A large public dataset of computerized cognitive test results with high-resolution speech quantification

Background: We present a publicly available normative dataset with results from the California Cognitive Assessment Battery (CCAB), to support research into cognition, aging, and speech and language biomarkers. This dataset is hosted on Open Science Framework (OSF) to facilitate collaborative and/or independent secondary analyses by other researchers.

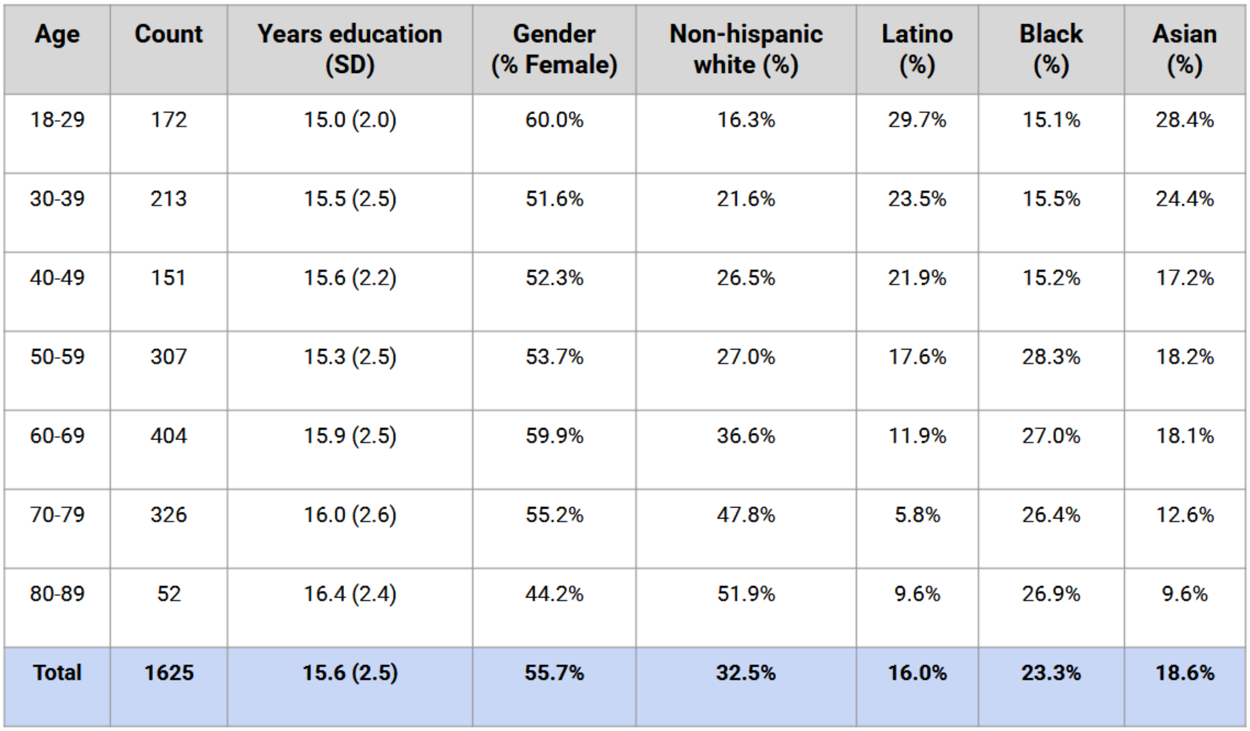
Methods: 1,631 participants (55.7% female, 37% white; mean age 55.2 ± 17.0 years; see Table 1) completed the CCAB normative protocol. Most completed ~6 total hours of supervised testing across three consecutive days, with identical tests administered on days 2 and 3 to measure retest learning effects; others completed a single two-hour test session. Some participants also completed follow-up assessments at one (N=480), two (N=244), and three (N=174) years post-enrollment. More than 95% of participants were tested in their homes using telemedical proctoring.

CCAB is an automated, computerized test battery with 22 verbal and 18 non-verbal tasks, administered on calibrated hardware with Presentation® software for sub-millisecond temporal precision. CCAB tasks span multiple cognitive domains and response modalities, including touch (e.g., trail making), mouse (e.g., choice reaction time), and speech (e.g., picture description). Extended demographic and health data were also collected, including psychological scales (e.g., GDS, CFQ, GAD-7), a vocabulary-based estimate of premorbid intelligence, and lifestyle and medical history.

For speech tasks, audio recordings were transcribed using consensus automatic speech recognition (CASR), which combines multiple ASR engines to achieve >98% word-level transcription accuracy. High-quality recordings and high-accuracy automated transcriptions enable fine-grained quantification of speech timing, fluency, syntax, and phonetics.

Results: Each individual test run (N > 100,000) was analyzed to produce summary performance metrics such as category-wise response counts, speech complexity measures, response timing, and speech acoustics. Results were collated by task, deidentified, and paired with metadata and data dictionaries following OSF best practices, and are available on OSF.

Conclusion: We invite researchers to explore and further analyze this large, high-resolution dataset for studies in neuropsychology, speech-language processing, longitudinal cognition, or machine learning. We will continue to update the OSF data repository (https://osf.io/x8u5z/) with new and more granular data.



Age	Count	Years education (SD)	Gender (% Female)	Non-hispanic white (%)	Latino (%)	Black (%)	Asian (%)
18-29	172	15.0 (2.0)	60.0%	16.3%	29.7%	15.1%	28.4%
30-39	213	15.5 (2.5)	51.6%	21.6%	23.5%	15.5%	24.4%
40-49	151	15.6 (2.2)	52.3%	26.5%	21.9%	15.2%	17.2%
50-59	307	15.3 (2.5)	53.7%	27.0%	17.6%	28.3%	18.2%
60-69	404	15.9 (2.5)	59.9%	36.6%	11.9%	27.0%	18.1%
70-79	326	16.0 (2.6)	55.2%	47.8%	5.8%	26.4%	12.6%
80-89	52	16.4 (2.4)	44.2%	51.9%	9.6%	26.9%	9.6%
Total	**1625**	**15.6 (2.5)**	**55.7%**	**32.5%**	**16.0%**	**23.3%**	**18.6%**

Figure 1. Gender, race, and education demographics information for the CCAB normative subject pool, split by age band.