Test–retest reliability of experimental language tasks in post-stroke aphasia: A large sample study with multiple sessions

by Zakariás Lilla | Christos Salis | Ágnes Lukács | Eötvös Loránd University, Bárczi Gusztáv Faculty of Special Needs Education, Budapest, Hungary; Semmelweis University, Rehabilitation Clinic, Budapest | Speech and Language Sciences, Newcastle University, Newcastle upon Tyne, UK | Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary

Abstract ID: 141
Event: SoA 2025 Copenhagen posters
Topic: Clinical and experimental work on aphasia and related disorders

Introduction and aims

Knowing that an assessment or treatment task has good test–retest reliability is crucial, as it indicates that the results obtained in a single session reliably reflect a person's ability and remain stable over time. In aphasia, it has gained increasing attention, particularly in the context of standardized assessments of language and cognition (e.g., working memory; DeDe et al., 2014), spoken discourse analysis (e.g., Stark et al., 2023), and eye-tracking studies using the visual-world paradigm (e.g., Mack et al., 2016). However, test–retest reliability data are still lacking for many commonly used experimental language tasks in aphasia, underscoring the need to assess their temporal stability. The aim of this study was to assess test–retest reliability of experimental language tasks designed to measure phonological, lexical, and semantic processing, based on the cognitive neuropsychological approach (Whitworth et al., 2014). Specifically, the tasks targeted auditory phonological analysis, the phonological input lexicon, and the semantic system involved in word processing, using phoneme identification, auditory lexical decision, and auditory animacy decision tasks.

Methods

Participants

Data from 55 adults with post-stroke aphasia (29 women; mean age = 57.95 years, SD =12.25 years; post-onset = 15.54 months, SD = 36.45 months) was included. Inclusion criteria were: aphasia due to stroke (diagnosed using the Western Aphasia Battery [WAB], Hungarian adaptation: Osmánné Sági, 1991), native Hungarian speaker, adequate hearing, physical ability to complete the tasks, and ≥50% accuracy in auditory word comprehension on the WAB. Exclusion criteria included major neurological or psychiatric disorders, moderate-to-severe hearing impairment, and global aphasia.

Procedures and design

Participants were assessed on four approximately consecutive days (session 1–4; mean

interval = 6.25 days) using the same set of three auditory tasks on each day. The experimental tasks assessed phonological, lexical, and semantic processing under low and high WM demand. Participants completed three structurally identical yes/no decision tasks: phoneme identification, lexical decision, and animacy decision. In the low WM condition, participants were asked to indicate whether (1) the heard phoneme string contained the phoneme /b/ or not (phoneme identification), (2) the phoneme string was a real word or a nonword (lexical decision), or (3) the word referred to a living or non-living entity (animacy decision). In the high WM condition, two auditory stimuli were presented, and participants indicated whether only one of them (vs. both or neither) (1) contained the phoneme /b/, (2) was a real word, or (3) referred to a living entity, depending on the task. The tasks were programmed in PsychoPy, and auditorily stimuli were presented as prerecorded audio files. Stimulus presentation lasted ~1 second in the low WM and ~2 seconds in the high WM condition. Participants had up to six seconds to respond, including stimulus presentation time. To assess test–retest reliability, intraclass correlation coefficients (ICCs) were calculated using a two-way random-effects model with absolute agreement, based on single measurements across all combinations of the four time points, for both accuracy (ACC) and reaction times (RTs) in each condition (Koo & Li, 2016).

Results

ACC was highest in the lexical decision task, ranging from 91–96% in the low WM condition and 70–80% in the high WM condition across the four time points. In the animacy decision task, ACC ranged from 86–91% in the low WM condition and 60–70% in the high WM condition. Performance was lowest in the phoneme identification task, with ACC ranging from 68–75% in the low WM condition and 54–60% in the high WM condition.

Except for the low WM condition in lexical decision, ICCs primarily ranged from moderate to excellent across tasks and measures. Phoneme identification demonstrated the highest reliability, particularly for ACC in the low WM condition (ICCs mostly > 0.80), with lower reliability in the high WM condition (ICCs = 0.47–0.82). RTs were consistently reliable, especially at later time points (ICCs = 0.78–0.89). Lexical decision accuracy exhibited poor to moderate reliability, especially in the low WM condition (ICCs = 0.03–0.69), whereas the high WM condition showed moderate to good reliability (ICCs = 0.62–0.72). RTs for lexical decision were also moderately reliable, particularly at later sessions (ICCs = 0.71–0.78, excluding session 1). Animacy decision accuracy and RTs showed fair to good reliability (ICCs = 0.45–0.85), with little difference between WM conditions.

For phoneme identification, confidence intervals (CIs) for ACC were narrow, especially at later time points (e.g., 0.70–0.89), indicating stability. Lexical decision accuracy had wider CIs, often crossing zero, suggesting low reliability for many comparisons. RTs had generally wider CIs than ACC, reflecting greater variability across sessions. Among the 12 cases (3 tasks x 2 conditions x 2 measures), ICCs were highest between session 2 and 3 in five cases,

between session 3 and 4 in four cases, between session 1 and 2 in two cases, and between session 2 and 4 in one case (Table 1). ICCs were higher when excluding session 1 in all but one case.

Discussion

The highest reliability was found for phoneme identification, followed by animacy decision, with lexical decision showing poor to moderate reliability for ACC. ICCs were highest when analyses excluded the first time point, suggesting greater stability at later time points, potentially due to initial adaptation or learning effects. Based on ICC and CI patterns, ACC was more reliable than RTs. Interestingly, task complexity did not appear to affect reliability. While one might expect higher complexity to introduce more measurement error, which typically decreases reliability, this was not observed. The higher complexity conditions did not show decreased reliability, likely because the increased between-subject variance in these conditions positively contributed to reliability (Hedge et al., 2018). Overall, these tasks were generally reliable, except in conditions where near-ceiling performance limited data variance. This highlights a common challenge in reliability studies: when performance is too high, reliability can be artificially inflated, as there is little room for variability. The second time point contributed most to increased reliability, with minimal improvement thereafter, suggesting that initial learning or adaptation effects stabilized over subsequent sessions.

References

DeDe, G., Ricca, M., Knilans, J., & Trubl, B. (2014). Construct validity and reliability of working memory tasks for people with aphasia. *Aphasiology, 28*(6), 692–712.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50,* 1166–1186.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Mack, J. E., Wei, A. Z. S., Gutierrez, S., & Thompson, C. K. (2016). Tracking sentence comprehension: Test-retest reliability in people with aphasia and unimpaired adults. *Journal of Neurolinguistics, 40*, 98–111.

Osmánné Sági J. (1991). Az afázia klasszifikációja és diagnosztikája. *Ideggyógyászati Szemle, 44,* 339–362.

Stark, B. C., Alexander, J. M., Hittson, A., Doub, A., Igleheart, M., Streander, T., & Jewell, E. (2023). Test–retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research, 66*(7),

2316–2345.

Whitworth, A., Webster, J., & Howard, D. (2014). A cognitive neuropsychological approach to assessment and intervention in aphasia: A clinician's guide. Psychology Press.