusively from the UK. Approximately two months later, these participants were re-invited via Prolific, and $n = 70$ returned (mean retest interval 2.25 months, SD 0.19). As for Cohort A, we refer to the first assessment as the *baseline assessment* and the second as the *follow-up assessment*. The short retest interval was chosen deliberately so that meaningful change in true cognition is unlikely in healthy older adults, providing a replication setting in which the stationarity assumption underlying the measurement-error framework is more strongly justified than in Cohort A.

We then repeated the analyses performed on Cohort A, including the demographic comparison to detect a potential attrition bias, the within-person change of composite

cognitive scores, the regression performance evaluation, the test-retest reliability estimation, the validation of the speech measurement model, and the MDC estimation. The speech-based prediction models trained on the original 2024 baseline observations and Cohort A follow-up observations (see above) were applied to Cohort B observations without retraining to generate out-of-sample speech-based predictions of the four composite cognitive scores.

4.9 Simulation

We used simulation to examine how measurement noise, reliability of predictions, machine learning regression performance, and non-cognitive influences on speech jointly determine the MDCs. We generated synthetic true cognitive scores and speech features, added measurement noise to both, then trained regression models to predict cognitive scores from speech features, and then estimated regression performance (R^2) as well as the MDCs. The simulation varied two generative parameters: variance of feature measurement noise (σ_η^2) and the proportion of feature variance that is unrelated to cognition $\rho \in [0, 1]$. Here, $\rho = 0$ means all feature variance is cognition-related, $\rho = 1$ means features contain no cognition-related signal. Cognitive measurement noise is fixed to $\sigma_\delta^2 = 0.273$, yielding a simulated test-retest reliability of cognitive assessment at 0.7, matching approximately the reliability observed in our empirical data.

Specifically, for subjects $i \in 1, \ldots, n$, we first generated latent cognition scores $c_i \sim \mathcal{N}(0, \sigma_c^2)$ and $k = 10$ true speech features $X_i \in \mathbb{R}^k$. The feature vectors were constructed as a mixture of a cognition-related component and a non-cognitive component:

$$X_i = \sqrt{1-\rho}\, X_i^{\text{cog}} + \sqrt{\rho}\, Z_i,$$

where X_i^{cog} was obtained by projecting c_i onto a fixed unit direction in feature space, and $Z_i \sim \mathcal{N}(\mathbf{0}, I_k)$ represents stable non-cognitive between-subject variation. Thus, ρ controlled how strongly cognition was reflected in the speech features.

We assumed true cognition was constant over time and drew two assessments ($t \in \{1, 2\}$) of observed cognitive scores y and feature vectors X^{obs} per subject by adding independent measurement noise:

$$y_{i,t} = c_i + \delta_{i,t}, \quad \delta_{i,t} \sim \mathcal{N}(0, \sigma_\delta^2), \qquad X_{i,t}^{\text{obs}} = X_i + \eta_{i,t}, \quad \eta_{i,t} \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_k).$$

$\text{Var}(c)$ was set such that $\text{Var}(y) = 0.91$ matching the variance of our empirical data in the *language* composite score of Cohort A.

We trained and evaluated a linear regression model predicting y_t from X_t^{obs} separately for each timepoint using 10-fold cross-validation, yielding an out-of-fold prediction $\hat{y}_{i,t}$ per subject and timepoint. Regression performance (R^2) was averaged across the two timepoints. Test-retest reliability of the cognitive assessment (y) and of the speech predictions ($\hat{y}$) were calculated using Pearson's r across the two assessments. The three MDCs were estimated as described in Section 4.7.

5 Data availability

The raw data of this study cannot be published fully due to the sensitive nature of voice recordings. However, the composite cognitive scores and speech-based predictions will be published on GitHub upon publication, allowing the replication of most results, including regression performance, test-retest reliabilities, and cognitive assessment and speech-based MDC estimates. In addition, the majority of participants have consented to the use of their data for further, non-commercial research. In the context of a scientific collaboration, the authors can make the audio recordings and transcripts from these participants available to researchers upon request.

6 Code availability

Our code for data preparation, feature extraction, machine learning regression, performance estimation, and results analysis will be made available online upon publication.

7 Acknowledgments

The authors gratefully acknowledge the Netherlands Cancer Institute Foundation / Antoni van Leeuwenhoek Hospital, together with the ACS Consortium Partners, for their support and for hosting the Amsterdam Cognition Scan test battery employed in the data collection for this study. This project was supported by the Swiss National Foundation (10000763) and Dementia Research Switzerland Synapsis Foundation (2021-PI02).

8 Author Contributions

NL and JH designed the study. JH collected and analyzed data, derived the theoretical measurement framework, implemented and evaluated the methodology, and drafted the manuscript. RD provided guidance and feedback on the statistical models. NL acquired funding, provided supervision and guidance, and oversaw the project. All the authors edited and reviewed the drafted manuscript.

9 Competing Interests

The authors declare no competing interests.

References

[1] Nichols, E., Steinmetz, J.D., Vollset, S.E., Fukutaki, K., Chalek, J., Abd-Allah, F., Abdoli, A., Abualhasan, A., Abu-Gharbieh, E., Akram, T.T., *et al.*: Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. The Lancet Public Health **7**(2), 105–125 (2022)

[2] Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., Cheng, F.: Alzheimer's disease drug development pipeline: 2024. Alzheimer's & Dementia: Translational Research & Clinical Interventions **10**(2), 12465 (2024)

[3] Van Dyck, C.H., Swanson, C.J., Aisen, P., Bateman, R.J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., *et al.*: Lecanemab in early alzheimer's disease. New England Journal of Medicine **388**(1), 9–21 (2023)

[4] Sims, J.R., Zimmer, J.A., Evans, C.D., Lu, M., Ardayfio, P., Sparks, J., Wessels, A.M., Shcherbinin, S., Wang, H., Monkul Nery, E.S., *et al.*: Donanemab in early symptomatic alzheimer disease: the trailblazer-alz 2 randomized clinical trial. Jama **330**(6), 512–527 (2023)

