ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

Test–Retest Reliability of Core Lexicon Analysis

Sarah Grace Dalton,[a] Robert Cavanaugh,[b] and Brielle C. Stark[c]

[a] Department of Communication Sciences and Special Education, University of Georgia, Athens [b] School of Health and Rehabilitation Sciences, MGH Institute of Health Professions, Boston, MA [c] Department of Speech, Language and Hearing Sciences, Indiana University Bloomington

ARTICLE INFO

Article History:
Received September 16, 2025
Revision received May 4, 2026
Accepted August 4, 2026

Editor-in-Chief: Catherine A. Off
Editor: Marion C. Leaman

https://doi.org/10.1044/2026_AJSLP-25-00438

ABSTRACT

Purpose: We examined the test–retest reliability of core lexicon analysis for healthy control participants (HCPs) and individuals with aphasia (IWAs) to facilitate research and clinical implementation of discourse analysis.
Method: Transcripts from 50 HCPs and 61 IWAs (from the AphasiaBank database) were used. All participants completed two testing sessions approximately 1 week apart. The AphasiaBank discourse protocol was administered, including a picture scene, two picture sequences, a story retell, and a procedural task. Computerized Language Analysis software was used to score each task for production of previously established core lexicon items. Test–retest reliability was assessed under a generalization theory framework to examine measurement stability, identify the magnitude of differences in core lexicon scores across time for both groups, and examine the contexts for which reliability is sufficient. Finally, minimal detectable change 95 (MDC_{95}) scores are reported for each task and group.
Results: Core lexicon scores demonstrated variable reliability across tasks for HCPs, from poor to good. In contrast, reliability coefficients demonstrated that core lexicon scores had good-to-excellent reliability for IWAs. Averaging multiple stimuli increased reliability, especially for HCPs. There were no substantial practice effects.
Conclusions: Core lexicon scores were sufficiently stable for IWAs to be used in group-level analyses. Additionally, reliability was excellent for longer stimuli or when averaging multiple shorter stimuli for IWA and sufficient for individual use. Using longer stimuli or averaging core lexicon scores across multiple stimuli is recommended when using as a repeated measure for HCPs to track healthy aging and monitor for changes in language ability. Our results, combined with previous research into validity, support the use of core lexicon analysis in clinical practice and research for HCPs and IWAs. Reporting MDC_{95} scores for each task and group facilitates clinical uptake of core lexicon analysis for use with clients as appropriate.
Supplemental Material: https://doi.org/10.23641/asha.34113471

Individuals with aphasia (IWAs) report that discourse-level communication (e.g., engaging in discussions; using language for work, volunteer, or learning; attending religious services) is the desired outcome following treatment (Worrall et al., 2011). Historically, barriers such as time, training, and resources have limited measurement of discourse-level communication in both research and clinical practice (Bryant et al., 2017; Cruice et al., 2020; Stark et al., 2021). Moreover, discourse measures are often excluded from field initiatives to develop core sets of outcome measures (Dietz & Boyle, 2018; Wallace et al., 2019). However, with a shift in recommended treatment outcomes away from discrete language impairments and toward functional communication (Brady et al., 2016), there has been renewed focus on developing discourse analysis tools appropriate for research and clinical use. While several standardized discourse tools with normative data have recently been published (e.g., Berube et al., 2019; Brisebois et al., 2023; Dalton et al., 2024), we lack discourse outcomes that are time efficient and clinically feasible, with well-established psychometric properties necessary for research and clinical use.

Correspondence to Sarah Grace Dalton: sgdalton@uga.edu. **Publisher Note:** This article is part of the Special Issue: Select Papers from the 54th Clinical Aphasiology Conference. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Core Lexicon Analysis

Core lexicon analysis is a standardized, normed discourse measure designed to measure how typical a speaker's word choices are when completing a discourse task, as compared to the word choices of a normative sample of speakers completing the same task (e.g., Dalton & Richardson, 2015). Using Dipper et al.'s (2021) theoretical discourse framework, which distinguishes linguistic, propositional, macrostructural/planning, and pragmatic levels of discourse, core lexicon analysis is situated within the linguistic level of production. Locating the measure at this level is important because different discourse levels are associated with distinct sources of variability and stability. At the linguistic level, performance may vary because of fluctuations in lexical access, semantic activation, retrieval efficiency, word-finding difficulty, and stimulus-specific vocabulary demands while remaining relatively stable as a function of underlying lexical–semantic knowledge and habitual word use within a discourse genre. Core lexicon therefore specifically indexes lexical–semantic selection and word-level informativeness rather than higher order planning or pragmatic organization. As a result, its psychometric properties must be evaluated with respect to the linguistic processes it is intended to capture.

Core lexicon checklists have been established for a number of discourse tasks including picture scenes (Cat Rescue, Cookie Theft), picture sequences (Refused Umbrella, Broken Window, Heungbu and Nolbu), story retells (Cinderella, Good Dog Carl, Picnic), and a procedure (making a peanut butter and jelly sandwich) in English (e.g., Dalton et al., 2020, 2024; Kim et al., 2019), French (Brisebois et al., 2026), Korean (Park et al., 2025), and Mandarin (Chen & Chang, 2024; Jiang et al., 2023). While the exact procedure for developing checklists has varied, all are based on the productions of healthy control participants (HCPs). The most common methods for developing core lexicon include identifying the 25–30 most frequently used lemmas from a normative sample (often separately by part of speech) or using a cutoff of words produced by a proportion of a normative sample (e.g., 50%). Regardless of the specific development procedures, core lexicon checklists identify the key semantic (e.g., for Cinderella: "dress," "fairy," "godmother") and structural (e.g., for Cinderella: "on," "and," "she/her") word-level elements that most speakers would be expected to produce when completing a discourse task.

Core Lexicon Psychometric Properties

To date, core lexicon analysis has demonstrated robust psychometric properties across multiple domains. Known groups validity (or the ability of a measure to differentiate groups when group membership is known) has been demonstrated between IWAs (both as a group and by aphasia subtype—including individuals with latent aphasia) and HCPs (Chen & Chang, 2024; Dalton & Richardson, 2015; Jiang et al., 2023; Kim et al., 2019), between individuals with Alzheimer's disease and HCPs (Fromm et al., 2024), between individuals with mild cognitive impairment and latent aphasia (Stark et al., 2025), between individuals with latent aphasia and all other aphasia subtypes (Dalton & Richardson, 2015), between individuals with Broca's aphasia and all other subtypes (Dalton & Richardson, 2015), and between individuals with fluent and nonfluent aphasia (Kim et al., 2019, 2021).

Concurrent validity, or the relationship between a newly developed measure and previously validated or gold standard measures collected at the same time point, has been demonstrated between core lexicon and other microlinguistic discourse measures such as correct information units (Brisebois et al., 2026; Chen & Chang, 2024) and lexical diversity (Brisebois et al., 2026; Chen & Chang, 2024; Kim & Wright, 2020). Convergent validity, or the relationship between a measure and other related measures, has been demonstrated between core lexicon and microlinguistic and macrolinguistic discourse measures such as main concepts (Brisebois et al., 2026; Dalton & Richardson, 2015); thematic units (Brisebois et al., 2026; Kim & Wright, 2020); and information units, coherence, and syntactic complexity (Kim & Wright, 2020). Evidence of predictive validity (the ability of a measure to predict future performance) was reported in a longitudinal study of IWAs where core lexicon scores, but not other naming measures, were significantly correlated across recovery from the acute to chronic stage (Kim et al., 2023). Finally, multiple studies have reported that both intrarater and interrater reliability is adequate for use in clinical settings (Brisebois et al., 2026; Kim & Wright, 2020).

The well-defined psychometric properties of core lexicon support its validity for use in clinical contexts. However, two fundamental psychometric criteria—short-term test–retest reliability (consistency of scores across multiple administrations over a period in which no true change is anticipated) and internal consistency (the extent to which items measure the same underlying construct)—have yet to be examined for core lexicon analysis. Establishing these properties is essential for determining whether changes in discourse performance reflect true change in lexical–semantic production rather than measurement variability. Reliable discourse indices are critical for assessment planning because they allow clinicians and researchers to select measures that can be meaningfully interpreted across sessions, support longitudinal monitoring, and inform treatment efficacy. In practice, evaluating these properties can be challenging, as validity and reliability are not static; using a tool in settings

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

or ways that differ from those described in norming studies may alter its psychometric behavior. Nevertheless, reporting test–retest reliability and internal consistency provides future users with confidence that the tool can be applied as described in the literature and allows them to judge whether adaptations required by their clinical context are likely to compromise interpretability. By defining these psychometric properties, users gain clearer guidance on how and when core lexicon analysis can be used appropriately. For example, while a tool with poor test–retest reliability may still be useful for diagnostic description or treatment planning, it should not be used to track recovery or response to intervention over time.

Clinicians may shorten discourse samples, omit certain elicitation prompts, or substitute alternative stimuli in response to time constraints, client fatigue, or communicative breakdowns. This concern is directly reflected in the discourse elicitation guidance of Dutta et al. (2025), who note that spoken discourse can be elicited through multiple structured and unstructured tasks across genres, that language performance differs depending on genre and task, and that task selection should be guided by the individual's needs and the intended assessment purpose. They further emphasize that structured tasks are often chosen when clinicians require greater control and a standardized method for repeatedly eliciting comparable samples over time, whereas different tasks impose different cognitive–linguistic demands and may yield different discourse outputs. Accordingly, modifications to elicitation procedures are unlikely to be psychometrically neutral and may alter the stability or interpretability of core lexicon scores relative to published norms. Explicitly defining reliability and related psychometric properties therefore allows clinicians to better judge whether such deviations are likely to meaningfully affect assessment validity or the interpretation of change over time. The present study addresses this issue directly by estimating core lexicon reliability within individual stimulus, across stimuli, and as a function of the number of stimuli administered in a decision study, thereby providing practical guidance for clinicians who may need to reduce sampling burden while still obtaining interpretable core lexicon scores.

Current Study

In the current study, we evaluate the test–retest reliability of core lexicon analysis in IWAs and HCPs across several prototypical monologic discourse tasks. We examined test–retest reliability both within individual discourse stimuli and across stimuli, evaluated the internal consistency of the core lexicon battery, and reported minimal detectable change (MDC) scores by group. Together, these analyses determine whether core lexicon analysis provides a stable and interpretable index of lexical–semantic production suitable for longitudinal monitoring and outcome measurement.

Method

Data and Transcripts

Transcripts from the NEURAL (Stark et al., 2023) and NEURAL-2 (Stark et al., 2025) corpora of the AphasiaBank database were retrieved for analysis. Ethical approval was received for the study (Institutional Review Board Number 14633 at Indiana University), and all participants provided informed consent. Exhaustive details about the procedures of each study can be found in each of the respective cited papers. Briefly, NEURAL and NEURAL-2 were conceived as short duration, no intervention test–retest designs, where each IWA and HCP was enrolled in two virtual testing sessions that occurred approximately 1 week apart, during which no other interventions or procedures occurred. In NEURAL, there was a single experimenter who performed all procedures; in NEURAL-2, two project managers performed all procedures. NEURAL-2 was intentionally designed to enroll individuals with mildest aphasia, whereas NEURAL was not designed with such a purpose. As part of a larger study, all participants completed the AphasiaBank elicitation protocol (MacWhinney et al., 2011), and all IWAs completed the Western Aphasia Battery–Bedside (WAB-B) to estimate aphasia severity. For all participants in NEURAL-2 and the majority of NEURAL, the AphasiaBank tasks were administered in the same order: Window, Umbrella, Cat, Cinderella, Sandwich. In NEURAL-2, an additional narrative task (which is not evaluated here because it does not have an established core lexicon) was administered after Cinderella and before Sandwich. In NEURAL, the earliest enrolled participants had tasks administered out of order, which was due to an experimenter error. For more specific information about each protocol, the reader can navigate to their respective corpora on AphasiaBank (NEURAL: https://talkbank.org/aphasia/access/English/Protocol/NEURAL.html; NEURAL-2: https://talkbank.org/aphasia/access/English/Protocol/NEURAL-2.html).

Participants

Across the two corpora, 61 IWAs, weighted toward individuals with latent and mild aphasia (average WAB-B aphasia quotient: 87.2; *SD* = 16.3), and 50 HCPs were included (see Table 1 for complete demographic data). The average age of participants across both groups was approximately 60 years, and both groups were highly educated. While the NEURAL-2 database now consists of a higher number of IWAs and HCPs, data for only some of the participants were available at the time of these analyses.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 1. Demographic data represented as mean (standard deviation) for healthy control participants and individuals with aphasia.

Variable	HCP	IWA
Age (years)	62.9 (9.3)	59.8 (12.2)
Sex	36 female	28 female
	17 male	35 male
Education (years)	17.2 (3.1)	16.8 (3.2)
Aphasia type	—	24 latent 21 anomic 5 conduction 1 Wernicke's 8 Broca's 3 transcortical motor 1 unclassified
WAB-B AQ	—	87.2 (16.3)

Note. Em dashes indicate data not applicable for the participant group. HCP = healthy control participant; IWA = individual with aphasia; WAB-B AQ = Western Aphasia Battery–Bedside Aphasia Quotient.

Core Lexicon Scoring

Previously developed core lexicon lists for five AphasiaBank discourse tasks were used, including the "Cinderella" story retell, the "Cat Rescue" picture scene, the "Broken Window" and "Refused Umbrella" picture sequences, and the "Sandwich" procedure (Dalton et al., 2020; Dalton & Richardson, 2015). The CORELEX command in the Computerized Language Analysis (CLAN; v02Jan25) software was used to analyze core lexicon production for each discourse task and visit. Development and validation of the CORELEX command have been reported (Dalton et al., 2022), with near-perfect reliability demonstrated between automated and manual scoring procedures across the structured AphasiaBank tasks (all intraclass correlations > .975). The current study followed the step-by-step procedures reported by Dalton et al. (2022), which are also available on the Core Lexicon webpage under the Discourse Analysis section of the AphasiaBank website (see Table 2 for specific commands).

Statistical Analysis

After core lexicon scores were extracted, we examined test–retest reliability under a generalizability theory (G-theory) framework (Brennan, 2001; ten Hove et al., 2022). We used G-theory for the following reasons: (a) Unlike the classical test theory, which combines all sources of measurement error into a single term, G-theory partitions variance into distinct sources. This partitioning reveals how much each facet of the measurement design (e.g., items, occasions, raters) contributes to measurement error, necessary for disentangling time point and stimuli effects; (b) G-theory allows us to tailor estimates of reliability to the universe of generalization; that is, the specific conditions and contexts we want our reliability estimates to apply, which is particularly useful for discourse assessments that are often modified for different research and clinical contexts; and (c) once variance components are estimated, decision studies can be used to estimate how changes to the measurement parameters (e.g., changing the number of discourse stimuli) would affect reliability under different use cases, such as deciding how many stimuli to use in an assessment. G-theory distinguishes between two types of reliability: Dependability coefficients describe absolute reliability (often called "agreement"—the extent to which scores match exactly), and generalizability coefficients describe relative reliability (often called "consistency"—whether scores preserve rank ordering across individuals). These coefficients are analogous to intraclass correlation coefficients (ICCs) in the classical test theory and can be interpreted according to guidelines suggested by Koo and Li (2016) where ICC < .5 is poor, ICC between .5 and .75 is moderate, ICC between .75 and .9 is good, and ICC > .9 is excellent, although it is also recommended that the context of future use of the measure should dictate the level of reliability required. For example, clinical decisions and useful MDC estimates often require ICC > .90, while reliability requirements for outcome measures used at the group level are less restrictive. Moreover, while we have focused on the point estimate of reliability, more conservative approaches suggest evaluating the lower tail of the 95% confidence interval (CI) to ensure that reliability estimates will exceed desirable thresholds.

Because core lexicon represents a proportion of words produced in a sample and some ceiling effects were evident in the healthy control participant (HCP) group, logistic mixed-effects models with a binomial distribution

Table 2. Computerized Language Analysis (CLAN) commands used to calculate core lexicon scores.

CLAN command	Function
chstring +q1 *.cha	Edits transcripts to disallow clinician-identified target words and allow revisions to be scored
mor *.chstr.cex	Reruns the mor grammar to identify parts of speech
gem +f*Story* + s*Story* + n + d1 + t*PAR + t%mor *.cex	Extracts the portion of the transcript with the relevant discourse task
Corelex +t*PAR + l*Story* *.cex	Calculates core lexicon production and overall score

Note. Italics indicate the portions of commands that must be changed to output data for a specific task.
*CLAN v02Jan25.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

were used to fit the data. To estimate dependability and generalizability coefficients, residual variance and variance components were extracted from the models following the observation-level methods reported in Nakagawa et al. (2017) as implemented in the insight R package (Lüdecke et al., 2019). The dispersion ratios for each model were approximately 1 (all $p > .05$), indicating that no models were over- or underdispersed, and the binomial model's variance assumptions were met. For all coefficients, 95% CIs were estimated via clustered bootstrapping with 1,000 iterations, requiring > 90% convergence rate. Data analysis was conducted using R (Version 4.5.1), including the lme4 (Bates et al., 2015) and tidyverse (Wickham et al., 2019) suite of packages. Code and de-identified data are available on the Open Science Framework at https://osf.io/3mwzv/. A summary of all statistical models is reported in Table 3.

Stimuli-Level Test–Retest Reliability

To examine the agreement of core lexicon for each of the five stimuli and for IWAs and healthy controls, we estimated the dependability coefficient taking the variance components σ^2 person / (σ^2 person + σ^2 time point + σ^2 residual) from a model with random intercepts for person and time point. Because the two-time-point design prevents separating systematic time point effects from the Person × Time Point interaction (typically required for a generalizability coefficient), we estimated an analogous generalizability coefficient by modeling time point as a fixed effect (see Supplemental Material S1), thereby accounting for systematic practice effects and returning an estimate of consistency.

Across-Stimuli Test–Retest Dependability

Second, we examined the dependability of core lexicon measures across stimuli for IWAs and HCPs using separate mixed-effects models. These models included random intercepts for person, stimuli, and time, along with Person × Stimuli, Person × Time, and Stimuli × Time interaction terms, allowing us to account for variability attributable to stimulus characteristics, measurement occasions, and differential performance patterns across individuals and time points. In this framework, variance associated with the Person × Stimuli interaction indexes the internal consistency of the core lexicon stimulus set, reflecting the extent to which the ranking of individuals' performance varies across different discourse stimuli.

Evaluating the Effects of Group on Dependability

Third, we evaluated the impact of group membership on dependability by fitting a combined model that included both groups and all stimuli, estimating reliability with and without group specified as a fixed effect. This approach allowed us to assess the extent to which variance attributable to group membership influenced the resulting variance components and, in turn, the dependability coefficient. This analysis was motivated by visual inspection of the data, which suggested comparable error around the trend lines for both groups, alongside a substantially narrower score range among HCPs. Because range restriction is known to attenuate reliability estimates even when absolute measurement error is similar, we examined whether controlling for group-related variance altered dependability estimates. Minimal change in the dependability coefficient would indicate that the groups can be reasonably collapsed for the purpose of estimating dependability.

Estimating the Effects of Averaging Stimuli Scores (Decision Study)

Because the reliability of each individual stimulus ranged from modest to excellent, we sought to understand how increasing the number of stimuli used to estimate core lexicon improved reliability. Understanding how reliability

Table 3. Dependability and generalizability coefficient definitions, descriptions, and variance formulas for each research aim.

Analysis	Focus of generalization	Fixed effects	Variance formula
Stimuli-level test–retest reliability	Dependability: estimating the absolute agreement across time points for each stimulus (i.e., agreement)	None	$\sigma^2_p / (\sigma^2_p + \sigma^2_t + \sigma^2_r)$
	Generalizability: estimating the relative agreement across time points for each stimulus (i.e., consistency)	Time point	$\sigma^2_p / (\sigma^2_p + \sigma^2_r)$
Across-stimuli reliability and internal consistency	Estimating the dependability across both stimuli and time points; estimating core lexicon internal consistency (σ^2_{pxs})	None	$\sigma^2_p / (\sigma^2_p + \sigma^2_s + \sigma^2_{pxs} + \sigma^2_t + \sigma^2_{txp} + \sigma^2_{txs} + \sigma^2_r)$
Group reliability	Impact of group (IWA vs. HCP) heterogeneity on dependability estimation	None vs. group	$\sigma^2_p / (\sigma^2_p + \sigma^2_s + \sigma^2_{pxs} + \sigma^2_t + \sigma^2_{txp} + \sigma^2_{txs} + \sigma^2_r)$
Decision study	Estimating the number of stimuli required to achieve adequate dependability	None	$\sigma^2_p / (\sigma^2_p + \sigma^2_s/n + \sigma^2_{pxs}/n + \sigma^2_t + \sigma^2_{txp} + \sigma^2_{txs}/n + \sigma^2_r)$

Note. Dependability is analogous to agreement; generalizability is analogous to consistency. Sigma = variance component; σ^2_p = participant; σ^2_s = stimuli; σ^2_{pxs} = Participant × Stimuli interaction; σ^2_t = time point; σ^2_{txp} = Time Point × Participant interaction; σ^2_{txs} = Time Point × Stimuli interaction; σ^2_r = residual. n = the number of stimuli averaged. IWA = individual with aphasia; HCP = healthy control participant.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

changes as a function of the number of stimuli used will provide guidance to both researchers and clinicians about how many stimuli may be required to obtain stable estimates of core lexicon for both IWAs and healthy controls. This is accomplished through a decision study in G-theory, in which reliability is evaluated across different permutations of the study design. Here, we estimated dependability using any one stimuli or an average of two, three, four, or all five possible stimuli. Since averaging random subsets of the five stimuli is an unlikely use case, we omitted the Participant × Stimuli interaction. This approach is therefore agnostic to which stimuli are being averaged; it reflects the effects of averaging "*n*" stimuli from any stimuli that are similar to those included in this study.

MDC

Finally, we estimated the MDC 95 (MDC_{95}) for each individual stimulus for IWAs and HCPs separately. For each stimulus, we calculated the standard error of measurement from the dependability coefficient using the standard formula ($SEM = \sqrt{(\text{variance} \times (1 - \text{dependability}))}$) and then derived MDC_{95} as $1.96 \times \sqrt{2} \times SEM$. Because of the nonlinear logistic transformation, a constant MDC on the logit scale corresponds to different absolute changes in raw scores depending on baseline performance. Therefore, we stratified baseline performance into ranges of approximately 5 points and calculated MDC_{95} values for each stratum using the baseline score closest to the scale midpoint within each range (to provide conservative estimates, particularly at the distribution tails). These stratified MDC_{95} values therefore represent the minimum number of items of improvement necessary to conclude with 95% confidence that an observed change exceeds measurement error.

Results

Stimuli-Level Test–Retest Reliability

The relationship between performance at Time 1 (test) and Time 2 (retest) is visualized in Figure 1. For IWAs, dependability estimates (the extent to which scores match exactly) ranged from .83 to .96, with all stimuli demonstrating good-to-excellent reliability according to established guidelines (Koo & Li, 2016). The Cinderella stimulus showed the highest reliability (.96, 95% CI [.93, .98]), followed by Umbrella (.88, 95% CI [.78, .92]), Cat (.88, 95% CI [.81, .92]), Sandwich (.85, 95% CI [.77, .89]), and Window (.83, 95% CI [.72, .88]).

In contrast, healthy controls showed more variable reliability across stimuli, with coefficients ranging from .49 to .88. The Cinderella stimulus again demonstrated the highest reliability (.88, 95% CI [.79, .93]), while Window showed the poorest reliability (.49, 95% CI [.35, .57]). The remaining stimuli showed moderate reliability: Cat (.64, 95% CI [.46, .74]), Sandwich (.59, 95% CI [.39, .69]), and Umbrella (.56, 95% CI [.39, .65])

Generalizability coefficients (whether scores preserve rank ordering across individuals) were largely unchanged from the dependability coefficients, indicating minimal shift in scores between time points. Consistent with this result, there was only a significant effect of time point for the Cinderella task for IWAs, and it was very small ($b = 0.12$, 95% CI [0.04, 0.20], $p = .004$). This coefficient was approximately 60% of the standard error of measurement (0.2), suggesting that, while statistically significant, the size of this practice effect is practically negligible. Taken together, the findings suggest that core lexicon performance is sufficiently stable across sessions without evidence of meaningful practice effects. All stimuli-level reliability coefficients are reported in Table 4.

Across-Stimuli Test–Retest Reliability

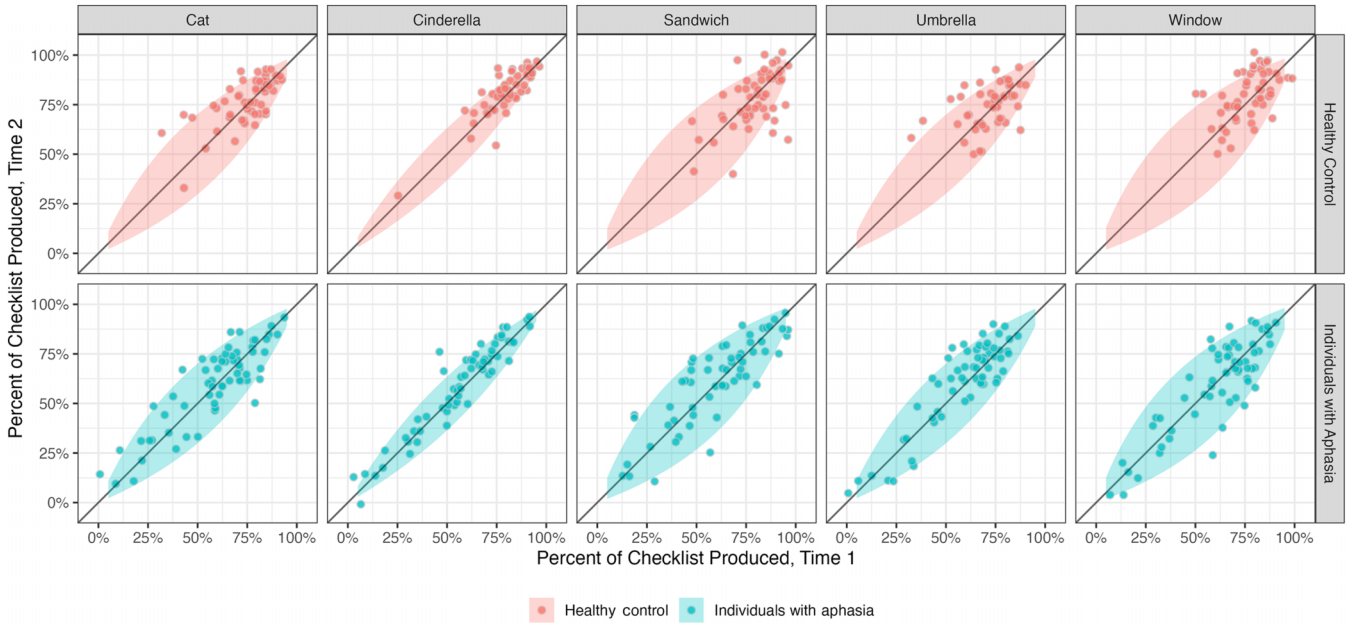
To examine the reliability of core lexicon as a general measure incorporating variability across different stimuli, we calculated dependability coefficients using the formula σ^2participant / (σ^2participant + σ^2stimuli + σ^2stimuli:participant + σ^2time + σ^2time:participant + σ^2time:stimuli + σ^2residual. For IWAs, the overall dependability coefficient was .82 (95% CI [.73, .86]), indicating good-to-moderate reliability even when accounting for stimulus variability. In contrast, HCPs demonstrated substantially lower overall reliability, with a dependability coefficient of .37 (95% CI [.24, .47]), falling well below conventional thresholds for acceptable reliability.

For IWAs, variance component analysis revealed that participant-level variance accounted for 81.69% of total variance, with stimuli contributing minimal variance (0.26%), Stimuli × Participant interactions contributing 8.35%, time contributing 0.23%, Time × Participant interactions contributing 0.91%, Time × Stimuli interactions contributing 0.01%, and residual error accounting for 8.55%. The relatively modest Stimuli × Participant interaction variance suggests good internal consistency, indicating that IWAs tend to perform consistently across different core lexicon stimuli but that some participants do better on specific stimuli.

For healthy controls, variance components showed participant-level variance accounting for 36.51% of total variance, with stimuli contributing 4.59%, Stimuli × Participant interactions contributing 28.89%, time contributing 0.54%, Time × Participant interactions contributing 1.35%, Time × Stimuli interactions contributing 0.00%, and residual

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 1. Relationship between core lexicon scores at Time 1 versus Time 2 for individuals with aphasia and healthy controls. Shaded areas represent 95% prediction intervals: Points within the shaded areas represent differences that do not exceed measurement error, while points outside the shaded areas represent changes beyond what is expected from measurement error alone.



error accounting for 28.11%. The substantially larger Stimuli × Participant interaction variance provides a clear reason for the large reduction in reliability between IWAs and HCPs when evaluating reliability across stimuli: The relative ranks of scores shift substantially between stimuli for HCP, and there is poorer internal consistency across this battery of core lexicon stimuli.

When analyzing all participants together (collapsed across groups), the overall reliability was good (dependability coefficient: .76, 95% CI [.67, .82]). When group differences were statistically controlled by including diagnosis as a fixed effect, reliability decreased to .71 (95% CI [.6, .78]). This modest 5-percentage-point decrease suggests that the majority of between-person variance in the collapsed model largely reflects individual differences within groups rather than systematic differences between groups. However, controlling for group (IWAs vs. HCPs) does reduce between-person variance to a degree.

Estimating the Effects of Averaging Stimuli Scores (Decision Study)

We conducted a decision study to examine the effects of averaging scores across stimuli. Results are visualized in Figure 2. For IWAs, single-stimulus reliability was good

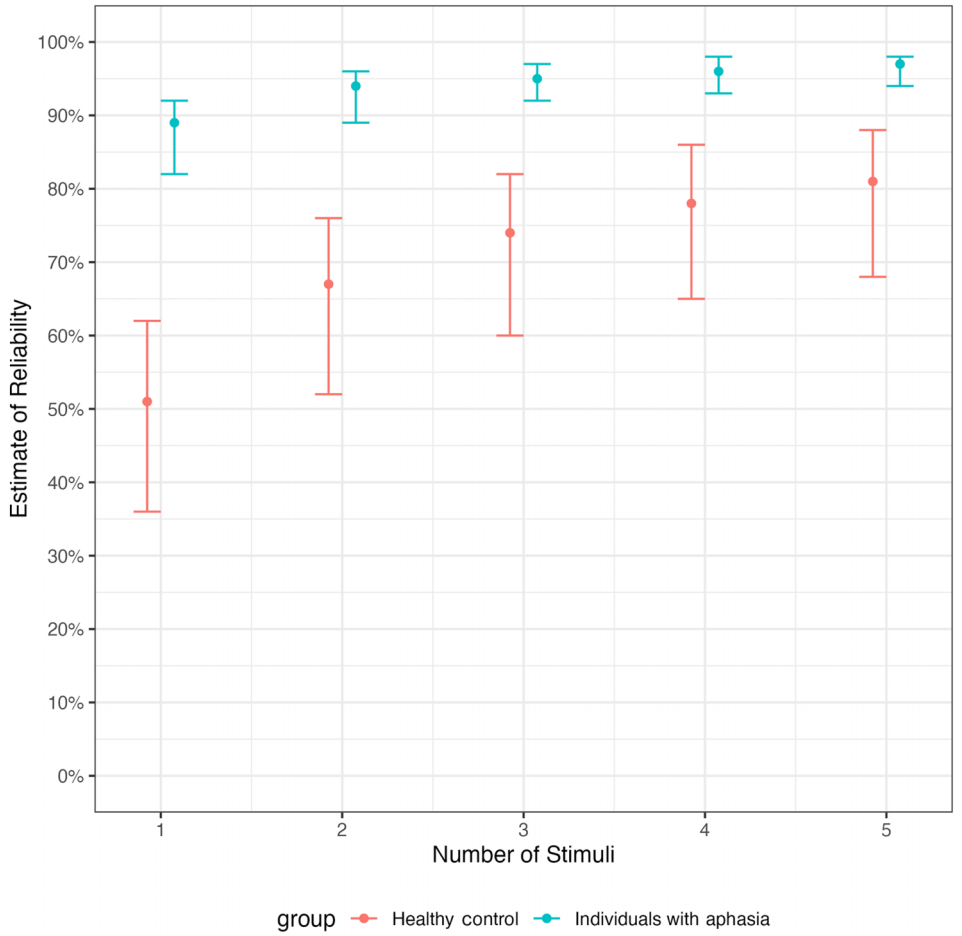
Table 4. Estimates of dependability and generalizability by stimulus for individuals with aphasia (IWAs) and healthy controls.

Stimuli	Group	Dependability	95% CI	Generalizability	95% CI
Cat	Control	.64	[.46, .74]	.65	[.47, .74]
Cat	IWA	.88	[.81, .92]	.88	[.81, .92]
Cinderella	Control	.88	[.79, .93]	.89	[.8, .93]
Cinderella	IWA	.96	[.93, .98]	.97	[.94, .98]
Sandwich	Control	.59	[.39, .69]	.60	[.42, .69]
Sandwich	IWA	.85	[.77, .89]	.85	[.77, .89]
Umbrella	Control	.56	[.39, .65]	.56	[.4, .66]
Umbrella	IWA	.88	[.78, .92]	.88	[.78, .92]
Window	Control	.49	[.35, .57]	.49	[.34, .58]
Window	IWA	.83	[.72, .88]	.83	[.71, .89]

Note. Dependability coefficients are analogous to single-rater agreement intraclass correlation coefficients, and generalizability coefficients are analogous to single-rater consistency intraclass correlation coefficients. CI = confidence interval.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 2. Effects of averaging additional stimuli on core lexicon dependability estimates.



(.89, 95% CI [.82, .92]) and increased modestly with two or more stimuli (.94, 95% CI [.89, .96]). In contrast, healthy controls showed substantial improvements in reliability from additional stimuli, with moderate-to-poor single-stimulus reliability (.51, 95% CI [.36, .62]) and obtaining good reliability only when averaging across four stimuli (.78 95% CI [.65, .86]). These findings suggest that, while single stimuli may be sufficient for reliable measurement in IWAs, assessments of healthy controls may require multiple stimuli to achieve adequate reliability for clinical or research purposes.[1]

MDC

To provide statistical benchmarks for interpreting individual change scores, we calculated MDC_{95} values for each stimulus within each group by baseline performance. Due to the binomial modeling approach and logistic link function, MDC_{95} values vary as a function of baseline performance, with larger absolute changes required in the middle of the scale and smaller changes at the extremes. This nonlinear effect is advantageous, as a 10-item improvement for someone with 0/35 items correct at baseline is likely to be considered having improved more than an individual who improved from 10/35 to 20/35 items. For stimulus-specific administration (using the same stimulus across time points), MDC_{95} values for IWAs ranged from four to 14 items depending on the stimulus and baseline performance level, representing 3.6%–25.1% of the maximum possible score. For example, when administering the Cat stimulus to an IWA, a patient who initially produced zero to five items would need to improve by at least five items to exceed measurement error, whereas a patient starting at 11–15 items would need to improve by eight items. These thresholds represent the minimum change that can be confidently attributed to true change rather than measurement variability and are essential for determining whether an individual patient has made clinically meaningful progress. A full

[1]Note that these results apply to the choice of how many stimuli to choose in general, but individual stimuli reliability coefficients still apply for decisions about specific stimuli.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

table of MDC_{95} thresholds for a range of baseline scores for IWA and HCP for all stimuli is reported in Supplemental Material S2. Healthy controls showed similar MDC_{95} values, ranging from four to 13 items (3.4%–25.5% of maximum scores). Although healthy controls showed similar absolute MDC_{95} values to IWAs, this similarity reflects offsetting effects of poorer reliability (lower ICC) and smaller standard deviations due to restricted score variability; thus, the comparable MDC_{95} thresholds mask underlying differences in measurement precision between groups.

Discussion

This study examined the test–retest reliability of core lexicon measures using G-theory, providing a nuanced view of measurement precision. To our knowledge, this is the first report of test–retest reliability and score consistency for core lexicon analysis. We evaluated reliability within individual stimulus, compared agreement versus consistency estimates, assessed overall reliability across multiple stimuli, and established MDC_{95} values for clinical interpretation.

The results showed that core lexicon demonstrated consistently strong reliability within each stimulus for IWAs, with coefficients ranging from .83 to .96. In contrast, reliability for HCPs was more variable and lower, ranging from .49 to .88. Differences between dependability (i.e., agreement) and generalizability (i.e., consistency) within stimuli were minimal. When examining reliability across all stimuli, the IWA group showed good internal consistency, as indicated by modest variance in the Stimuli × Participant interaction (8.4%). In contrast, the HCP group showed poor internal consistency, reflected in a low dependability coefficient (.37) and a substantially higher interaction variance (29%). Finally, the decision study indicated that a single stimulus was sufficient to achieve good reliability for IWA (.89), whereas for HCP, data from four stimuli were necessary to reach good reliability levels (.78). These findings, combined with prior validity and reliability research (Brisebois et al., 2026; Chen & Chang, 2024; Dalton et al., 2024; Dalton & Richardson, 2015; Jiang et al., 2023; Kim et al., 2019, 2021, 2023; Kim & Wright, 2020; Park et al., 2025), support the clinical and research use of core lexicon for IWA, while the lower stability for HCP suggests that it may be less suitable for tracking healthy aging, unless multiple stimuli are used.

Our results extend prior work on discourse reliability. Brookshire and Nicholas (1994) showed that the stability of connected speech measures (words per minute [WPM], percent correct information units [%CIUs], or CIUs/min) improved as discourse stimuli number increased, with the greatest gains in test–retest reliability between a sample elicited from one stimulus (Boston Diagnostic Aphasia Examination Cookie Theft) and a sample elicited from five stimuli (two picture scenes, a picture sequence, a procedure, and a personal narrative). However, even with the five-stimuli sample, HCP in their study showed more variable performance on discourse measures than IWA—a finding echoed in our results for HCP here, where longer, lexically rich tasks (e.g., Cinderella) and averaging across stimuli yielded stronger core lexicon score reliability. Boyle (2014) similarly demonstrated that, when averaging across the same five stimuli investigated by Brookshire and Nicholas, several word-retrieval measures were stable enough for group-level research (Pearson product–moment correlations above .7), although only a subset (e.g., WPM, CIUs/min) was sufficiently reliable for individual-level clinical decisions (Pearson product–moment correlations above .9). A further investigation of the four picture-based AphasiaBank tasks similarly concluded that neither the two picture sequences combined nor the picture scene or Cinderella retell yielded samples with sufficient stability to be used alone when investigating word retrieval errors. However, when all four stimuli were combined, the sample was sufficiently stable to measure some word-retrieval errors for group-level research or individual clinical decisions (Boyle, 2015).

In addition, our recent work has shown that both sample length (measured as number of tokens) and aphasia severity contribute to variability in test–retest reliability; however, neither accounts for a large proportion of the observed effects, and results are measure dependent (Stark et al., 2023). In some cases, longer samples were associated with higher reliability (e.g., %CIUs in aphasia), whereas in others, particularly in HCPs, shorter samples yielded comparable or even higher reliability for certain metrics (e.g., mean length of utterance [MLU], verbs per utterance). These effects were also task dependent, with some stimuli (e.g., Cinderella narratives) showing stronger length-related effects than others (e.g., picture description), underscoring that reliability cannot be fully explained by a monotonic relationship to sample length. Given the lack of a clear pattern for sample length, additional research is needed to understand what other variables (e.g., elicitation methods, analysis methods) may be driving these findings. However, that is outside the scope of the current study, as we did not include sample length in our analysis.

There are some critical differences between the current study and the above research. First, participants ranged from 10–12 IWAs (Boyle, 2014, 2015) to 20 IWAs and HCPs (Brookshire & Nicholas, 1994), to 23 IWAs and 24 HCPs (Stark et al., 2023) in previous research, while the current study included 61 IWAs and 50 HCPs. This larger participant sample may provide a more accurate estimate of test–retest reliability. Additionally, while Stark et al. (2023) used a statistical analysis well suited for investigating test–retest reliability, other work has reported mean change

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

scores and Pearson product–moment correlations (Boyle, 2014, 2015), or change scores and "correlation coefficients," without specifying the type (Brookshire & Nicholas, 1994). It is likely that neither analysis was optimized to directly examine the magnitude of change over time needed for short-interval test–retest reliability. Finally, the characteristics of the discourse measures may have impacted our findings. All of the measures previously investigated (except the percent accurate and complete main concepts reported by Boyle, 2014) are "open-set" measures where there is no maximum score. In contrast, core lexicon analysis is a "closed-set" measure consisting of specific items as determined by a normative sample. While this development method comes with some inherent limitations, it may also improve test–retest reliability by narrowing the focus of measurement.

Importantly, reliability reflects both within-speaker consistency across sessions and the degree of between-participant variability captured by a task. Shorter tasks such as Sandwich and Window may therefore show lower reliability for different reasons: Sandwich elicits a highly constrained and predictable set of lexical items, which may limit between-speaker variability and attenuate reliability estimates despite relatively stable within-speaker performance, whereas Window allows greater flexibility in event framing and lexical choice, potentially increasing within-speaker variability across sessions and reducing consistency. In contrast, Cinderella appears to promote higher reliability because it combines strong within-speaker consistency with sufficient between-speaker variability. As a well-known narrative with a canonical event structure, it reliably cues a stable core lexicon and temporal framework across sessions while still permitting meaningful individual differences in narrative detail and style. Although Cinderella also tends to elicit longer samples, prior findings indicate that its higher reliability is better explained by narrative scaffolding and preserved variance than by sample length alone.

Stark et al. (2023) additionally showed that lexical and informativeness measures (e.g., CIUs, tokens, type–token ratio) demonstrate higher stability across tasks and time points than syntactic or grammatical indices (e.g., MLU, noun/verb ratio, open class/closed class word ratio). Our findings corroborate these insights: Core lexicon, as a lexical measure, showed consistently high stability in IWA, paralleling the reliability of CIUs and other lexical metrics. Taken together, these studies chart a trajectory in discourse measurement: early evidence on the role of stimuli number (Brookshire & Nicholas, 1994), recognition of measure-specific stability and the need for MDC thresholds (Boyle, 2014), broad evidence favoring lexical over structural measures (Boyle, 2014, 2015; Stark et al., 2023), and the variable effects of sample length and aphasia severity on reliability across tasks (Stark et al., 2023). By establishing test–retest reliability and conducting the decision study, our work builds on prior research by demonstrating the potential for core lexicon analysis to measure true change, rather than measurement error, over time and by task. Within this trajectory, the present findings confirm core lexicon as a stable discourse index in short test–retest windows and provide the psychometric detail necessary for clinical translation.

Reliability and Internal Consistency

At the level of individual discourse stimuli, core lexicon analysis demonstrated consistently good-to-excellent test–retest reliability for IWA. This pattern dovetails with the broader movement toward standardized discourse elicitation and scoring protocols, such as AphasiaBank and CLAN, which have been argued to enable stable measurement of fluency and lexical discourse behaviors across repeated samples (Stark et al., 2023). These findings also align with prior work showing that other context-relevant lexical behaviors, such as correct information units, demonstrate high stability across administrations (Boyle, 2014; Stark et al., 2023). In contrast, grammatical and syntactic indices have often shown weaker reliability across both IWAs and HCPs (Stark et al., 2023). Thus, the present results reinforce the conclusion that lexical-level discourse indices, including core lexicon, provide a more reliable signal for repeated measurement than structural measures of language.

Stimulus characteristics also appeared to influence reliability. Longer and more lexically rich tasks, such as Cinderella, produced the strongest reliability estimates, consistent with evidence that greater language sample "mass" is associated with greater stability (Brookshire & Nicholas, 1994; Stark et al., 2023). Our decision study further supports this mass-stability principle, showing that averaging across multiple stimuli improved core lexicon reliability, particularly for HCP. This echoes recent recommendations for optimizing discourse protocols by increasing sample length and aggregating across tasks to minimize measurement error (Marini et al., 2011; Stark, 2019; Stark & Fukuyama, 2021). Importantly, this finding has group-specific implications: While core lexicon demonstrated high reliability with even a single stimulus for IWA, core lexicon scores had lower single-task stability for HCP. If discourse analysis is to be used as a sensitive metric for detecting age-related decline in HCP, administration of longer samples and/or multiple discourse tasks may be necessary to achieve reliable measurement.

When considering reliability across multiple discourse stimuli, core lexicon had high internal consistency for IWA. Variance decomposition revealed low Stimuli × Participant interaction variance, suggesting that core lexicon performance in aphasia, as a lexical–semantic discourse measure,

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

is relatively stable across the discourse contexts examined here rather than shifting substantially with task or genre. This should not be interpreted to mean that discourse production in IWA is generally task invariant. Rather, it suggests that core lexicon may capture a relatively stable component of aphasia-related lexical access difficulty across elicitation contexts, supporting its use as a generalizable index of lexical retrieval impairment. In contrast, core lexicon had lower internal consistency for HCP, with larger Stimuli × Participant interaction variance indicating greater cross-stimulus inconsistency. This pattern suggests that core lexicon performance in neurotypical speakers may be more sensitive to task demands and genre, although an alternative explanation—that HCP performance reflects ceiling effects and restricted score variance rather than true group differences in internal consistency—cannot be ruled out. Regardless of the underlying mechanism, these findings remain compatible with previous work showing that discourse performance more broadly varies with elicitation task in both healthy speakers and IWA (see Dutta et al., 2025, for a review), because such task effects may be stronger for other discourse dimensions than for core lexicon specifically.

Our finding that reliability fell from .76 to .71 when controlling for the diagnostic group suggests that a small, but nontrivial portion of between-person variance is attributable to group membership rather than individual differences within groups and that diagnostic group affects the variance structure underlying reliability estimates. In other words, the presence or absence of aphasia is associated with different proportions of person, stimulus, and error variance, resulting in different reliability coefficients for core lexicon (as evidenced by the breakdown of variance components in the separate models). This distinction underscores the importance of evaluating psychometric properties separately in clinical and nonclinical groups, rather than assuming that reliability estimates will generalize across populations. More broadly, it highlights that validity and reliability are not merely static traits of an instrument but are also sample dependent. Thus, applying core lexicon in contexts that differ substantially from those in which its properties were established—for example, in progressive neurodegenerative populations, in non-English speakers, or with tasks beyond the AphasiaBank discourse set—may yield very different psychometric behavior. This aligns with longstanding cautions in psychometric theory: A tool validated in one population or setting may not retain the same reliability or validity when transported wholesale into another (Cicchetti, 1994; Koo & Li, 2016).

Clinical Implications

The present study provides stimulus-specific MDC_{95} thresholds that can be directly applied to clinical practice. For IWA, MDC_{95} values ranged from 4 to 14 points or approximately 3%–25% of maximum scores, offering usable cut-points for interpreting longitudinal changes that take into account baseline performance (see Supplemental Material S2 for stimuli-specific MDC_{95} scores and guidelines for use). Changes larger than these thresholds can therefore be interpreted as true improvement rather than measurement error, aligning with the Cochrane-driven shift toward functional communication outcomes and repeated-measure designs sensitive to real-world change. Importantly, we found minimal evidence of practice effects in IWA, supporting the repeated administration of core lexicon as a stable outcome measure for monitoring recovery and treatment response. This psychometric property complements the validity evidence already demonstrated for core lexicon, including convergent, known-groups, and criterion validity (Dalton & Richardson, 2015; Kim et al., 2019; Kim & Wright, 2020), strengthening confidence in its clinical applicability.

In contrast, flexible administration that substitutes different stimuli across time points requires substantially larger relative changes to surpass MDC_{95}, cautioning against stimulus substitution in longitudinal designs unless averaging is applied. Consistent with our stimulus-level analyses, longer and richer tasks such as Cinderella provided the most reliable measurement and may be sufficient on their own for IWAs. When longer tasks such as Cinderella are not feasible, clinicians and researchers should consider averaging at least two of the shorter stimuli to ensure sufficient reliability when using core lexicon with IWA. Averaging four or more stimuli may be necessary for reliable estimates of core lexicon with HCPs. We do note one limitation here: Estimates of reliability when averaging more than one stimuli assumes exchangeability between the stimuli used in this study—which is unlikely to fully hold given the differences in stimuli genre and length and variation in stimuli-specific reliability. Therefore, any specific combination of two or three stimuli will deviate somewhat from these average estimates depending on which stimuli are selected.

Historically, discourse was excluded from the Research Outcome Measurement in Aphasia consensus of standard core aphasia outcome measures (Wallace et al., 2019) in part because of limited psychometric evidence (e.g., Pritchard et al., 2017). Scholars have advocated for the creation of a discourse-specific core outcome set (Dietz & Boyle, 2018). By quantifying test–retest reliability and MDC_{95}, our study helps address this gap and aligns with ongoing efforts to standardize elicitation protocols and report psychometric properties necessary for clinical adoption (Boyle, 2014, 2015; Stark et al., 2023). The take-home message parallels prior work: Psychometric properties of discourse-related variables must be estimated at the group level, particularly for IWA, to ensure valid interpretation.

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Finally, while core lexicon can be scored automatically using the CLAN software as in this study, it may also be evaluated perceptually (e.g., as a checklist), facilitating its use in routine clinical practice (Dalton, Hubbard, & Richardson, 2020; Kim et al., 2019 Richardson & Dalton, 2020). Recent evidence demonstrated good-to-excellent reliability of in-session scoring compared to traditional transcript scoring on a similar, checklist-based measure of content word fluency (Alyahya, 2025), although this should be confirmed directly with core lexicon analysis. By providing reliability benchmarks alongside previously established validity evidence, this study brings core lexicon analysis closer to bridging the well-documented implementation gap: IWAs consistently report that discourse-level outcomes are among their most valued treatment goals (Worrall et al., 2011), yet clinicians cite time and training barriers to using discourse measures (Bryant et al., 2017; Cruice et al., 2020; Stark et al., 2021).

Core lexicon offers a clinically realistic path forward, a standardized yet flexible measure that can be deployed to track meaningful change in monologic discourse production. At the same time, it is important to underscore that validity and reliability are not fixed attributes of a tool. They are functions of how and where a measure is used (Cronbach & Meehl, 1955; Messick, 1995). Applying core lexicon analysis in contexts that diverge substantially from those in which its properties were established—for example, in progressive neurodegenerative populations, in non-English speakers, or with discourse tasks beyond the AphasiaBank set—may yield different psychometric behavior. Careful attention to context, population, and administration protocols remains critical for the responsible implementation of core lexicon analysis into research and clinical practice.

Based on the present findings and prior work (e.g., Stark et al., 2023), we can offer preliminary guidance for stimulus selection in clinical settings. For IWAs, core lexicon scores demonstrated good reliability even for single stimuli, with excellent reliability achieved when averaging two stimuli; accordingly, clinicians may use either one sufficiently rich narrative task or average at least two shorter stimuli to improve measurement precision when time is limited. Although Cinderella consistently showed the highest reliability in this study, this should not be interpreted as a recommendation for that stimulus specifically but rather as evidence that narratives, which elicit a sufficiently large, lexically diverse sample with broad core lexicon coverage, are more likely to yield stable measurement than shorter or less constrained tasks.

Limitations and Future Directions

Test–retest reliability was only examined across two, relatively short time points. While this is standard practice in psychometrics, it limits our ability to generalize these estimates of reliability across longer time windows. It also precludes our ability to estimate variance attributable to Person × Occasion interactions, constraining our ability to generalize these reliability estimates to future time points. Nonetheless, given the excellent reliability of core lexicon analysis, particularly for IWAs, we do not view this limitation as overly concerning.

In addition, our sample of IWAs may not fully represent the range of aphasia severity. Although the sample includes individuals diagnosed with Broca's, conduction, transcortical motor, and Wernicke's aphasia, it is predominantly made up of individuals with latent and anomic aphasia—given the NEURAL-2 study's intentional focus on recruiting individuals with mild aphasia. As such, our findings are not guaranteed to reflect the reliability of core lexicon for individuals with severe aphasia. However, inclusion of a greater number of individuals with severe aphasia is likely to increase the between-subjects variance in core lexicon scores, which is likely to increase reliability or offset minor increases in measurement error. In contrast, this restricted range may have a greater impact at different levels of discourse measurement (e.g., pragmatics, macrostructure/planning, propositional) and should therefore be addressed directly for the specific measure under investigation.

Furthermore, the individuals in this study were primarily monolingual speakers of English, and there was insufficient variability in the sample to investigate the impacts of bi-/multilingual status on reliability and internal consistency. While we would not expect significant differences in these results for an investigation of monolingual speakers of a language other than English, it is possible that individuals who are bilingual or multilingual could demonstrate different patterns of reliability and internal consistency. This could potentially be driven by factors such as language proficiency and/or language use. This is an empirical question that should be tested in future psychometric research of core lexicon analysis.

Finally, while core lexicon is a useful measure of lexical–semantic selection, informativeness, and lexical cohesion in discourse (because it includes functors such as "so," "there," etc.), it has only been developed in monologue discourse to date. There are marked differences in task demands and language production in discourse monologue versus dialogue (e.g., Leaman & Edmonds, 2023), which may limit the extent to which core lexicon reflects language use in everyday communication. However, strong correlations between core lexicon and aphasia severity (Chen & Chang, 2024; Kim et al., 2019, 2021), main concept analysis (Brisebois et al., 2026; Dalton & Richardson, 2015), and other discourse measures (Brisebois et al., 2026; Chen & Chang, 2024; Kim & Wright, 2020) suggest that, although the task demands are different, core lexicon

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

as a measure deployed in discourse monologue is still a good reflection of lexical–semantic access in general.

Future research should build on these findings by extending reliability analyses across a broader range of populations, tasks, and measurement frameworks. While this study focused on individuals with poststroke aphasia completing structured discourse tasks in English, the psychometric properties of core lexicon may differ in progressive conditions such as primary progressive aphasia or dementia, in multilingual speakers, or in more naturalistic conversational contexts. Cross-linguistic development and validation of core lexicon checklists are also critical steps toward global clinical applicability, given the sociolinguistic variability in lexical norms. Finally, clinical implementation research is needed to test the feasibility and reliability of perceptual, "live" core lexicon scoring in therapy settings. These directions will not only refine core lexicon analysis as a robust psychometric tool but also address the long-standing gap that has kept discourse measures from inclusion in core outcome sets for aphasia.

Conclusions

Overall, this study supports core lexicon analysis as a promising and clinically meaningful discourse measure, particularly for IWAs. The findings show that core lexicon performance is generally stable over short retest intervals, with especially strong reliability in the aphasia group and improved consistency when longer discourse samples or multiple tasks are combined. This suggests that core lexicon analysis can be used with confidence in research and, in many contexts, in clinical practice to monitor discourse production over time. At the same time, the variability seen in HCPs highlights that task selection and sampling method matter greatly when interpreting repeated measures. Taken together, the results strengthen the psychometric foundation of core lexicon analysis, provide practical guidance through MDC_{95} values, and support its growing role as a feasible tool for tracking meaningful changes in functional language ability.

Data Availability Statement

The data sets generated and/or analyzed during the current study are available on the Open Science Framework: https://osf.io/3mwzv/.

Acknowledgments

The larger NEURAL study was supported by an ASHFoundation New Investigator grant awarded to Brielle C. Stark.

References

Alyahya, R. S. W. (2025). Development and real-time clinical application of new transcription-less discourse assessment approaches for Arabic speakers with aphasia. *International Journal of Language & Communication Disorders, 60*(3), Article e70043. https://doi.org/10.1111/1460-6984.70043

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, S., Wright, A., Tippett, D. C., & Hillis, A. E. (2019). Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American Journal of Speech-Language Pathology, 28*(1S), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57*(3), 966–978. https://doi.org/10.1044/2014_JSLHR-L-13-0171

Boyle, M. (2015). Stability of word-retrieval errors with the AphasiaBank stimuli. *American Journal of Speech-Language Pathology, 24*(4), S953–S960. https://doi.org/10.1044/2015_AJSLP-14-0152

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews, 2016*(6), Article CD000425. https://doi.org/10.1002/14651858.CD000425.pub4

Brisebois, A., Brambati, S. M., Jutras, C., Rochon, E., Leonard, C., Zumbansen, A., Anglade, C., & Marcotte, K. (2023). Adaptation and reliability of the Cinderella story retell task in Canadian French persons without brain injury. *American Journal of Speech-Language Pathology, 32*(6), 2871–2888. https://doi.org/10.1044/2023_AJSLP-23-00101

Brennan, R. L. (2001). Advanced topics in univariate generalizability theory. In *Generalizability theory* (pp. 141–177). Springer. https://doi.org/10.1007/978-1-4757-3456-0_5

Brisebois, A., Brambati, S. M., Marois, É., & Marcotte, K. (2026). Core lexicon in Laurentian French speakers without brain injury: Development, validation, and reliability. *Journal of Speech, Language, and Hearing Research, 69*(3), 1091–1104. https://doi.org/10.1044/2025_JSLHR-24-00713

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37*(2), 399–407. https://doi.org/10.1044/jshr.3702.399

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Chen, J., & Chang, H. (2024). Core lexicon analysis of spoken discourse production by Mandarin Chinese speakers with aphasia. *Aphasiology, 38*(11), 1766–1785. https://doi.org/10.1080/02687038.2024.2325169

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication*

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Disorders, 55(3), 417–442. https://doi.org/10.1111/1460-6984.12528

Dalton, S. G., AL Harbi, M., Berube, S., & Hubbard, H. I. (2024). Development of main concept and core lexicon checklists for the original and modern Cookie Theft stimuli. *Aphasiology, 38*(12), 1975–1999. https://doi.org/10.1080/02687038.2024.2340794

Dalton, S. G. H., Hubbard, H. I., & Richardson, J. D. (2020). Moving toward non-transcription based discourse analysis in stable and progressive aphasia. *Seminars in Speech and Language, 41*(1), 032–044. https://doi.org/10.1055/s-0039-3400990

Dalton, S. G., & Richardson, J. D. (2015). Core-lexicon and main-concept production during picture-sequence description in adults without brain damage and adults with aphasia. *American Journal of Speech-Language Pathology, 24*(4), S923-S938. https://doi.org/10.1044/2015_AJSLP-14-0161

Dalton, S. G., Stark, B. C., Fromm, D., Apple, K., MacWhinney, B., Rensch, A., & Rowedder, M. (2022). Validation of an automated procedure for calculating core lexicon from transcripts. *Journal of Speech, Language, and Hearing Research, 65*(8), 2996–3003. https://doi.org/10.1044/2022_JSLHR-21-00473

Dalton, S. G. H., Kim, H., Richardson, J. D., & Wright, H. H. (2020). A compendium of core lexicon checklists. *Seminars in Speech and Language, 41*(1), 45–060. https://doi.org/10.1055/s-0039-3400972

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology, 32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021). Creating a theoretical framework to underpin discourse assessment and intervention in aphasia. *Brain Sciences, 11*(2), Article 183. https://doi.org/10.3390/brainsci11020183

Dutta, M., Murray, L. L., Park, H., Burklow, E., Bose, A., Kim, H., Greenslade, K., Ramage, A. E., Combs, C., Balasubramanian, A., & Casilio, M. (2025). Let's chat about spoken discourse: A tutorial to support use of spoken discourse analysis when providing aphasia clinical services. *American Journal of Speech-Language Pathology, 34*(4), 1931–1966. https://doi.org/10.1044/2025_AJSLP-24-00534

Fromm, D., Dalton, S. G., Brick, A., Olaiya, G., Hill, S., Greenhouse, J., & MacWhinney, B. (2024). The case of the cookie jar: Differences in typical language use in dementia. *Journal of Alzheimer's Disease, 100*(4), 1417–1434. https://doi.org/10.3233/JAD-230844

Jiang, Y.-E., Liao, X.-Y., & Liu, N. (2023). Applying core lexicon analysis in patients with anomic aphasia: Based on Mandarin AphasiaBank. *International Journal of Language & Communication Disorders, 58*(6), 1875–1886. https://doi.org/10.1111/1460-6984.12864

Kim, H., Berube, S., & Hillis, A. E. (2023). Core lexicon in aphasia: A longitudinal study. *Aphasiology, 37*(10), 1679–1691. https://doi.org/10.1080/02687038.2022.2121598

Kim, H., Kintz, S., & Wright, H. H. (2021). Development of a measure of function word use in narrative discourse: Core lexicon analysis in aphasia. *International Journal of Language & Communication Disorders, 56*(1), 6–19. https://doi.org/10.1111/1460-6984.12567

Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders, 54*(1), 62–78. https://doi.org/10.1111/1460-6984.12432

Kim, H., & Wright, H. H. (2020). Concurrent validity and reliability of the core lexicon measure as a measure of word retrieval ability in aphasia narratives. *American Journal of Speech-Language Pathology, 29*(1), 101–110. https://doi.org/10.1044/2019_AJSLP-19-0063

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. https://doi.org/10.1016/j.jcm.2016.02.012

Leaman, M. C., & Edmonds, L. A. (2023). Analyzing language in the picnic scene picture and in conversation: The type of discourse sample we choose influences findings in people with aphasia. *American Journal of Speech-Language Pathology, 32*(4), 1413–1430. https://doi.org/10.1044/2023_AJSLP-22-00279

Lüdecke D., Waggoner, P. D., & Makowski, D. (2019). insight: A unified interface to access information from model objects in R. *Journal of Open Source Software, 4*(38), Article 1412. https://doi.org/10.21105/joss.01412

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology, 25*(11), 1372–1392. https://doi.org/10.1080/02687038.2011.584690

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), Article 20170213. https://doi.org/10.1098/rsif.2017.0213

Park, J., Choi, S., Kim, S. E., Kim, H., & Sung, J. E. (2025). Development of core lexicon lists for cognitively healthy adults across age groups in a Korean storytelling task. *Communication Sciences and Disorders, 30*(3), 579–592. https://doi.org/10.12963/csd.250116

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders, 52*(6), 689–732. https://doi.org/10.1111/1460-6984.12318

Richardson, J. D., & Dalton, S. G. H. (2020). Main concepts for two picture description tasks: An addition to Richardson and Dalton, 2016. *Aphasiology, 34*(1), 119–136. https://doi.org/10.1080/02687038.2018.1561417

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology, 28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

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. https://doi.org/10.1044/2023_JSLHR-22-00266

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research, 64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Stark, B. C., & Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience, 36*(5), 562–585. https://doi.org/10.1080/23273798.2020.1862258

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Stark, B. C., Meinert, K., Urena, K., Oeding, G., Fromm, D., & MacWhinney, B. (2025). Introducing the NEURAL research lab data set for studies of discourse and gesture in aphasia and cognitively healthy aging adults. *Journal of Speech, Language, and Hearing Research, 68*(11), 5543–5556. https://doi.org/10.1044/2025_JSLHR-24-00732

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods, 27*(4), 650–666. https://doi.org/10.1037/met0000391

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., Babbitt, E., Bose, A., Brady, M., Cherney, L. R., Copland, D., Cruice, M., Enderby, P., Hersh, D., Howe, T., Kelly, H., Kiran, S., Laska, A.-C., Marshall, J., . . . Webster, J. (2019). A core outcome set for aphasia treatment research: The ROMA consensus statement. *International Journal of Stroke, 14*(2), 180–185. https://doi.org/10.1177/1747493018806200

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pederson, T. L., Miller, E., Bache, S. M., Müller, K. Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D., Ferguson, A., & Davidson, B. (2011). What people with aphasia want: Their goals according to the ICF. *Aphasiology, 25*(3), 309–322. https://doi.org/10.1080/02687038.2010.508530

Downloaded from: https://pubs.asha.org Carnegie Mellon University on 10/14/2026, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions