# Analytical and Clinical Validation of Digital Language Assessments

Jessica Robin [1], Mengdan Xu [1], William Simpson [1,2], Jekaterina Novikova [1]

(1) Winterlight Labs, Toronto, ON, Canada, (2) Department of Psychiatry and Behavioural Neuroscience, McMaster University, Hamilton, ON, Canada

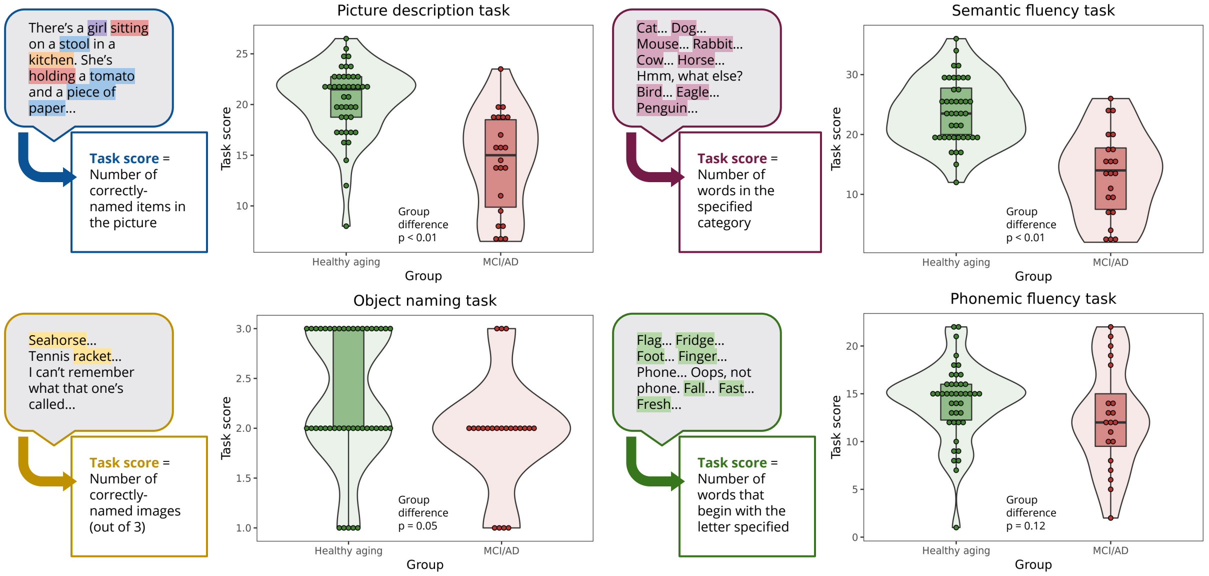
## Background

Digital tools offer new possibilities for cognitive assessment that may be more sensitive to cognitive changes and less burdensome to patients.[1,2] These novel technologies require both analytical and clinical forms of validation to ensure they are fit for purpose.[3,4] As described in the V3 framework, analytical validation verifies that a measure is accurately measuring the outcome of interest.[3] Clinical validation serves to test the relationship of a given outcome measure with a clinical condition or symptom. In this study, we evaluate the analytical validity (i.e. how accurate are the automated scores?) and clinical validity (i.e. are the scores sensitive to clinical differences?) of digital language assessments in older adults. To accomplish this goal, we test the properties of automated versions of standard assessments and their outcome scores from four language tasks.

## Methods

- The Winterlight App provides a range of digital language assessments including standard neuropsychological language tests such as: picture description, object naming, phonemic fluency and semantic fluency.
- In each task, participants are guided through the task and prompted to make verbal responses, describing a picture, naming objects displayed on a screen, or naming as many words as possible in a minute that fit into a certain category (i.e. animals or words that start with the letter F).
- Verbal responses are recorded by the app, transcribed and analyzed, generating >500 variables that describe the acoustic and linguistic characteristics of the speech recording.
- For each task, a standard score is generated reflecting performance on the task following standard scoring practices.
- For analytical validation, two trained human raters manually scored 150-200 recordings each made by healthy older adults (MoCA scores >= 26) for each of the speech tasks.
- Pearson correlations were computed between the raters and the automated scores, and between the two human raters, for comparison.
- For clinical validation, scores on each task were compared between groups of healthy older adults (MoCA scores >= 26, N = 43) and those with cognitive impairment due to mild cognitive impairment (MCI) or early Alzheimer's disease (AD) (N = 22) using linear regression models with factors of group, age, sex and years of education.

**Figure 1: Clinical validation of task scores for detecting language changes in MCI/early AD**



There's a girl sitting on a stool in a kitchen. She's holding a tomato and a piece of paper…

**Task score** = Number of correctly-named items in the picture

Cat… Dog… Mouse… Rabbit… Cow… Horse… Hmm, what else? Bird… Eagle… Penguin…

**Task score** = Number of words in the specified category

Seahorse… Tennis racket… I can't remember what that one's called…

**Task score** = Number of correctly-named images (out of 3)

Flag… Fridge… Foot… Finger… Phone… Oops, not phone. Fall… Fast… Fresh…

**Task score** = Number of words that begin with the letter specified

**Table 1: Analytical validation to assess agreement between manual and automated scoring**

| Language score | Agreement between automated scores and human raters (*r*) | Inter-rater agreement between two human raters (*r*) |
|---|---|---|
| Picture description | 0.66-0.71 | 0.76 |
| Object Naming | 0.63-0.66 | 0.92 |
| Semantic Fluency | 0.67-0.83 | 0.96 |
| Phonemic Fluency | 0.93-0.96 | 0.99 |

## Conclusions

This study evaluates the use of digital language assessments and automated scoring for assessing language abilities in older adults. Agreement between automated scores and human scorers was highest for phonemic fluency, and comparable to interrater agreement for picture description. Picture description and semantic fluency scores were the most sensitive to differences between healthy participants and those with MCI or early AD. Overall, a digital version of the picture description task appears to be as reliable as human scoring and the most sensitive to detecting cognitive impairment, supporting the utility of digital assessments to assess cognition. Digital speech assessments can be used remotely, enabling faster, safer and less burdensome screening and monitoring for dementia.

## References

(1) Kourtis, L. C., Regele, O. B., Wright, J. M. & Jones, G. B. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. npj Digital Med 2, 9 (2019).
(2) Dagum, P. Digital biomarkers of cognitive function. npj Digital Med 1, 10 (2018).
(3) Goldsack, J. C. et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). npj Digit. Med. 3, 55 (2020).
(4) Robin, J. et al. Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit Biomark 99–108 (2020) doi:10.1159/000510820.

WINTERLIGHT

Presented at ISCTM 17th Annual Scientific Meeting. April 6-9, 2021

McMaster University