[5] Ngandu, T., Lehtisalo, J., Solomon, A., Levälahti, E., Ahtiluoto, S., Antikainen, R., Bäckman, L., Hänninen, T., Jula, A., Laatikainen, T., *et al.*: A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (finger): a randomised controlled trial. The Lancet **385**(9984), 2255–2263 (2015)

[6] Livingston, G., Huntley, J., Liu, K.Y., Costafreda, S.G., Selbæk, G., Alladi, S., Ames, D., Banerjee, S., Burns, A., Brayne, C., *et al.*: Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. The lancet **404**(10452), 572–628 (2024)

[7] Mendes, A.J., Ribaldi, F., Sayin, O., Khachvani, G., Mulargia, R., Volpara, G., Remoli, G., Nencha, U., Gianonni-Luza, S., Cappa, S., et al.: Single-domain and multidomain lifestyle interventions for the prevention of cognitive decline in older adults who are cognitively unimpaired: a systematic review and network meta-analysis. The Lancet Healthy Longevity (2025)

[8] Lang, L., Clifford, A., Wei, L., Zhang, D., Leung, D., Augustine, G., Danat, I.M., Zhou, W., Copeland, J.R., Anstey, K.J., *et al.*: Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. BMJ open **7**(2), 011146 (2017)

[9] Bradford, A., Kunik, M.E., Schulz, P., Williams, S.P., Singh, H.: Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. Alzheimer Disease & Associated Disorders **23**(4), 306–314 (2009)

[10] Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., *et al.*: Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The lancet neurology **12**(2), 207–216 (2013)

[11] Lezak, M.D., Howieson, D.B., Bigler, E.D., Tranel, D.: Neuropsychological Assessment. Oxford University Press, ??? (2012)

[12] Feeney, J., Savva, G.M., O'Regan, C., King-Kallimanis, B., Cronin, H., Kenny, R.A.: Measurement error, reliability, and minimum detectable change in the mini-mental state examination, montreal cognitive assessment, and color trails test among community living middle-aged and older adults. Journal of Alzheimer's Disease **53**(3), 1107–1114 (2016)

[13] Lindvall, E., Abzhandadze, T., Quinn, T.J., Sunnerhagen, K.S., Lundström, E.: Is the difference real, is the difference relevant: the minimal detectable and clinically important changes in the montreal cognitive assessment. Cerebral Circulation-Cognition and Behavior **6**, 100222 (2024)

[14] Calamia, M., Markon, K., Tranel, D.: Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. The Clinical Neuropsychologist **26**(4), 543–570 (2012)

[15] Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R., Markesbery, W.R.: Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study. Jama **275**(7), 528–532 (1996)

[16] Ahmed, S., Haigh, A.-M.F., Jager, C.A., Garrard, P.: Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. Brain **136**(12), 3727–3737 (2013)

[17] Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's disease **49**(2), 407–422 (2015)

[18] Petti, U., Baker, S., Korhonen, A.: A systematic literature review of automatic alzheimer's disease detection from speech and language. Journal of the American Medical Informatics Association **27**(11), 1784–1797 (2020)

[19] Wisler, A.A., Fletcher, A.R., McAuliffe, M.J.: Predicting montreal cognitive assessment scores from measures of speech and language. Journal of Speech, Language, and Hearing Research **63**(6), 1752–1761 (2020)

[20] Ostrand, R., Gunstad, J.: Using automatic assessment of speech production to predict current and future cognitive function in older adults. Journal of Geriatric Psychiatry and Neurology **34**(5), 357–369 (2021)

[21] Heitz, J., Engler, I.M., Langer, N.: Towards a speech-based digital biomarker for cognitive impairment: speech as a proxy for cognitive assessment. npj Digital Medicine (2026)

[22] Hastie, T., Tibshirani, R., Friedman, J., et al.: The elements of statistical learning. Springer (2009)

[23] Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., Brambati, S.M.: Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review. Aphasiology **34**(6), 723–755 (2020)

[24] Shakeri, A., Farmanbar, M.: Natural language processing in alzheimer's disease research: Systematic review of methods, data, and efficacy. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **17**(1), 70082 (2025)

[25] Mahon, E., Lachman, M.E.: Voice biomarkers in middle and later adulthood as predictors of cognitive changes. Frontiers in psychology **15**, 1422376 (2024)

[26] Thomas, J.A., Burkhardt, H.A., Chaudhry, S., Ngo, A.D., Sharma, S., Zhang, L., Au, R., Hosseini Ghomi, R.: Assessing the utility of language and voice biomarkers to predict cognitive impairment in the framingham heart study cognitive aging cohort data. Journal of Alzheimer's Disease **76**(3), 905–922 (2020)

[27] Stegmann, G., Hahn, S., Bhandari, S., Kawabata, K., Shefner, J., Duncan, C.J., Liss, J., Berisha, V., Mueller, K.: Automated semantic relevance as an indicator of cognitive decline: out-of-sample validation on a large-scale longitudinal dataset. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **14**(1), 12294 (2022)

[28] Berg, R.L., Boer, C., Zwan, M.D., Jutten, R.J., Liere, M., Glind, M.-C.A., Dubbelman, M.A., Schlüter, L.M., Harten, A.C., Teunissen, C.E., *et al.*: Digital remote assessment of speech acoustics in cognitively unimpaired adults: feasibility, reliability and associations with amyloid pathology. Alzheimer's Research & Therapy **16**(1), 176 (2024)

[29] Robin, J., Xu, M., Balagopalan, A., Novikova, J., Kahn, L., Oday, A., Hejrati, M., Hashemifar, S., Negahdar, M., Simpson, W., *et al.*: Automated detection of progressive speech changes in early alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **15**(2), 12445 (2023)

[30] Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research **30**, 457–500 (2007)

[31] Gollan, T.H., Montoya, R.I., Cera, C., Sandoval, T.C.: More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. Journal of memory and language **58**(3), 787–814 (2008)

[32] Jacewicz, E., Fox, R.A., O'Neill, C., Salmons, J.: Articulation rate across dialect, age, and gender. Language variation and change **21**(2), 233–256 (2009)

[33] Le Dorze, G., BÉDard, C.: Effects of age and education on the lexico-semantic content of connected speech in adults. Journal of communication disorders **31**(1), 53–71 (1998)

[34] Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology **54**(1), 547–577 (2003)

[35] Tombaugh, T.N., McIntyre, N.J.: The mini-mental state examination: a comprehensive review. Journal of the American Geriatrics Society **40**(9), 922–935 (1992)

[36] De Vet, H.C., Terwee, C.B., Mokkink, L.B., Knol, D.L.: Measurement in Medicine: a Practical Guide. Cambridge university press, ??? (2011)

[37] Mistridis, P., Krumm, S., Monsch, A.U., Berres, M., Taylor, K.I.: The 12 years preceding mild cognitive impairment due to alzheimer's disease: the temporal emergence of cognitive decline. Journal of Alzheimer's Disease **48**(4), 1095–1107 (2015)

[38] Mortamais, M., Ash, J.A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Pose, C., Albala, B., Ropacki, M., Ritchie, C.W., *et al.*: Detecting cognitive changes in preclinical alzheimer's disease: A review of its feasibility. Alzheimer's & dementia **13**(4), 468–492 (2017)

[39] Vonk, J.M., Bouteloup, V., Mangin, J.-F., Dubois, B., Blanc, F., Gabelle, A., Ceccaldi, M., Annweiler, C., Krolak-Salmon, P., Belin, C., *et al.*: Semantic loss marks early alzheimer's disease-related neurodegeneration in older adults without dementia. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **12**(1), 12066 (2020)

[40] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[41] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[42] Almaghrabi, S.A., Thewlis, D., Thwaites, S., Rogasch, N.C., Lau, S., Clark, S.R., Baumert, M.: The reproducibility of bio-acoustic features is associated with sample duration, speech task, and gender. IEEE Transactions on Neural Systems and Rehabilitation Engineering **30**, 167–175 (2022)

[43] Feng, F., Zhang, Z., Tang, L., Qian, H., Yang, L.-Z., Jiang, H., Li, H.: Test-retest reliability of acoustic and linguistic measures of speech tasks. Computer Speech & Language **83**, 101547 (2024)

[44] Stegmann, G.M., Hahn, S., Liss, J., Shefner, J., Rutkove, S.B., Kawabata, K., Bhandari, S., Shelton, K., Duncan, C.J., Berisha, V.: Repeatability of commonly used speech and language features for clinical applications. Digital biomarkers **4**(3), 109–122 (2020)

[45] Çokal, D., Villalba, M., Rui, H., Palominos, C.F., Böke, A., Homan, P., Heusinger, K., Kambeitz, J., Hinzen, W.: What is the retest reliability of computationally extractable speech and language markers? Computer Speech & Language, 101981 (2026)

[46] Tröger, J., Baykara, E., Zhao, J., Ter Huurne, D., Possemis, N., Mallick, E., Schäfer, S., Schwed, L., Mina, M., Linz, N., *et al.*: Validation of the remote automated ki: e speech biomarker for cognition in mild cognitive impairment: verification and validation following dime v3 framework. Digital biomarkers **6**(3), 107–116 (2022)

[47] UNESCO, I.: Isced fields of education and training 2013 (isced-f 2013): manual to accompany the international standard classification of education 2011. Montreal: UNESCO IFS (2014)

[48] Adler, N.E., Epel, E.S., Castellazzo, G., Ickovics, J.R.: Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. Health psychology **19**(6), 586 (2000)

[49] Goodglass, H., Kaplan, E.: The Assessment of Aphasia and Related Disorders. Lea & Febiger, Philadelphia (1972)

[50] Kertesz, A.: The western aphasia battery. (No Title) (1982)

[51] Feenstra, H.E., Vermeulen, I.E., Murre, J.M., Schagen, S.B.: Online self-administered cognitive testing using the amsterdam cognition scan: establishing psychometric properties and normative data. Journal of Medical Internet Research **20**(5), 192 (2018)

[52] Kaplan, E., Goodglass, H., Weintraub, S.: Boston naming test. The Clinical Neuropsychologist (1983)

[53] Lee Meeuw Kjoe, P.R., Vermeulen, I.E., Rentergem, J.A., Wall, E., Schagen, S.: Standardized item selection for alternate computerized versions of rey auditory verbal learning test (-based) word lists. Journal of Clinical and Experimental Neuropsychology **44**(9), 681–701 (2022)

[54] Student: The probable error of a mean. Biometrika, 1–25 (1908)

[55] Kim, H.-Y.: Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. Restorative dentistry & endodontics **42**(2), 152–155 (2017)

[56] Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., et al.: Google usm: Scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037 (2023)

[57] Llera, A., Brammer, M., Oakley, B., Tillmann, J., Zabihi, M., Amelink, J., Mei, T., Charman, T., Ecker, C., Dell'Acqua, F., *et al.*: Evaluation of data imputation strategies in complex, deeply-phenotyped data sets: the case of the eu-aims longitudinal european autism project. BMC Medical Research Methodology **22**(1), 229 (2022)

[58] Mukherjee, S., Mez, J., Trittschuh, E.H., Saykin, A.J., Gibbons, L.E., Fardo, D.W., Wessels, M., Bauman, J., Moore, M., Choi, S.-E., *et al.*: Genetic data and cognitively defined late-onset alzheimer's disease subgroups. Molecular psychiatry **25**(11), 2942–2951 (2020)

[59] Chen, F.F.: Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural equation modeling: a multidisciplinary journal **14**(3), 464–504 (2007)

[60] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., *et al.*: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing **7**(2), 190–202 (2015)

[61] Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. Advances in neural information processing systems **9** (1996)

[62] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

[63] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[64] Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)

[65] Lord, F.M., Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company, Reading, MA (1968)

[66] Bentler, P.M.: Comparative fit indexes in structural models. Psychological bulletin **107**(2), 238 (1990)

[67] Tucker, L.R., Lewis, C.: A reliability coefficient for maximum likelihood factor

analysis. Psychometrika **38**(1), 1–10 (1973)

[68] Hu, L.-t., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal **6**(1), 1–55 (1999)

[69] Igolkina, A.A., Meshcheryakov, G.: semopy: A python package for structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal **0**(0), 1–12 (2020) https://doi.org/10.1080/10705511.2019.1704289 https://doi.org/10.1080/10705511.2019.1704289

[70] Brunet, É., *et al.*: Le Vocabulaire de Jean Giraudoux Structure et Évolution. Slatkine, ??? (1978)

[71] Honoré, A., *et al.*: Some simple measures of richness of vocabulary. Association for literary and linguistic computing bulletin **7**(2), 172–177 (1979)

[72] Covington, M.A., McFall, J.D.: Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics **17**(2), 94–100 (2010)

[73] Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)

[74] Masrani, V., Murray, G., Field, T., Carenini, G.: Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In: BioNLP 2017, pp. 232–237 (2017)

[75] Parsapoor, M., Alam, M.R., Mihailidis, A.: Performance of machine learning algorithms for dementia assessment: impacts of language tasks, recording media, and modalities. BMC medical informatics and decision making **23**(1), 45 (2023)

[76] Brysbaert, M., New, B.: Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. Behavior research methods **41**(4), 977–990 (2009)

[77] Brysbaert, M., Mandera, P., McCormick, S.F., Keuleers, E.: Word prevalence norms for 62,000 english lemmas. Behavior research methods **51**, 467–479 (2019)

[78] Hoffman, P., Lambon Ralph, M.A., Rogers, T.T.: Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. Behavior research methods **45**, 718–730 (2013)

[79] Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods **46**, 904–911 (2014)

Supplementary Section A Additional results

Table S1: Cohort A ($n = 294$) regression performance: R^2 and MSE of the prediction model (mean over both assessments) for all composite score targets. We report point estimates with 95% bootstrap CI. Note that the variance in true scores (Var(y)) contributes to the R^2 calculation, with lower Var(y) producing lower R^2 values given identical errors.

	Target	Cookie Theft	Picnic Scene	Journaling	Avg. Prediction	Var(y)
R^2	Language	0.19 [0.08, 0.28]	0.23 [0.13, 0.30]	0.15 [0.02, 0.25]	0.25 [0.16, 0.31]	0.90
	Memory	0.13 [0.04, 0.21]	0.13 [0.04, 0.22]	0.09 [-0.00, 0.17]	0.14 [0.05, 0.21]	1.12
	Executive Function	0.11 [0.03, 0.18]	0.17 [0.09, 0.24]	0.13 [0.04, 0.21]	0.18 [0.10, 0.24]	0.96
	Speed	0.04 [-0.04, 0.11]	0.09 [0.02, 0.16]	0.06 [-0.03, 0.13]	0.09 [0.03, 0.15]	0.89
MSE	Language	0.73 [0.61, 0.86]	0.70 [0.59, 0.82]	0.77 [0.65, 0.90]	0.68 [0.58, 0.80]	0.90
	Memory	0.97 [0.80, 1.16]	0.96 [0.79, 1.15]	1.02 [0.84, 1.21]	0.96 [0.79, 1.15]	1.12
	Executive Function	0.85 [0.69, 1.03]	0.79 [0.64, 0.96]	0.83 [0.68, 0.99]	0.79 [0.64, 0.95]	0.96
	Speed	0.86 [0.69, 1.04]	0.81 [0.65, 0.98]	0.84 [0.68, 1.03]	0.81 [0.65, 0.99]	0.89

Table S2: Test-retest reliability (ICC) of the cognitive composite scores and the speech-based predictions between baseline and follow-up (Cohort A). We report point estimates with 95% bootstrap CI.

	Language	*Memory*	*Executive Function*	*Speed*
Cognitive composite score	0.69 [0.62, 0.74]	0.68 [0.61, 0.74]	0.71 [0.65, 0.76]	0.76 [0.71, 0.81]
Predictions: Cookie Theft	0.65 [0.58, 0.71]	0.78 [0.73, 0.82]	0.61 [0.54, 0.68]	0.60 [0.52, 0.67]
Predictions: Picnic Scene	0.75 [0.70, 0.80]	0.84 [0.80, 0.87]	0.68 [0.62, 0.74]	0.68 [0.62, 0.74]
Predictions: Journaling	0.57 [0.49, 0.64]	0.76 [0.70, 0.80]	0.55 [0.46, 0.62]	0.56 [0.48, 0.63]
Predictions: Avg. Prediction	0.80 [0.76, 0.84]	0.89 [0.86, 0.91]	0.77 [0.71, 0.81]	0.74 [0.68, 0.78]

Table S3: Cohort A: likelihood-ratio tests for the nested model comparisons across all cognitive target scores. For the three nested alternatives (top three rows per task), a significant p-value rejects the alternative in favor of the full model. For the two stationarity relaxations (bottom two rows per task), a non-significant p-value is consistent with stationary residual variances.

		Language			Memory			Executive function			Speed		
Task	Comparison	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p
Cookie Theft	$\lambda = 1$ vs. Full	128.155	1	< 0.001	141.427	1	< 0.001	167.824	1	< 0.001	226.943	1	< 0.001
	No β vs. Full	78.453	1	< 0.001	135.023	1	< 0.001	92.510	1	< 0.001	108.939	1	< 0.001
	$\lambda = 1$, no β vs. Full	128.155	2	< 0.001	141.786	2	< 0.001	167.824	2	< 0.001	227.033	2	< 0.001
	Full vs. free σ^2_ϵ by wave	4.565	1	0.033	0.022	1	0.883	1.131	1	0.288	0.009	1	0.925
	Full vs. free σ^2_δ by wave	0.340	1	0.560	2.032	1	0.154	0.374	1	0.541	0.007	1	0.932
Picnic Scene	$\lambda = 1$ vs. Full	122.495	1	< 0.001	139.283	1	< 0.001	158.714	1	< 0.001	219.267	1	< 0.001
	No β vs. Full	108.024	1	< 0.001	137.495	1	< 0.001	113.399	1	< 0.001	148.013	1	< 0.001
	$\lambda = 1$, no β vs. Full	122.647	2	< 0.001	140.748	2	< 0.001	158.714	2	< 0.001	219.420	2	< 0.001
	Full vs. free σ^2_ϵ by wave	1.001	1	0.317	0.385	1	0.535	0.173	1	0.677	0.223	1	0.637
	Full vs. free σ^2_δ by wave	0.155	1	0.694	2.212	1	0.137	0.161	1	0.688	0.222	1	0.638
Journaling	$\lambda = 1$ vs. Full	124.765	1	< 0.001	146.403	1	< 0.001	163.561	1	< 0.001	225.620	1	< 0.001
	No β vs. Full	64.554	1	< 0.001	146.196	1	< 0.001	65.019	1	< 0.001	93.007	1	< 0.001
	$\lambda = 1$, no β vs. Full	124.959	2	< 0.001	148.438	2	< 0.001	163.562	2	< 0.001	226.211	2	< 0.001
	Full vs. free σ^2_ϵ by wave	3.142	1	0.076	1.599	1	0.206	0.125	1	0.724	0.249	1	0.618
	Full vs. free σ^2_δ by wave	0.313	1	0.576	3.131	1	0.077	0.183	1	0.669	0.020	1	0.888

Table S4: Test-retest reliability (ICC) comparing the Cookie Theft and Picnic Scene speech predictions of the same participant at the same assessment. Results are given for the *language* composite cognitive score, and computed across the entire cross-sectional data set and within each longitudinal cohort. We report point estimates with 95% CI.

Cohort	# Observations	ICC [95% CI]
Entire data	1455	0.81 [0.80, 0.83]
Cohort A	599	0.74 [0.70, 0.78]
Cohort B	140	0.83 [0.77, 0.88]

Table S5: Longitudinal change in speech-based predictions across assessments in Cohort B. We compared the baseline and follow-up assessment of the same participants, testing for significant differences with paired t-tests. Results were computed separately for each task's predictions and the average prediction across tasks.

Cognitive domain	Task	Baseline	Follow-up	% Improv. / % Decline	Difference	p-val
Language	Cookie Theft	0.43 ± 0.42	0.44 ± 0.44	48.6% / 51.4%	0.01 ± 0.30	0.89
	Picnic Scene	0.35 ± 0.48	0.35 ± 0.48	51.4% / 48.6%	0.00 ± 0.27	0.91
	Journaling	0.36 ± 0.42	0.37 ± 0.48	57.1% / 42.9%	0.01 ± 0.33	0.71
	Avg. Prediction	0.38 ± 0.41	0.39 ± 0.42	50.0% / 50.0%	0.01 ± 0.20	0.74
Memory	Cookie Theft	0.26 ± 0.38	0.24 ± 0.41	40.0% / 60.0%	-0.02 ± 0.24	0.55
	Picnic Scene	0.22 ± 0.42	0.22 ± 0.47	55.7% / 44.3%	0.00 ± 0.22	0.86
	Journaling	0.24 ± 0.42	0.26 ± 0.44	55.7% / 44.3%	0.02 ± 0.20	0.37
	Avg. Prediction	0.24 ± 0.38	0.24 ± 0.42	48.6% / 51.4%	0.00 ± 0.16	0.88
Executive Function	Cookie Theft	0.32 ± 0.39	0.35 ± 0.37	58.6% / 41.4%	0.03 ± 0.24	0.38
	Picnic Scene	0.29 ± 0.44	0.29 ± 0.44	51.4% / 48.6%	-0.01 ± 0.22	0.78
	Journaling	0.31 ± 0.37	0.33 ± 0.38	55.7% / 44.3%	0.02 ± 0.29	0.65
	Avg. Prediction	0.31 ± 0.37	0.32 ± 0.36	52.9% / 47.1%	0.01 ± 0.16	0.56
Speed	Cookie Theft	0.22 ± 0.33	0.24 ± 0.32	54.3% / 45.7%	0.02 ± 0.24	0.41
	Picnic Scene	0.19 ± 0.37	0.17 ± 0.37	44.3% / 55.7%	-0.01 ± 0.19	0.6
	Journaling	0.23 ± 0.32	0.21 ± 0.36	45.7% / 54.3%	-0.02 ± 0.27	0.49
	Avg. Prediction	0.21 ± 0.32	0.21 ± 0.31	48.6% / 51.4%	-0.00 ± 0.17	0.86

Supplementary Section B Results of replication analysis on Cohort B

Table S6: Attrition analysis Cohort B: Baseline demographic and cognitive characteristics of the $n = 103$ participants newly recruited for replication. We compared the participants who returned for a follow-up assessment ($n = 70$, Cohort B) and those who did not ($n = 33$).

	Non-returners	Cohort B (Baseline)	p-value
# Participants	33	70	
Age	66.24 ± 5.06	65.34 ± 5.94	0.46
Gender			0.67
Female	19 (57.6%)	44 (62.9%)	
Male	14 (42.4%)	26 (37.1%)	
Country			1.00
UK	33 (100.0%)	70 (100.0%)	
Education			0.18
High Education	24 (72.7%)	47 (67.1%)	
Low Education	9 (27.3%)	23 (32.9%)	
Socioeconomic Status	5.88 ± 1.54	5.73 ± 1.51	0.66
Mean Transcript Length (# words)	163.42 ± 57.08	187.06 ± 75.06	0.12
Mean Audio Length (s)	76.09 ± 23.60	85.86 ± 30.09	0.11
Mean Cognitive Score	0.24 ± 0.66	0.24 ± 0.73	0.99
Language	0.20 ± 0.91	0.36 ± 0.92	0.41
Executive Function	0.23 ± 0.95	0.21 ± 0.80	0.91
Memory	0.35 ± 0.92	0.33 ± 0.89	0.91
Speed	0.19 ± 0.91	0.07 ± 0.99	0.58

Table S7: Summary of composite cognitive scores for Cohort B. We report mean and standard deviation of score in each assessment and their difference.

	Baseline	Follow-up	% Improvement / % Decline	Difference	p-val
Memory	0.33 ± 0.89	0.23 ± 0.90	40.0% / 60.0%	-0.10 ± 0.62	0.18
Language	0.36 ± 0.93	0.53 ± 0.97	61.4% / 38.6%	0.17 ± 0.71	0.05
Speed	0.07 ± 1.00	0.24 ± 0.95	60.0% / 40.0%	0.17 ± 0.50	0.01
Executive Function	0.21 ± 0.81	0.40 ± 0.84	72.9% / 27.1%	0.20 ± 0.51	0.00

Table S8: Cohort B ($n = 70$) regression performance: R^2 and MSE of the prediction model (mean over both assessments) for all composite score targets. We report point estimates with 95% bootstrap CI. Note that the variance in true scores ($\mathrm{Var}(y)$) contributes to the R^2 calculation, with lower $\mathrm{Var}(y)$ producing lower R^2 values given identical errors.

	Target	Cookie Theft	Picnic Scene	Journaling	Avg. Prediction	$\mathrm{Var}(y)$
R^2	Language	0.30 [0.11, 0.45]	0.32 [0.11, 0.48]	0.22 [-0.01, 0.37]	0.32 [0.14, 0.46]	0.88
	Memory	0.16 [-0.04, 0.30]	0.19 [-0.06, 0.35]	0.21 [-0.01, 0.38]	0.21 [0.00, 0.35]	0.79
	Executive Function	0.12 [-0.13, 0.28]	0.16 [-0.11, 0.33]	0.06 [-0.20, 0.24]	0.15 [-0.08, 0.30]	0.67
	Speed	0.04 [-0.15, 0.19]	0.03 [-0.18, 0.18]	-0.00 [-0.24, 0.17]	0.04 [-0.15, 0.19]	0.93
MSE	Language	0.62 [0.38, 0.93]	0.60 [0.37, 0.89]	0.69 [0.45, 0.99]	0.60 [0.37, 0.89]	0.88
	Memory	0.67 [0.50, 0.86]	0.65 [0.49, 0.83]	0.62 [0.45, 0.83]	0.63 [0.46, 0.81]	0.79
	Executive Function	0.59 [0.42, 0.79]	0.56 [0.41, 0.74]	0.63 [0.44, 0.84]	0.57 [0.40, 0.76]	0.67
	Speed	0.89 [0.51, 1.40]	0.90 [0.54, 1.38]	0.93 [0.54, 1.45]	0.89 [0.51, 1.39]	0.93

Table S9: Test-retest reliability (ICC) of the cognitive composite scores and the speech-based predictions between baseline and follow-up (Cohort B). Point estimates with 95% bootstrap CI.

	Language	*Memory*	*Executive Function*	*Speed*
Cognitive composite score	0.72 [0.58, 0.82]	0.76 [0.64, 0.84]	0.81 [0.71, 0.88]	0.87 [0.79, 0.92]
Predictions: Cookie Theft	0.75 [0.63, 0.84]	0.82 [0.72, 0.88]	0.79 [0.69, 0.87]	0.73 [0.60, 0.82]
Predictions: Picnic Scene	0.84 [0.75, 0.90]	0.88 [0.82, 0.93]	0.87 [0.80, 0.92]	0.86 [0.79, 0.91]
Predictions: Journaling	0.74 [0.61, 0.83]	0.89 [0.83, 0.93]	0.71 [0.57, 0.81]	0.68 [0.53, 0.79]
Predictions: Avg. Prediction	0.89 [0.82, 0.93]	0.92 [0.88, 0.95]	0.90 [0.85, 0.94]	0.86 [0.78, 0.91]

Table S10: Absolute fit indices of the speech measurement model for each speech task and composite score for Cohort B. CFI and TLI above 0.95 and SRMR below 0.08 indicate adequate fit. The corresponding likelihood-ratio tests are reported in Supplementary Table S11.

	Language			Memory			Executive Function			Speed		
	TLI	CFI	SRMR	TLI	CFI	SRMR	TLI	CFI	SRMR	TLI	CFI	SRMR
Cookie Theft	1.03	1.02	0.03	1.01	1.01	0.04	0.97	0.98	0.06	0.99	1.00	0.05
Picnic Scene	1.03	1.02	0.02	0.99	0.99	0.05	0.99	1.00	0.05	1.02	1.01	0.03
Journaling	1.01	1.00	0.06	1.01	1.01	0.03	0.94	0.96	0.06	0.95	0.97	0.09

Table S11: Cohort B: likelihood-ratio tests for the nested model comparisons across all cognitive target scores. For the three nested alternatives (top three rows per task), a significant p-value rejects the alternative in favor of the full model. For the two stationarity relaxations (bottom two rows per task), a non-significant p-value is consistent with stationary residual variances.

		Language			Memory			Executive function			Speed		
Task	Comparison	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p
Cookie Theft	$\lambda = 1$ vs. Full	35.101	1	< 0.001	46.940	1	< 0.001	60.191	1	< 0.001	90.980	1	< 0.001
	No β vs. Full	14.424	1	< 0.001	60.594	1	< 0.001	57.504	1	< 0.001	49.978	1	< 0.001
	$\lambda = 1$, no β vs. Full	35.101	2	< 0.001	47.770	2	< 0.001	62.167	2	< 0.001	92.077	2	< 0.001
	Full vs. free σ^2_ε by wave	0.517	1	0.472	1.579	1	0.209	0.171	1	0.679	0.004	1	0.948
	Full vs. free σ^2_δ by wave	0.599	1	0.439	0.103	1	0.748	0.456	1	0.499	0.868	1	0.351
Picnic Scene	$\lambda = 1$ vs. Full	28.552	1	< 0.001	41.827	1	< 0.001	54.052	1	< 0.001	88.909	1	< 0.001
	No β vs. Full	19.435	1	< 0.001	43.907	1	< 0.001	78.766	1	< 0.001	91.442	1	< 0.001
	$\lambda = 1$, no β vs. Full	28.552	2	< 0.001	44.141	2	< 0.001	58.884	2	< 0.001	93.137	2	< 0.001
	Full vs. free σ^2_ε by wave	0.311	1	0.577	4.712	1	0.030	0.036	1	0.849	0.042	1	0.838
	Full vs. free σ^2_δ by wave	0.813	1	0.367	0.322	1	0.571	0.445	1	0.505	0.562	1	0.453
Journaling	$\lambda = 1$ vs. Full	34.382	1	< 0.001	42.836	1	< 0.001	63.373	1	< 0.001	91.392	1	< 0.001
	No β vs. Full	27.934	1	< 0.001	44.219	1	< 0.001	41.780	1	< 0.001	41.391	1	< 0.001
	$\lambda = 1$, no β vs. Full	34.382	2	< 0.001	44.602	2	< 0.001	65.613	2	< 0.001	92.970	2	< 0.001
	Full vs. free σ^2_ε by wave	1.931	1	0.165	0.492	1	0.483	0.319	1	0.572	3.055	1	0.081
	Full vs. free σ^2_δ by wave	1.043	1	0.307	0.006	1	0.939	0.187	1	0.666	0.576	1	0.448

Table S12: Bootstrap test results assessing whether the speech-based MDC is significantly larger than the cognitive assessment's MDC on Cohort B. For each task and composite cognitive score, we report the 95% CI of the difference (speech MDC minus cognitive MDC) across 10000 bootstrap replicates, and the proportion of replicates where the speech MDC is larger (less sensitive) than the cognitive MDC (% larger). An asterisk (*) on the CI indicates that the CI excludes zero, i.e. the speech MDC is significantly larger than the cognitive MDC.

	Language		Memory		Executive Function		Speed	
	95% CI	% larger	95% CI	% larger	95% CI	% larger	95% CI	% larger
Journaling	-0.14 – 2.16	93.9%	-0.35 – 0.93	64.1%	0.53 – 8.63*	99.9%	1.53 – 86.55*	100.0%
Cookie Theft	-0.29 – 1.06	81.8%	0.07 – 2.11*	98.6%	0.11 – 4.05*	98.5%	1.18 – 32.42*	100.0%
Picnic Scene	-0.55 – 0.51	39.0%	-0.22 – 1.08	81.5%	-0.19 – 2.01	91.8%	0.67 – 14.62*	100.0%
Avg. Speech Pred.	-0.75 – 0.24	12.9%	-0.45 – 0.57	37.5%	-0.30 – 1.99	86.8%	0.60 – 24.84*	100.0%

Supplementary Section C Methodological details

Table S13: Statistics for the original cognitive scores (ACS tests and standardized language tests). Reproduced with adaptations from Heitz et al. [21]. RAVLT: Rey Auditory Verbal Learning Test.

Test	Outcome measure
Trail Making Test A	Completion time
Trail Making Test B	Completion time
RAVLT (Learning)	Total number of correct words
RAVLT (Recall)	Total number of correct words
RAVLT (Recognition)	Total number of correct words
Visual Reaction Time	Mean reaction time
Corsi Block-tapping Test	Total number of correctly repeated sequences
Grooved Pegboard	Completion time
Digit Span (forward)	Total number of correctly repeated sequences
Digit Span (backward)	Total number of correctly repeated sequences
Clicking speed test	Completion time
Mouse dragging speed test	Completion time
Phonemic Fluency (”F” fluency)	Total number of valid words
Semantic Fluency (Category fluency)	Total number of valid words
Boston Naming Test	Total number of correctly named objects

Table S14: Standardized factor loadings from the confirmatory factor analysis (CFA) of the cognitive tests onto the four composite scores: Memory, Language, Speed, and Executive Function. These loadings were obtained in a confirmatory factor analysis in our previous work [21], based on the complete baseline assessment of 1003 participants.

Factor	Indicator	Loading
Memory	RAVLT Learning	0.79
	RAVLT Delayed Recall	0.85
	RAVLT Recognition	0.59
Language	Semantic Fluency	0.77
	Phonemic Fluency	0.44
	Boston Naming Test	0.41
Speed	Visual Reaction Time	0.31
	Grooved Pegboard	0.83
	Click Skill	0.68
	Drag Skill	0.70
	Trail Making Test A	0.49
Executive Function	Trail Making Test A	0.36
	Trail Making Test B	0.87
	WAIS III Digit Span Backward	0.28
	WAIS III Digit Span Forward	0.36

Table S15: Linguistic and acoustic features extracted for the spontaneous speech audio recordings and corresponding transcripts. Reproduced from Heitz et al. [21].

Feature group **Feature name**	**Description**
Linguistic	
verb_noun_ratio	Ratio of verbs to nouns
subordinate_coordinate_con-junction_ratio	Ratio of subordinate to coordinate conjunctions
adverb_ratio	Ratio of adverbs to all words
noun_ratio	Ratio of nouns to all words
verb_ratio	Ratio of verbs to all words
pronoun_ratio	Ratio of pronouns to all words
personal_pronoun_ratio	Ratio of personal pronouns to all words
determiner_ratio	Ratio of determiners to all words
preposition_ratio	Ratio of prepositions to all words
verb_present_participle_ratio	Ratio of verbs (present participle) to all words
verb_modal_ratio	Ratio of modal verbs to all words
verb_third_person_singular_ratio	Ratio of verbs in 3rd person singular to all words
interjection_ratio	Ratio of interjections (mostly hesitations, such as uh) to all words
NP ->PRP	Count based on constituency parsing
ROOT ->FRAG	Count based on constituency parsing
VP ->VBG	Count based on constituency parsing
VP ->VBG_PP	Count based on constituency parsing
VP ->VBD_NP	Count based on constituency parsing
INTJ ->UH	Count based on constituency parsing
NP_ratio	Ratio of NP constituents
VP_ratio	Ratio of VP constituents
avg_n_words_in_NP	Average number of words in a noun phrase
n_words	Number of words in the transcript
avg_word_length	Average word length (number of letters)
words_not_in_dict_ratio	Number of words not in an English dictionary
brunets_index	Brunét's index [70], a measure of lexical richness.
honores_statistic	Honoré Statistic [71], a measure of lexical richness
moving_average_type_token_ratio	The moving average type-token ratio [72] with a window size of 20 words
n_unique_in_50	Average number of unique words in a 50 word sample [19]

flesch_kincaid	Flesch-Kincaid Grade Level Formula Kincaid et al. [73], a measure of readability.
avg_distance_between_utterances	The average cosine distance between two sentences, based on Masrani et al. [74]'s implementation
prop_utterance_dist_below_05	Proportion of sentence pairs with cosine distance <= 0.5, based on Masrani et al. [74].
propositional_density	Propositional density [75]: Ratio of verbs, adjectives, adverbs, prepositions, and conjunctions to all words
content_density	Content density [75]: Ratio of nouns, verbs, adjectives, and adverbs to all words
frequency_noun	Mean word frequency of nouns Brysbaert and New [76]
age_of_acquisition_noun	Mean age of acquisition of nouns Brysbaert et al. [77]
familiarity_noun	Mean word familiarity of nouns Brysbaert et al. [77]
ambiguity_noun	Mean ambiguity of nouns Hoffman et al. [78]
noun_concreteness	Mean concreteness of words Brysbaert et al. [79]
Audio — Hand-crafted	
audio_length	Audio length in seconds
fraction_of_pause	Fraction of pause segments over audio length
mean_duration	Mean pause duration
std_duration	Standard deviation of pause durations
speech_rate	Number of spoken phonemes divided by the total audio length
articulation_rate	Number of spoken phonemes divided by the duration of voice-activated segments
Audio — eGeMAPS [60]	
F0semitoneFrom27.5Hz_sma3nz _meanRisingSlope	The mean rising slope of pitch (logarithmic F0 on a semitone frequency scale) smoothed over time
loudness_sma3_percentile20.0	The 20th percentile loudness (perceived signal intensity)
loudness_sma3_percentile50.0	The 50th percentile loudness (perceived signal intensity)
loudness_sma3_percentile80.0	The 80th percentile loudness (perceived signal intensity)
loudness_sma3_meanRisingSlope	The mean rising slope of loudness
spectralFlux_sma3 _amean	Mean Spectral flux (difference of the spectra of two consecutive frames)
spectralFlux_sma3 _stddevNorm	Standard deviation of spectral flux
mfcc1_sma3_stddevNorm	Standard deviation of Mel-Frequency Cepstral Coefficient 1
mfcc3_sma3_amean	Mean of Mel-Frequency Cepstral Coefficient 3
mfcc3_sma3_stddevNorm	Standard deviation of Mel-Frequency Cepstral Coefficient 3
F1bandwidth_sma3nz_stddevNorm	Standard deviation of bandwidth of Formant 1
F1amplitudeLogRelF0_sma3nz_amean	Mean of the relative energy of Formant 1 relative to Pitch (F0) frequency
alphaRatioV_sma3nz_stddevNorm	Standard deviation of the alpha ratio (ratio between energy high and low frequency bands) in voiced regions
spectralFluxV_sma3nz_stddevNorm	Standard deviation of spectral flux in voiced regions
mfcc1V_sma3nz_stddevNorm	Standard deviation of MFCC 1 in voiced regions
alphaRatioUV_sma3nz_amean	Mean of the alpha ratio (ratio between energy high and low frequency bands) in unvoiced regions
spectralFluxUV_sma3nz_amean	Mean of spectral flux in unvoiced regions
loudnessPeaksPerSec	Number of loudness peaks per second
VoicedSegmentsPerSec	Number voiced segments per second
MeanUnvoicedSegmentLength	The mean length of the unvoiced segments
equivalentSoundLevel_dBp	Equivalent sound level (LEq)

Supplementary Section D Regression algorithm hyperparameter tuning

For each model family (Ridge regression, Support Vector Regression, and XGBoost), we performed randomized hyperparameter search with 200 sampled configurations, evaluated using a 3-fold cross-validation scheme on the subset of 2024 baseline participants who did not return for re-assessment. Analyses were implemented in Python 3.12 using `scikit-learn` 1.8.0 (`Ridge`, `SVR`, `LinearSVR`) and `xgboost` 3.2.0 (`XGBRegressor`). The Ridge grid varied the regularization strength α over 200 values log-spaced in $[10^{-4}, 10^{5}]$. The SVR family considered two kernels (`linear` and `rbf`), with 200 configurations sampled per kernel: for the linear kernel, `C` varied over 20 values log-spaced in $[10^{-3}, 10^{1}]$, `epsilon` over 15 values log-spaced in $[10^{-3}, 10^{-0.3}]$, and the loss function over {`epsilon_insensitive`, `squared_epsilon_insensitive`}; for the RBF kernel, `C` varied over 20 values log-spaced in $[10^{-2}, 10^{2}]$, `epsilon` over 15 values log-spaced in $[10^{-3}, 10^{-0.3}]$, and γ over {scale, auto} plus 10 values log-spaced in $[10^{-4}, 10^{-1}]$. The XGBoost grid varied `n_estimators` $\in \{200, 400, 800, 1200\}$, `learning_rate` $\in \{0.01, 0.03, 0.05, 0.1\}$, `max_depth` $\in \{2, 3, 4, 5, 6\}$, `min_child_weight` $\in \{1, 3, 5, 7, 10\}$, `subsample` $\in \{0.6, 0.8, 1.0\}$, `colsample_bytree` $\in \{0.5, 0.7, 0.9, 1.0\}$, `reg_alpha` $\in \{0, 0.001, 0.01, 0.1, 1\}$, `reg_lambda` $\in \{0.5, 1, 3, 5, 10\}$, and `gamma` $\in \{0, 0.01, 0.1, 0.3, 1\}$. All searches used a fixed random seed for reproducibility.

The best hyper parameters for each model family are provided in Table S16.

Table S16: Best hyperparameters for each model family.

Model family	Hyperparameters
SVR RBF	gamma: 0.001, epsilon: 0.321471781635738, C: 3.359818286283781
Ridge	alpha: 361.23426997094305
SVR linear	loss: squared_epsilon_insensitive, epsilon: 0.0015590395868560912, C: 0.001623776739188721
XGBoost	subsample: 0.8, reg_lambda: 1, reg_alpha: 0.01, n_estimators: 400, min_child_weight: 1, max_depth: 4, learning_rate: 0.01, gamma: 1, colsample_bytree: 0.7

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf