Exploring a Burst Design to Improve Psychometric Properties of a Remote Digital Cognitive Assessment for Clinical Trials in AD

Johannes Tröger[1], Louisa Schwed[1], Nicklas Linz[1], Alexandra König[1], Simona Schäfer[1]

[1] ki:elements, Saarbrücken, Germany

Methodological Question

We investigate whether a burst design—averaging multiple repeated remote cognitive assessments over a short period—can improve psychometric properties such as reliability, measurement error, and sensitivity to change compared to single-timepoint assessments. This approach aims to enhance the utility of digital cognitive assessments as reliable endpoints in Alzheimer's disease clinical trials.

Introduction

Digital cognitive assessments (DCAs) offer a promising avenue for measuring cognitive function remotely and at scale, particularly in clinical trials for Alzheimer's disease (AD). However, single-instance remote assessments often suffer from poor psychometric properties due to high intra-individual variability and novelty effects, resulting in low test–retest reliability and reduced sensitivity to change. These limitations undermine their utility as clinical trial endpoints, where cognitive changes are subtle yet crucial for early decision-making.

To address this challenge, we explore a burst design—averaging multiple repeated assessments over a short period—to improve reliability by smoothing out day-to-day performance fluctuations. We evaluate this approach using the SB-C, a validated digital cognitive screener commonly employed in AD clinical research. Its remote, low-burden, and automated nature makes it well-suited for cost-effective repeated administration. We present results from a healthy control sample to investigate whether burst design aggregation improves psychometric properties and enhances suitability for endpoint use in clinical trials.

Methods

We conducted a fully remote study using the Mili ki:elements app to administer the SB-C through the crowdsourcing platform Prolific. A total of 70 healthy adult participants were enrolled (mean age 45.57 ± 18.54 years; 46 female). Each participant completed three parallel versions of the SB-C at three separate timepoints, spaced one week apart. The SB-C is based on automatic speech analysis of audio recordings from two widely used neuropsychological tasks: Semantic Verbal Fluency (SVF) and the Rey Auditory Verbal Learning Test (RAVLT).

To evaluate the psychometric properties of a burst design versus a traditional application, we compared key metrics across individual and aggregated timepoints. This involved computing the intra-class correlation coefficient (ICC) using a two–way mixed–effects model to assess test–retest reliability and determining the within-subject standard deviation (SD) to quantify individual variability. We then calculated the standard error of measurement (SEM), from which we derived the minimal detectable change at 95% confidence (MDC). We evaluated whether aggregating two consecutive timepoints (i.e., mean(t1,t2) vs. t3 and t1 vs. mean(t2,t3)) improved reliability (ICC) and reduced SEM and MDC, compared to using single assessments (i.e., t1 vs. t3). This allowed us to quantify the potential benefits of a burst design for increasing signal stability and interpretability in repeated cognitive testing.

Results

Results showed that when using a single timepoint, the minimal detectable change (MDC) for the overall SB-C cognition z-score was approximately 0.80, with an SEM of 0.30 and ICC of 0.81—indicating that a substantial change would be required to confidently detect a true signal beyond measurement noise.

In contrast, aggregating two adjacent timepoints (t1,t2 and t2,t3) significantly reduced the MDC (to 0.62 and 0.55 z-score units), lowered the SEM (to 0.22 and 0.20), and slightly increased the ICC (to 0.86 and 0.85). This improvement enables the detection of smaller cognitive changes with the same 95% confidence level.

Conclusion

These findings support the utility of a burst assessment design in improving the psychometric properties of remote DCAs. By aggregating data from repeated administrations of the SB-C, we observed a marked reduction in measurement error and an increase in reliability, as indicated by lower SEM and MDC values. This enhancement enables more sensitive detection of cognitive changes, which is critical for evaluating subtle treatment effects in Alzheimer's disease clinical trials. Our results suggest that incorporating burst designs into remote cognitive assessment protocols can significantly improve signal quality and strengthen their viability as clinical trial endpoints.

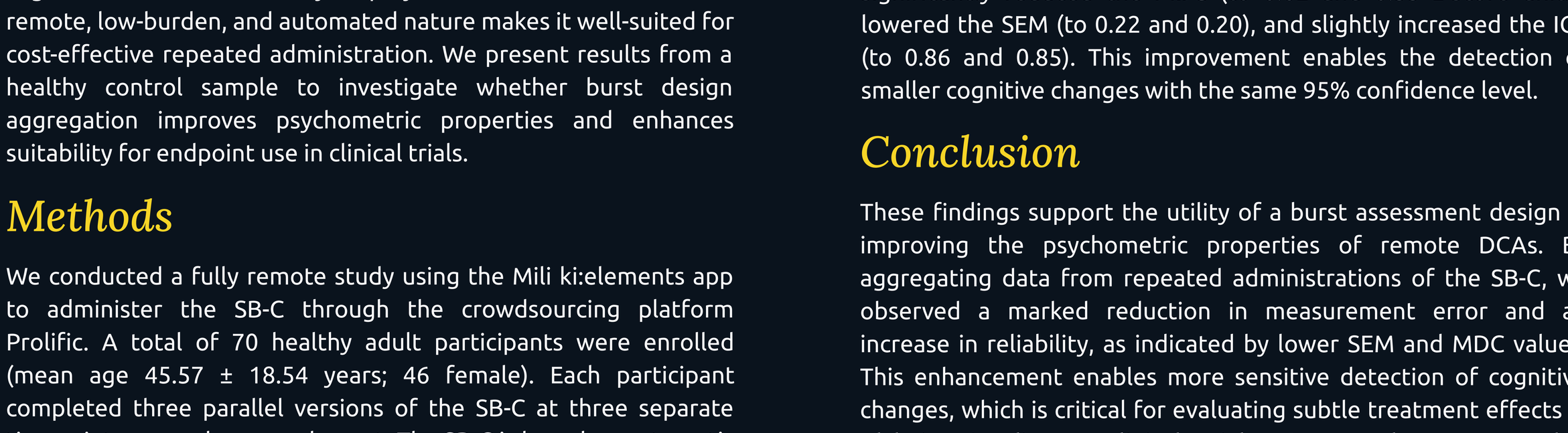

Figure: Scatterplots depicting the relationship between two consecutive assessments; left panel in yellow showy t1-t3 whereas the two right panels in purple show how the relationship becomes stronger by using a second timepoint for meaning (ICC (mean(t1_t2) & t3) or ICC (mean(t2_t3) & t1)).