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		
		$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p	$\Delta\chi^2$	Δdf	p
Task	Comparison												
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