Research Article

The Severity-Calibrated Aphasia Naming Test

Grant M. Walker,[a] Julius Fridriksson,[b] Argye E. Hillis,[c] Dirk B. den Ouden,[b] Leonardo Bonilha,[d] and Gregory Hickok[a,e]

[a] Department of Cognitive Sciences, University of California, Irvine [b] Department of Communication Sciences and Disorders, University of South Carolina, Columbia [c] Departments of Neurology, Physical Medicine & Rehabilitation, and Cognitive Science, Johns Hopkins Medicine, Baltimore, MA [d] Department of Neurology, Emory University, Atlanta, GA [e] Department of Language Science, University of California, Irvine

ARTICLE INFO

Article History:
Received March 10, 2022
Revision received May 29, 2022
Accepted July 26, 2022

Editor-in-Chief: Katherine C. Hustad
Editor: Brent Archer

https://doi.org/10.1044/2022_AJSLP-22-00071

ABSTRACT

Purpose: We present a 20-item naming test, the Severity-Calibrated Aphasia Naming Test (SCANT), that can serve as a proxy measure for an aphasia severity scale that is derived from a thorough test battery of connected speech production, single-word production, speech repetition, and auditory verbal comprehension.
Method: We use lasso regression and cross-validation to identify an optimal subset from a set of 174 pictures to be named for prediction of aphasia severity, based on data from 200 participants with left-hemisphere stroke who were quasirandomly selected to represent the full impairment scale. Data from 20 healthy controls (i.e., participant caretakers/spouses) were also analyzed. We examine interrater reliability, test–retest reliability, sensitivity and specificity to the presence of aphasia, sensitivity to therapy gains, and external validity (i.e., correlation with aphasia severity measures) for the SCANT.
Results: The SCANT has extremely high interrater reliability, and it is sensitive and specific to the presence of aphasia. We demonstrate the superiority of predictions based on the SCANT over those based on the full set of naming items. We estimate a 15% reduction in power when using the SCANT score versus the full test battery's aphasia severity score as an outcome measure; for example, to maintain the same power to detect a significant group average change in aphasia severity, a study with 25 participants using the full test battery to measure treatment effectiveness would require 30 participants if the SCANT were to be used as the testing instrument instead.
Conclusion: We provide a linear model to convert SCANT scores to aphasia severity scores, and we identify a change score cutoff of four SCANT items to obtain a high degree of confidence based on test–retest SCANT data and the modeled relation between SCANT and aphasia severity scores.
Supplemental Material: https://doi.org/10.23641/asha.21476871

Evaluating language abilities remains a challenge in clinical settings due to the limited time available to clinicians to reach a diagnosis. In the United States, this time constraint is largely driven by policies set by third-party reimbursors, which often provide relatively minimal funding to support baseline testing. Early clinical assessment is also complicated by our evolving views of the various components making up the construct of "language ability." Language processing may be divided into core and supporting subprocesses (e.g., auditory comprehension, social intelligence, word finding, syntactic operations, semantic memory, working memory, articulation), each with their own associated tests and measurement scales, or alternatively, language processing may be gauged on a holistic scale of general severity. Although these approaches may provide complementary information, each approach has its advantages and disadvantages.

A Standard Measure of Aphasia Severity

The Western Aphasia Battery (WAB) is a popular standardized assessment for evaluating speech and language impairment, particularly after stroke (for a review of research and clinical applications, see Kertesz, 2022). The initial conceptualization of the scale was presented in

Correspondence to Grant M. Walker: grantw@uci.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

the work of Kertesz and Poole (1974), with minor revisions of several test items, additional (optional) subtests of related functional domains (i.e., reading, writing, and apraxia), and updated stimulus and scoring materials and instruction manuals being published in subsequent versions (Kertesz, 1982, 2007). The most recent version (Kertesz, 2007) is known as the Western Aphasia Battery–Revised (WAB-R), but the scores from different versions are essentially comparable. Unless an explicit distinction is made, we use the term *WAB* inclusively to refer to both versions of the test battery. The portion of the test battery that addresses speech processing is composed of 10 tasks assessing five domains: (a) fluency of spontaneous speech, (b) information content of spontaneous speech, (c) word finding, (d) speech repetition, and (e) auditory verbal comprehension. A composite summary score called the *Aphasia Quotient* (AQ) is calculated as a weighted average of these five subscores.

The AQ has been criticized as a measure of aphasia severity on several grounds, with the ordinal nature of the measure, the singular standard error of measurement (*SEM*) for the full range of ability, and the particular weighting of subtests being notable psychometric concerns (Hula et al., 2010; Risser & Spreen, 1985). Despite the AQ's compositional nature encompassing both receptive and expressive domains, principal component and factor analysis investigations have found that the majority of variance in the subscores can be explained by a single underlying factor (Hula et al., 2010; Kertesz & Phipps, 1977). Applying a formal psychometric measurement model (i.e., a Rasch model) to the WAB, Hula et al. (2010) found that, with the exclusion of a few test items and the application of a partial credit scoring scheme, responses to WAB items (across subtests) can be adequately modeled as a function of a single, unidimensional latent trait (i.e., aphasia severity). After the adjustments to the measurement scale, the Rasch-based WAB measure and AQ still had r^2 = .90 (W. Hula, personal communication, January 31, 2022). The AQ shares most of its variance with this modeled latent trait. Although the Rasch-based measure may have stronger claims to validity as a gold standard of aphasia severity, the use of this statistic requires rescoring of the WAB and application of a computational model that is not publicly available. On the other hand, the AQ has frequently been analyzed and reported in aphasia treatment and neuroimaging research (Kertesz, 2022), with sensible and convincing results.

Another limitation of the WAB is that it sometimes classifies individuals with relatively mild aphasia who show impairments on discourse measures and endorse language deficits in daily life as not having aphasia (Cunningham & Haley, 2020; Dede & Salis, 2019). It also fails to detect language deficits in individuals with pure agrammatism, manifest as difficulty formulating grammatically correct sentences (Caramazza & Hillis, 1989; Miceli et al., 1983; Nespoulous et al., 1988), because it does not include a sensitive test for agrammatism. A person who can produce only simple subject–verb and verb phrases, even with hesitations, might be given a 4, 6, or 9 (out of 10) on the Fluency scale of the WAB, given that the sample on which scoring is based is limited to responses to simple, open-ended questions. Even among experienced speech-language pathologists, there is low interrater reliability in measuring Fluency, one of the two subtests of Spontaneous Speech (Trupe, 1984). Because the subjective scale of Fluency is multidimensional, variability between raters in terms of how they weigh grammaticality, phrase length, pauses, rate of speech, and articulation often results in markedly different scores for the same individual. Although this weakness mostly affects "classification" (aphasia type), it can also influence severity and diagnosis (presence or absence of aphasia), given the weight of the Fluency score in calculating AQ.

Despite its limitations, the AQ (along with other measures from the WAB) has some consensus support among rehabilitation experts as being a preferred outcome measure for aphasia treatment research (Wallace et al., 2019). While the AQ may not represent a "gold" standard of aphasia severity in terms of being an absolutely optimal measure, it is an accepted standard that is in widespread use and is familiar to many practitioners. Survey respondents from the United States, the United Kingdom, Canada, New Zealand, Australia, and Chile indicated that the WAB was the most frequently used aphasia assessment across all clinical and research settings (Kiran et al., 2018). Given its wide use and the deep scrutiny it has received (Kertesz, 2022), the WAB AQ is a reasonable choice for a standard measure of aphasia severity, despite its noted limitations. Of course, any new test that is calibrated to the WAB AQ would be expected to inherit these limitations.

A Proxy Measure of Aphasia Severity

One of the primary drawbacks of the WAB, however, is that it takes a relatively long time to administer. To complete all the subtests that contribute to the AQ takes approximately 45 min. While this may fit within a single session in a clinical setting, there is ultimately an opportunity cost in focusing so much on impairment-based assessment (Tierney-Hendricks et al., 2021). The WAB-R (Kertesz, 2007) includes a bedside short form that takes somewhere between 15 and 20 min to administer, but it is not clear how widely it is used in clinical settings nor how its psychometric properties differ from the full test battery (El Hachioui et al., 2017).

The Quick Aphasia Battery (QAB; Wilson et al., 2018) fills a gap between comprehensive evaluations of aphasic impairments and fast aphasia screening tests. The QAB was designed to take about 15 min to administer, and there are three parallel forms. The QAB evaluates expressive and receptive functions that are known to dissociate from one another in the aphasia population (e.g., word comprehension and sentence comprehension), and it yields an alternative composite measure of aphasia severity based on a more even weighting of expressive and receptive functions than the WAB AQ, which is biased toward expressive functions. In a sample of 16 participants with chronic aphasia, the QAB Aphasia Severity measure and WAB AQ had $r^2 = .88$.

The Comprehensive Aphasia Test (Howard et al., 2009; Swinburn et al., 2004) provides an alternative approach to the WAB, eschewing syndrome-based assessment to focus on the degrees of impairment within multiple psycholinguistic domains. This newer test battery incorporates decades of research into balancing the known features of testing materials that influence performance on structured tasks. In addition to language impairment, this test battery assesses frequently co-occurring cognitive impairments and the functional impact of the aphasia on quality of life. It is supported by normative data and provides guidance for therapy planning and prognosis. It can also take a long time to administer and score, about 1–2 hr. The Comprehensive Aphasia Test includes a statistic representing the overall severity of language impairment, the modality mean, which is an average score across spoken and written language comprehension, repetition, spoken naming, spoken picture description, reading aloud, writing, and written picture description (Howard et al., 2009). We were unable to find a direct comparison of this score with the WAB AQ in the literature as evidence of concurrent validity. However, Howard et al. (2009) reported that among 64 people with chronic aphasia, this score shared 73.3% of its variance ($r^2 = .733$) with the therapists' impairment rating from Enderby's (1997) therapy outcome measures.

Notwithstanding standardized test batteries of comprehensive language functions, assessment of naming in aphasia treatment studies has become commonplace (Conroy et al., 2018; Evans et al., 2021; Fridriksson et al., 2018; Howard et al., 1985; Leonard et al., 2008; Martin et al., 2020; Raymer et al., 1993). Some have raised the concern that picture naming lacks ecological validity, because most adult language experience involves connected speech in the context of a conversation. However, one-word utterances serve as the building blocks for syntactic and semantic structure in child language development (Bloom, 1976; McNeill, 1970). As Dore (1975) explains, one-word utterances are "primitive speech acts" that serve important ecological functions including labeling, repeating, answering, requesting, calling, greeting, protesting, and practicing. Assessing the integrity of this foundational production process is an efficient way to gain insight into the language system as a whole. At first glance, picture naming may seem to tap into only a narrow range of speech and language ability; however, naming has several features that could make it an appealing option to target for aphasia assessment. First, naming difficulty is a common clinical sign in almost all types of aphasia (Kohn & Goodglass, 1985), allowing the detection of aphasia independent of type. Second, naming has proven to be a highly sensitive measure to the presence of aphasia (Calero et al., 2002; MacOir et al., 2021). Third, naming tests are strongly correlated with overall aphasia severity as measured by more comprehensive assessments such as the WAB (MacWhinney et al., 2011; Mirman et al., 2010). Similarly, Fergadiotis et al. (2018) found that confrontation naming scores were strongly related to informativeness of monologic discourse. Fourth, the brain's naming network—as it is understood from functional imaging (Price, 2000), electrocorticography (Saravani et al., 2019), direct cortical stimulation mapping (Corina et al., 2010), and lesion studies (DeLeon et al., 2007)—aligns quite well with the language network as a whole, consistent with the first three points noted above. Finally, naming tests are easy to administer and easy to score, providing substantial practical advantages for clinicians. The goal of this study was to develop and evaluate a new naming test that (a) optimally predicts aphasia severity as measured by a standard comprehensive aphasia assessment tool (WAB AQ) and (b) does so in a time-efficient manner. We call this new naming test the Severity-Calibrated Aphasia Naming Test (SCANT).

Method

Participants

We examined archived data from three large-scale studies of left-hemisphere stroke and aphasia: (a) an R01-funded study of psycholinguistic factors in aphasia at MossRehab in Philadelphia, PA, that created the Moss Aphasia Psycholinguistic Project Database (MAPPD; Mirman et al., 2010); (b) an R01-funded study of neuroimaging and lesion mapping in aphasia (LESMAP) at University of South Carolina (Yourganov et al., 2015); and (c) a P50-funded study of predictors of language outcomes after rehabilitation (POLAR), also at University of South Carolina (Spell et al., 2020). From each database, we included all participants with (a) a single, left-hemisphere stroke without other degenerative neurological or psychiatric comorbidities and correctable vision; (b) a complete first administration of the Philadelphia Naming Test (PNT; Roach et al., 1996) with at least one complete naming attempt; and (c) a WAB AQ score, yielding a total N = 360 participants (MAPPD n = 183; LESMAP n = 89; POLAR n = 88) for the SCANT construction and modeling. Eighteen individuals participated in both the LESMAP and POLAR studies; because the data were collected at different times, years apart, both sets were included for analysis. After constructing the SCANT, we examined SCANT scores from 98 POLAR participants with test–retest data available. Ten of these participants had incomplete naming tests at enrolment, preventing their inclusion in the cohort for SCANT construction, but they had complete pairs of naming tests later in the longitudinal POLAR study. We also examined SCANT scores from 20 healthy control participants in the MAPPD database (caregivers and spouses of participants with aphasia) for analysis of sensitivity and specificity for the presence of aphasia. Table 1 presents descriptive statistics for demographic and clinical variables from each group, including the breakdown of aphasia types assigned by the WAB. All studies were approved by their respective institutional review boards, and all participants provided informed consent.

Testing Materials, Data, and Statistical Analysis

All participants were administered the PNT (Roach et al., 1996) upon enrolment in each study, along with the WAB (Kertesz, 1982) for MAPPD participants and the WAB-R (Kertesz, 2007) for LESMAP and POLAR participants. The PNT consists of 175 black-and-white line drawings to be named, and we examined item-level accuracy data (i.e., whether each item was named correctly or incorrectly by each participant). Because the item "Eskimo" was replaced with the item "umbrella" in the POLAR study for cultural sensitivity reasons (Bernstein-Ellis et al., 2021), this trial was excluded from all analyses. The WAB AQ was obtained for each participant.

In addition to baseline administrations of PNT and WAB-R tests upon enrolment, the POLAR study also included multiple administrations of the full PNT (~92% of test–retest administrations were within 1–3 days, whereas the remaining ~8% were within 4–21 days, with no change in ability expected) and WAB AQ (all test–retest administrations were within 1.25–3.14 years). After constructing the SCANT, we examined the 249 test–retest pairs of the SCANT scores that were available from 98 participants (one to six test–retest pairs per participant). We also examined the 39 test–retest pairs of WAB AQ scores that were available (one test–retest pair per participant).

Interrater reliability for SCANT scoring was examined using the published reliability data from the POLAR study (Walker et al., 2021). Five pairs of raters (speech-language pathology master's degree students) scored nine PNTs that were selected to represent the full range of ability. We report Cronbach's alpha and the total number of scoring disagreements for the SCANT items.

Table 1. Clinical and demographic information for the participants included in each group.

Database	MAPPD	LESMAP	POLAR	POLAR	POLAR	MAPPD
Purpose	SCANT construction	SCANT construction	SCANT construction	WAB AQ test–retest	SCANT test–retest	Healthy controls
Participants (*n*)	183	89	88	39	98	20
Sex (F/M)	82/101	36/53	41/47	12/27	39/59	13/7
Age (years)	58 (24–79)	62 (36–83)	61 (35–80)	60 (35–76)	60 (29–80)	NA
Education (years)[a]	12 (6–22)	NA	16 (12–20)	16 (12–20)	16 (12–20)	NA
Months poststroke	11 (1–181)	22 (6–276)	35 (10–241)	33 (12–241)	29 (10–241)	NA
Race						
African American	85	10	11	10	21	NA
Asian	0	0	0	0	1	NA
White	98	79	67	29	76	NA
Speech motor deficit	37	36	53	28	60	0
Aphasia type						
Anomia	68	27	24	9	29	0
Broca's	44	32	28	22	39	0
Conduction	30	10	11	4	15	0
Global	1	7	0	1	1	0
Transcortical motor	2	0	1	0	1	0
Transcortical sensory	2	0	0	0	0	0
Wernicke's	23	6	4	2	6	0
None	13	7	20	1	7	20
WAB AQ	77 (25–98)	68 (16–97)	78 (22–100)	54 (20–100)	67 (20–100)	NA
PNT accuracy	.67 (.01–.98)	.41 (.00–.96)	.73 (.00–.99)	.30 (.00–.98)	.55 (.00–.99)	.99 (.90–1.00)

Note. The median of continuous measures is reported with the range in parentheses. MAPPD = Moss Aphasia Psycholinguistic Project Database; LESMAP = lesion mapping in aphasia; POLAR = predictors of language outcomes after rehabilitation; SCANT = Severity-Calibrated Aphasia Naming Test; WAB AQ = Western Aphasia Battery Aphasia Quotient (Kertesz, 1982); F = female; M = male; PNT = Philadelphia Naming Test; NA = not applicable.

[a]There are 23 missing values, *n* = 160 for the "Education (years)" variable.

All data processing and statistical analysis were handled in the MATLAB (R2021b) programming environment (The MathWorks, Inc., 2021), using both standard functions and custom scripts. Standard functions for statistical analysis included *lasso* for lasso regression to identify useful predictors, *ICC* (intraclass correlation coefficient) for test–retest reliability, *corr* for linear association strength, *regstats* for linear regression analysis and partial correlations, *perfcurv* for receiver operator characteristic analysis (i.e., sensitivity and specificity), *fexact* for Fisher's exact test comparing frequencies in a contingency table, *ttest* for paired Student's *t* test comparing group means, and *r_test_paired* for paired test of correlation coefficients. Custom scripts were used for nested cross-validation routines. These scripts are available in Supplemental Materials S1, S2, and S3.

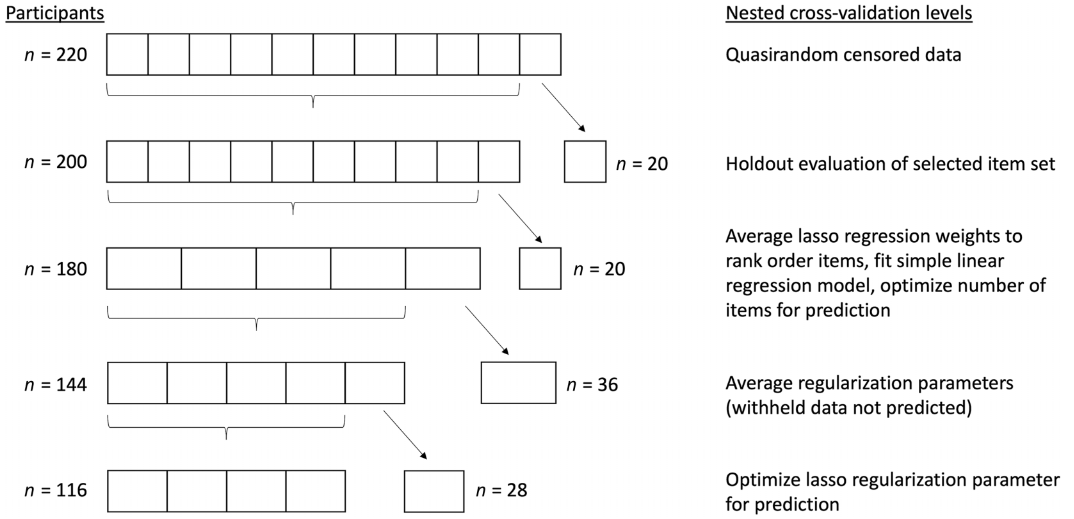
SCANT Item Selection

Our goal was to select a subset of PNT items that optimized out-of-sample predictions of WAB AQ using a simple linear regression model with the sum of correctly named items as the independent variable. To do this, we used a feature selection technique from machine learning: lasso regression with nested cross-validation and quasirandom censored partitions. Lasso regression serves the same purpose as stepwise regression (i.e., eliminating redundant or ineffective predictors) but evaluates all predictors simultaneously rather than sequentially. Cross-validation is a method that partitions data into training and testing sets to evaluate the generalizability of a model that has been fit to one set of data to a new set of data (i.e., to avoid overfitting noise in the training data). This approach helps to ensure that the test items selected by lasso regression will be similarly predictive for new participants. Quasirandom sampling with censoring is a method of selecting participants to ensure evenly distributed coverage over a scale (i.e., to avoid "clumping" that is inherent with pure random selection). Participants along the full scale of impairment should be evenly represented during test construction and evaluation. The following paragraphs explain these steps in more detail.

Within the lasso regression paradigm, we treated each participant–item combination as a binary, independent variable (i.e., taking the value 1 or 0 to indicate if that item was named correctly or incorrectly by that participant) and each participant's WAB AQ as the dependent variable. A coefficient was estimated for each item, along with a regularization parameter that drives as many coefficients to zero as possible while maintaining an optimal prediction accuracy for WAB AQ. Because the optimization procedure depended on random samples of data, we repeated the procedure 5 times, each time excluding a subset of participants, to obtain an average weighting of items that was less dependent on any particular sample.

Cross-validation (i.e., partitioning the data into training and testing sets for out-of-sample predictions) was used (a) to optimize the regularization parameter (i.e., ranking the items based on importance for prediction), (b) to minimize the number of items required for optimal predictions using the accuracy rate (i.e., to determine the length of the SCANT), and (c) to estimate the simple linear regression model (i.e., to convert SCANT scores into WAB AQ scores). Because cross-validation was used to optimize multiple parameters, the partitions were nested to avoid contamination across partitions (see Figure 1).

Figure 1. A diagram of the nested cross-validation schematic used to construct and evaluate the Severity-Calibrated Aphasia Naming Test. Aside from the initial holdout set, all cross-validation partitions were used as training data for predictions of withheld testing data. The withheld data were not predicted within the step of averaging regularization parameters; the data were withheld to create variability in the regularization training set.


Participants
Nested cross-validation levels
n = 220
Quasirandom censored data
n = 200
n = 20
Holdout evaluation of selected item set
n = 180
n = 20
Average lasso regression weights to rank order items, fit simple linear regression model, optimize number of items for prediction
n = 144
n = 36
Average regularization parameters (withheld data not predicted)
n = 116
n = 28
Optimize lasso regularization parameter for prediction

That is, if any data were used to estimate a part of the model, they could not be used again to evaluate an elaborated form of the model.

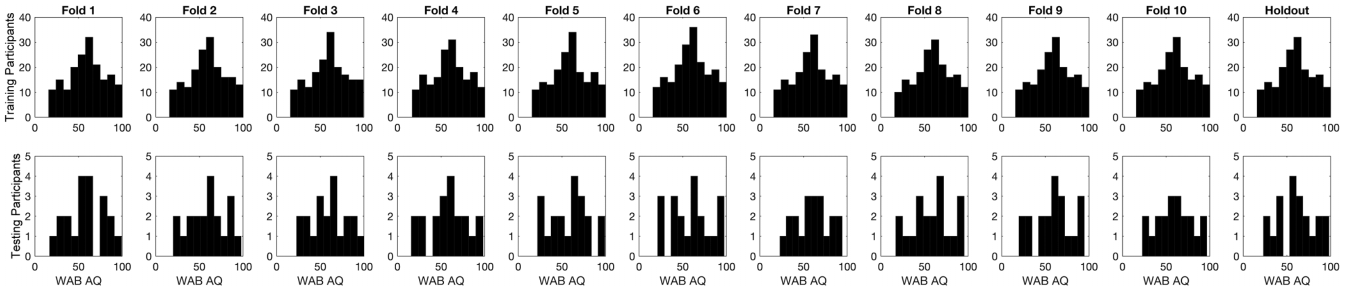
Participants with mild impairments were overrepresented in the full data set, but we wanted our testing sets for cross-validation to each include an approximately uniform distribution over the accuracy scale, so as not to bias our item selection procedure toward any subgroup of participants. To ensure this, we used quasirandom assignment of participants to partitions (i.e., based on the participant with the closest WAB AQ to a low-discrepancy sequence that avoids the "clumping" inherent in random selection), followed by censoring to remove participants with overrepresented accuracy rates (Joe & Kuo, 2003; Sobol', 1967). The required number of participants to remove is somewhat arbitrary; we used graphical analysis of the distributions within each partition to strike a balance between maximizing inclusion and approximating uniformity of WAB AQ within the testing partitions (see Figure 2). We used data from 200 participants to construct the SCANT (i.e., 10-fold cross-validation) and data from another 20 holdout participants to evaluate its prediction accuracy.

Specifically, to construct the SCANT, for each of the 10 folds at the top level of cross-validation, we averaged the item-level coefficients from the training lasso regression models to assign a value indicating each item's importance for prediction. Items were then sorted by these values, and lists of increasing length of items with highest importance were used to calculate naming accuracy and estimate a simple linear model with the training data. This regression model was then applied to the testing data to estimate prediction error (root-mean-square error) for each item set size, within each cross-validation fold. These prediction errors were then averaged over cross-validation folds, and the minimum number of items that had less average cross-validation prediction error than that of the full test was identified. Finally, we averaged the cross-validation training lasso regression weights over all 10 folds to sort the items by importance and select the top number of indicated items to construct the SCANT. The entire procedure was repeated 5 times to assess the reliability of the resulting item sets.

Predicting WAB AQ Scores From SCANT Scores

A simple linear regression model was constructed to predict WAB AQ scores from SCANT scores based on the 200 participants included in the cross-validation analysis (i.e., the same participants whose data were used to select the items). This model was then applied to the data from the 20 holdout participants, which was not used for item selection or model fitting. To compare predictions based on the SCANT with predictions based on the full PNT, a similar linear regression model based on PNT accuracy was constructed using the same cross-validation training data that were used for the SCANT model and again applied to the data from the holdout participants. We report the coefficient of determination (r^2) and the mean absolute error (MAE) between predicted and observed scores as measures of prediction accuracy. To determine whether the SCANT items have an especially strong relationship with WAB AQ compared with other possible sets of items, we calculated a permutation *p* value for the observed MAE when using the SCANT scores to predict WAB AQ scores; that is, we compared the observed MAE to a distribution of MAEs produced by 100,000 randomly constructed 20-item short forms. The *p* value is calculated as the proportion of these random item sets that exhibits a stronger relationship with WAB AQ than the SCANT items exhibit. In addition to examining prediction accuracy for the holdout sample, we used leave-one-out cross-validation (LOOCV) with the full data set of 360 participants (i.e., refitting the linear regression model to predict each holdout participant) to estimate prediction accuracy in a much larger group that included the original testing holdout participants (n = 20), the overrepresented holdout participants (n = 140), and the participants whose data were used to select the SCANT items (n = 200), again for both the SCANT and the PNT.

Figure 2. The distributions of WAB AQ within the training and testing sets for the 10 cross-validation folds and the holdout data set. WAB AQ = Western Aphasia Battery Aphasia Quotient.



Interpreting Longitudinal Change in SCANT Scores

We created percentile scores to help interpret longitudinal absolute change in SCANT scores based on the test–retest data from the POLAR study. Because the distribution of baseline SCANT scores (i.e., Time 1) was not uniform in the full data set (n = 249 test–retest pairs), we used 1,000 random permutations drawn uniformly from 11 bins of baseline SCANT accuracy rates, with 13 test–retest pairs randomly selected per bin (n = 143 test–retest pairs). The number of test–retest pairs per bin was selected based on the minimum frequency observed in a histogram of the full data set. For each permutation, we calculated the percentile score (i.e., the proportion of test–retest change scores that were less than the observed change score) for each possible level of change and then averaged over permutations to yield the final percentile scores.

In addition to examining test–retest reliability under conditions of no expected change, we also examined the sensitivity of SCANT scores to changes associated with therapy. There were 75 POLAR participants who had complete SCANT data sets at enrolment and immediately after the first phase of a therapy program. Participants were randomly assigned to semantically oriented therapy (n = 37) or phonologically oriented therapy (n = 38). The items on the PNT (and SCANT) were not treated specifically for naming. Details of the therapy protocols can be found in the work of Spell et al. (2020). There were at least two issues regarding treatment-induced change that were of interest in this study. The first issue: Do SCANT scores change in response to therapy and by how much? To address this question, we compared individual changes in SCANT scores in response to treatment to individual changes in PNT scores, the original outcome measure for the study. The second issue: Are changes in SCANT scores related to changes in WAB AQ scores? Unfortunately, of the 39 participants with test–retest available for WAB AQ, only 13 participants had complete SCANT data at the corresponding assessment time points. Therefore, we consider our investigation of this second issue to be preliminary.

To investigate the sensitivity of SCANT scores to therapy-induced change, we compared different methods for classifying participants as significantly improved or unimproved. Specifically, we compared SCANT change thresholds of four, three, or two items with the corresponding thresholds indicated by a Fisher's exact test when identifying significant score improvements for the PNT. Participants with fewer SCANT errors than the given threshold were excluded, as they would be unable to exhibit the requisite improvement in principle, yielding groups with 52, 56, and 63 participants, respectively. The alpha criteria for significance of the Fisher's exact tests were matched to the SCANT test–retest percentile scores so that each pair of thresholds approximated the same purported Type I error rates (alpha = .07, .17, and .4, respectively). After categorizing each participant as improved or unimproved with each method (SCANT or PNT), there were two ways to construct contingency tables and apply Fisher's exact test at the group level. First, we tested the null hypothesis that the two classification methods indicate the same rate of improvement among participants, ignoring the identity of which participants improved. A significant result implies that the classification methods yield different rates of improvement due to therapy. Second, we evaluated the agreement of the classification methods for each individual participant. Specifically, we tested the null hypothesis that the classification of a participant as improved or unimproved by one method is independent of the other method's classification of the same participant. A significant result indicates that the individual classifications made by each method are related to one another.

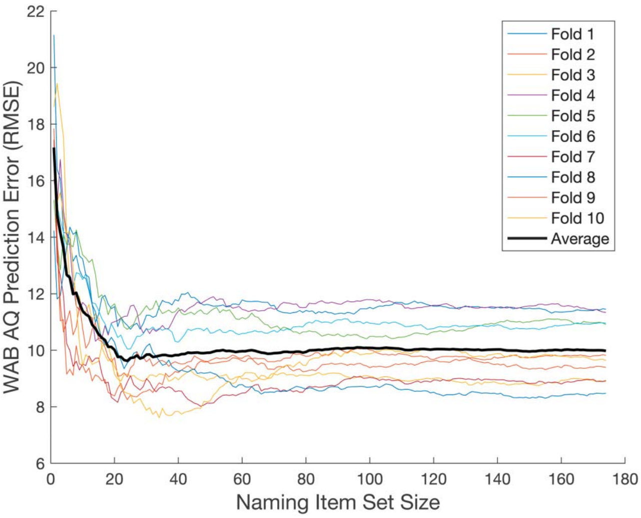
Results

Item Selection

The minimum number of items that yielded less average cross-validation prediction error than that of the full test was 20 items (see Figure 3). These 20 items constitute the SCANT and are listed in Appendix A. The SCANT picture stimuli and score sheet are available in Supplemental Materials S4.

The item selection procedure yielded reasonably reliable sets of items. The five generated sets shared 50%–70% of their items in pairwise comparisons. On average, the SCANT items appeared 3.55 times across the five generated sets. The 40 unique items that were selected in total, along with their selection frequencies on the five generated sets, are listed in Appendix B. The log lexical frequency in movie and television transcripts, the number of phonological segments, and the phonological neighborhood density of each selected item, as reported by Walker et al. (2018), is also presented in Appendix B. While a detailed investigation into the theoretical forces motivating item selection is beyond the scope of this article, we note that, compared with items that were never selected, the SCANT items have significantly higher average lexical frequency (6.66 vs. 6.18, p = .028), but not significantly different numbers of phonological segments (4.08 vs. 4.46, p = .20) nor phonological neighborhood densities (2.50 vs. 2.53, p = .23). The SCANT items were significantly easier to name by participants with aphasia on average than items that were not selected (62% vs. 55% mean accuracy, p = .0076). We suspect that the high frequency of the items provides control for noisy measurement due to variable linguistic experience, allowing the impact of the

Figure 3. WAB AQ prediction accuracy (RMSE) as a function of naming item set size for each of the 10 cross-validation folds and the grand average. WAB AQ = Western Aphasia Battery Aphasia Quotient; RMSE = root-mean-square error.



injury to drive the performance rather than environmental factors.

Interrater Reliability

There was extremely high interrater reliability for the SCANT (Cronbach's alpha = .99). Among 180 trials from nine participants with aphasia scored by five pairs of raters, there was only one disagreement regarding the correctness of a response.

Sensitivity and Specificity for Presence of Aphasia

A differential diagnosis chart is presented in Table 2, categorizing Healthy Control participants, Stroke—No Aphasia by WAB (NABW) participants (WAB AQ ≥ 93.8), and Stroke—Aphasia participants (WAB AQ < 93.8) based on an observed SCANT score. SCANT scores of 17 or less indicate stroke aphasia detectable by the WAB. A SCANT score of 18 indicates left-hemisphere stroke, with a 63% probability that the WAB will detect aphasia. A SCANT score of 19 provides the least information to discriminate between the groups because people without aphasia tend to have higher scores and people with aphasia tend to have lower scores; there is only a 33% chance that the WAB will detect aphasia. The maximum SCANT score of 20 indicates there is less than a 4% chance that the person has stroke aphasia detectable by WAB, though a stroke with mild aphasic symptoms (NABW) is still possible. The group-level naming patterns are consistent with the expected gradient of ability observed in the discourse analysis (Dalton & Richardson, 2015; Fromm et al., 2017; Richardson et al., 2018), with the highest scores in Healthy Control participants, slightly lower scores in Stroke—NABW participants, and the lowest scores in Stroke—Aphasia participants.

Table 2. Differential diagnosis chart indicating the probability of belonging to each group based on the SCANT score.

SCANT score	Healthy control	Stroke—NABW	Stroke—Aphasia
20	50%	46%	4%
19	34%	34%	32%
18	0%	37%	63%
17 or less	0%	0%	100%

Note. SCANT = Severity-Calibrated Aphasia Naming Test; NABW = no aphasia by Western Aphasia Battery (i.e., Aphasia Quotient ≥ 93.8).

The area under the receiver operator characteristic curve discriminating Stroke—Aphasia participants from Healthy Control participants based on SCANT score was .95 (bootstrap 95% CI [.91, .98]). This means that a randomly selected Healthy Control participant has a 95% chance of having a higher SCANT score than a randomly selected participant with aphasia. A cutoff score of one error to discriminate Stroke—Aphasia participants from Healthy Control participants yielded a sensitivity of 93% and a specificity of 90%; a cutoff score of two errors yielded a sensitivity of 83% and a specificity of 100%. That is, two errors are enough to rule out a Healthy Control participant, but 7% of participants with stroke aphasia (detectable by WAB) still achieved the maximum SCANT score. For comparison, among 27 healthy controls and 266 people with aphasia, the Comprehensive Aphasia Test (Swinburn et al., 2004) correctly categorized 93.1% of individuals into their respective groups (Bruce & Edmundson, 2009); among 20 healthy controls and 83 people with aphasia, the SCANT correctly categorized 92.2% of individuals into their respective groups.

Only two items elicited errors from Healthy Control participants: *man* and *lion*. Each item elicited an error from a single control participant. Three items elicited errors from Stroke—NABW participants: *football*, *bone*, and *banana*. *Football* and *bone* each elicited an error from a single participant, and *banana* elicited errors from two participants. Among Stroke—Aphasia participants, the relative frequencies of errors elicited by each item ranged from a minimum of 25% (*nose*) to a median of 45% (*queen, banana*) to a maximum of 64% (*helicopter*), meaning that there was an appreciable range of difficulty among the items, but no items were consistently named correctly or incorrectly.

Predicting WAB AQ Scores From SCANT Scores

The simple linear regression model to predict WAB AQ scores from SCANT scores based on the 200 participants included in the cross-validation analysis yielded:

$$\text{WAB AQ} = 31.07 + (2.86 \times \text{SCANT}) \qquad (1)$$

Within the holdout data set, predictions of WAB AQ using the SCANT (see Figure 4, left) were more accurate on average than predictions using the full PNT (see Figure 4, right), with SCANT predictions yielding a higher r^2 (.86 for SCANT vs. .81 for PNT, p = .256) and lower MAE (6.64 for SCANT vs. 7.45 for PNT, p = .226). The permutation p value for the SCANT was .0175, meaning that the SCANT had lower out-of-sample prediction error for the holdout data than 98.25% of randomly constructed 20-item short-form tests. In other words, responses to the SCANT items yielded unusually good predictions of WAB AQ compared with responses to other possible sets of items. Within the full data set, again, the SCANT yielded better predictions of WAB AQ (see Figure 5, left) than the full PNT (see Figure 5, right), with higher r^2 (.83 for SCANT vs. .80 for PNT, p = .0092) and lower MAE (7.11 for SCANT vs. 7.56 for PNT, p = .0215). Thus, while the modest improvement in prediction accuracy was not significant in the holdout set of 20 participants, the average improvement was reliable when considering the larger group.

The coefficient of determination, r^2, between SCANT scores and WAB AQ scores was .86 in the holdout set and .83 in the full data set, so a population estimate of .85 is well supported by the evidence. This means that a clinical study using SCANT scores as an outcome measure would have 15% less power than the same study using WAB AQ as an outcome measure (Armstrong, 1996). That is, any estimated power that depends on normal distributions and is conducted on WAB AQ scores can be multiplied by 0.85 to obtain the power for a test of SCANT scores. Alternatively, any estimated sample size that yields a given power for detecting differences in WAB AQ can be multiplied by 1.15 to obtain the corresponding sample size for a test of SCANT scores that has the same power. For example, to maintain the same power to detect an effect, a study that uses WAB AQ as an outcome measure with 25 participants would need to have 30 participants if SCANT scores were to be used as the outcome measure instead. Given that the WAB requires a substantial number of materials and can take up to an hour to administer and even more time to score, a simple naming accuracy test that takes less than 5 min to administer and score could improve participant compliance and research throughput.

Interpreting Longitudinal Change in SCANT Scores

We present test–retest differences for WAB AQ (i.e., predicting WAB AQ at Time 2 from WAB AQ at Time 1)

Figure 4. Predictions of WAB AQ scores in the holdout data set from SCANT scores (left) and PNT scores (right). The diagonal line is the identity line. SCANT = Severity-Calibrated Aphasia Naming Test; PNT = Philadelphia Naming Test; WAB AQ = Western Aphasia Battery Aphasia Quotient; MAE = mean absolute error.

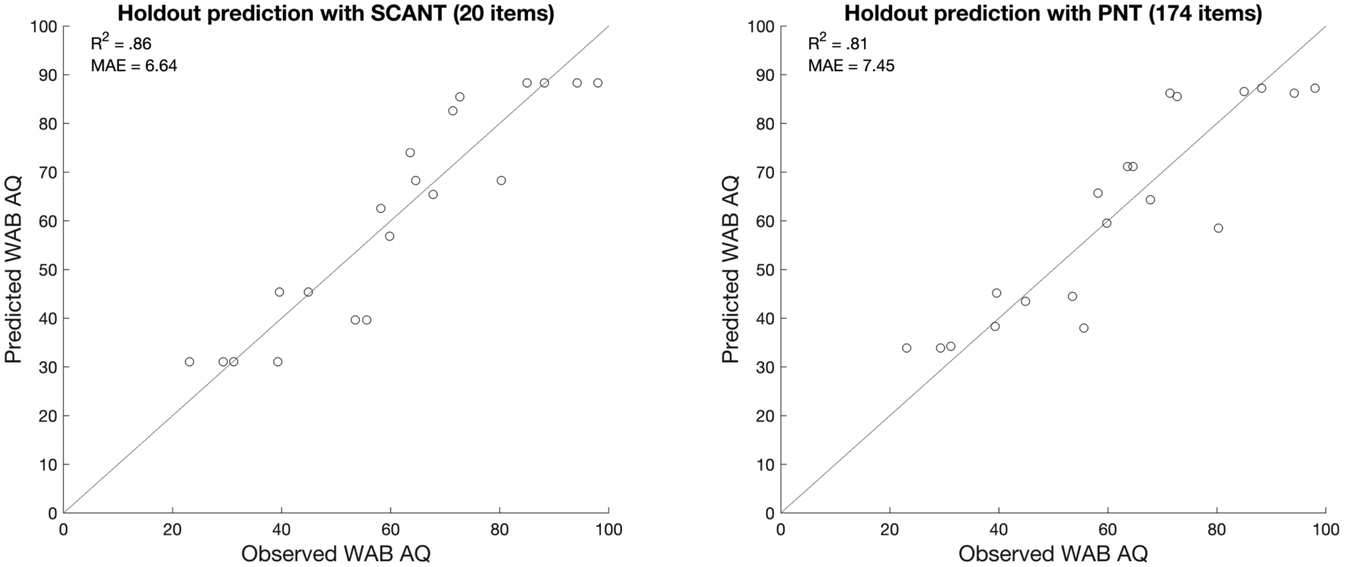
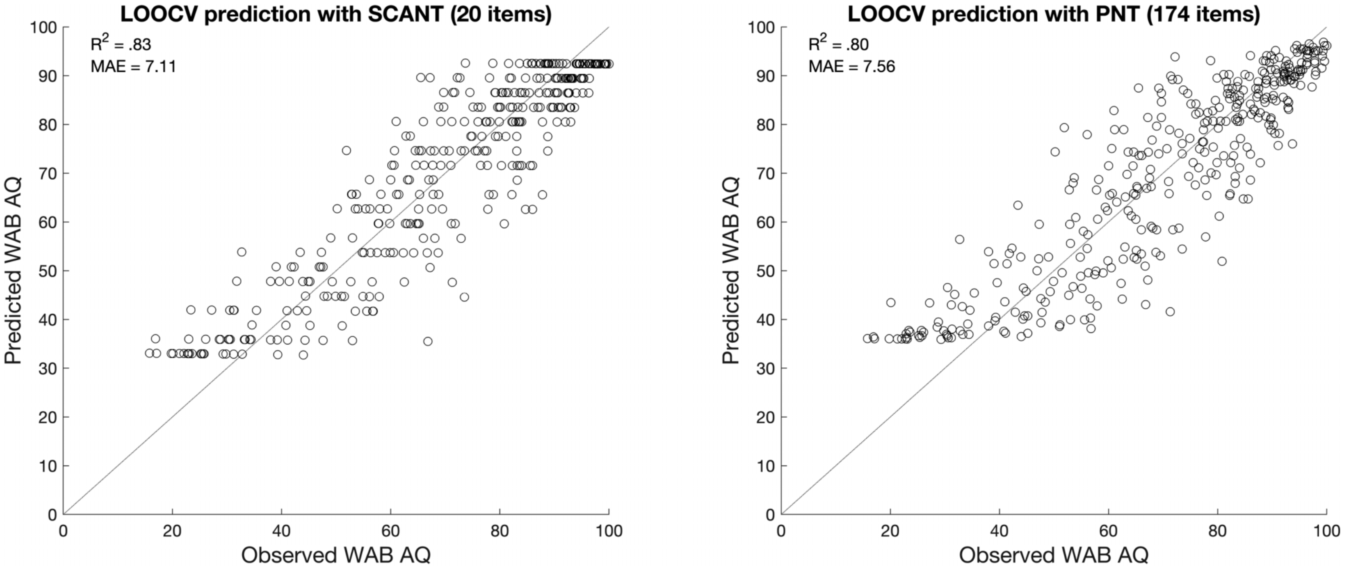


Figure 5. Leave-one-out cross-validation (LOOCV) predictions of WAB AQ scores in the full data set from SCANT scores (left) and PNT scores (right). The diagonal line is the identity line. SCANT = Severity-Calibrated Aphasia Naming Test; PNT = Philadelphia Naming Test; WAB AQ = Western Aphasia Battery Aphasia Quotient; MAE = mean absolute error.

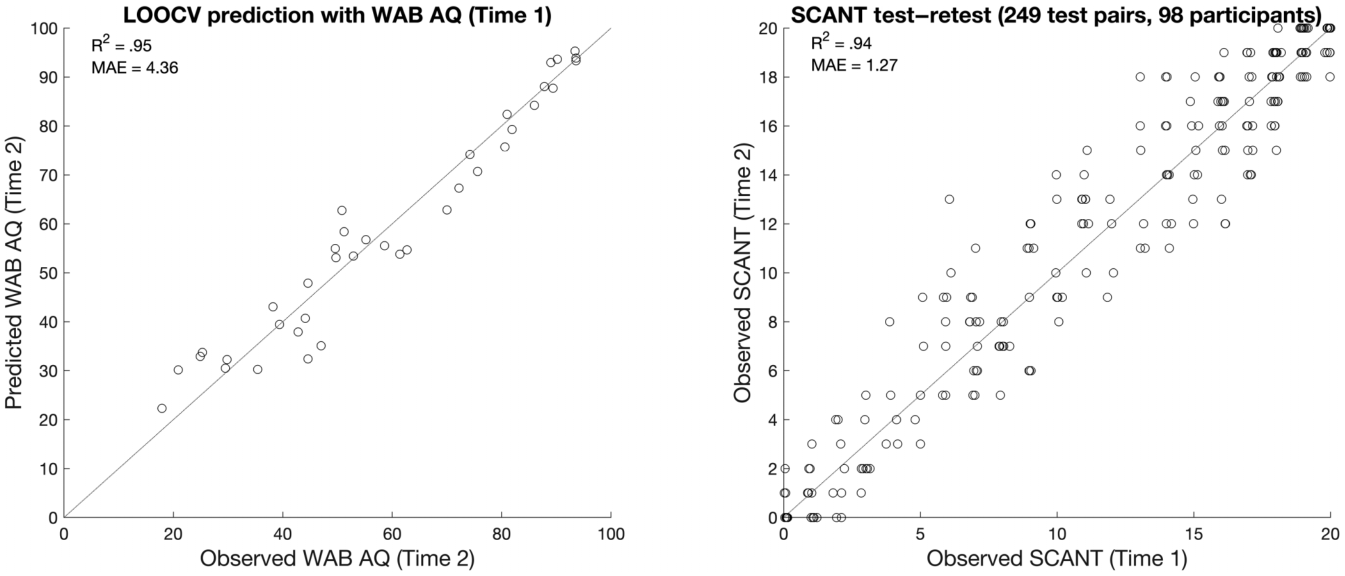


for comparison with predicting WAB AQ from a proxy measure such as the SCANT. To clarify, no statistical inferences in this study are based on this estimate of test–retest reliability for WAB AQ; it is simply provided for context. There was a significant difference in average WAB AQ between Time 1 and Time 2 for the POLAR participants (μ_1 = 57.94, μ_2 = 59.85, $t(38)$ = 2.25, p = .03), as expected due to the long interval between testing (that included therapy). Therefore, the relationship between observed WAB AQ scores at Time 2 and LOOCV predictions of WAB AQ scores at Time 2 based on WAB AQ scores at Time 1 are shown in the left panel of Figure 6, using the linear model to account for group average differences. There was no significant difference between test–retest pairs of SCANT scores (μ_1 = 11.95, μ_2 = 12.11, $t(248)$ = 1.44, p = .15), as expected due to the relatively short interval between administrations (approximately 1 week). The SCANT test–retest scores (i.e., Time 1 and Time 2) are shown in the right panel of Figure 6. Both WAB AQ and SCANT scores have high test–retest reliability (r^2 = .95 and .94, respectively). Notably, the test–retest reliability estimate for WAB AQ obtained in this study is in strong agreement with other estimates provided in the literature (Bond, 2019; Kertesz, 1979; Shewan & Kertesz, 1980).

The SCANT change percentile scores are presented in Table 3. An observed change of four items in an individual participant's SCANT scores over time would be larger than about 93% of observed SCANT test–retest scores. It would also correspond to an expected change of 0–23 WAB AQ points, based on a point estimate of 11.44 WAB AQ points, ± 1 *SD* of residual error for the difference in two, independent WAB AQ scores predicted from two, independent SCANT scores (i.e., each SCANT score combined with the linear model yields a normal distribution of predicted WAB AQ scores, and taking the difference of those WAB AQ score distributions leads to another normal distribution with standard deviation = $(2 \times 8.19^2)^{0.5}$ = 11.56 WAB AQ points, assuming the standard deviation of residuals from the holdout data set). In other words, a change of four SCANT items in an individual participant represents a larger-than-expected change in the SCANT score and a likely nonzero change in WAB AQ scores. Notably, the interval estimate for WAB AQ change is wide, because the error from two separate predictions of WAB AQ at different time points accrues. This means that estimates of the *magnitude* of WAB AQ change are not particularly reliable. While the magnitudes of SCANT score changes are reliably measurable given the high test–retest reliability, at present, the SCANT is only useful as a screening tool to determine whether or not an appreciable change in WAB AQ is likely. For investigators who require a precise estimate of the magnitude of WAB AQ change, administering the WAB itself would be the better option.

The benchmark of four items is suggested to achieve a reasonably high degree of confidence, around 93%, which may be relaxed when Type I error (i.e., inferring a real change occurred when it did not) is less of a concern. To understand the implications of using relaxed thresholds

Figure 6. Predicted longitudinal change in WAB AQ scores (left); the linear regression model is accounting for a significant average improvement (1.91 WAB AQ points) over time (approximately one to three years, including therapy). Test–retest changes in SCANT scores (right); no average change is expected or observed (approximately one-week interval, no therapy). A random jitter is applied to the *x*-axis for visualization of overlapping data points. The diagonal line is the identity line. LOOCV = leave-one-out cross-validation; WAB AQ = Western Aphasia Battery Aphasia Quotient; SCANT = Severity-Calibrated Aphasia Naming Test; MAE = mean absolute error.



for interpreting SCANT change scores, we compared using four-, three-, and two-item SCANT change score thresholds with using Fisher's exact tests and PNT change scores in response to the first phase of the POLAR therapy program. The SCANT had a 10%–33% reduction in the rate of identified improvements compared with the more powerful method, depending on the confidence level of the threshold, but these differences in improvement rates between categorization methods were not significant (p = .58, .81, and .86, respectively). The two methods of categorizing change scores agreed on 87%, 82%, and 75% of individual participants, respectively (see Table 4). Classification as improved or unimproved by one method was significantly related to the other method's classification, for each of the investigated threshold pairs (p = .0059, .0038, and .0027, respectively). Of course, there were some disagreements about which individual participants improved. Dice similarity coefficients for the sets of improved

Table 3. Changes in SCANT scores and the corresponding percentile (i.e., the proportion of test–retest pairs with a smaller difference than the observed change).

SCANT score change	Percentile
0	0
1	22.64
2	60.38
3	82.73
4	92.76
5	98.30
6–7	99.30
8	100.00

Note. Percentiles were constructed from 143 test–retest pairs sampled uniformly from the range of baseline SCANT scores. SCANT = Severity-Calibrated Aphasia Naming Test.

Table 4. Contingency tables comparing inference methods for identifying significant improvement in response to 3 weeks of aphasia therapy, with thresholds for significance shown in parentheses.

		SCANT (4 items)	
Threshold 1	Status	Unimproved	Improved
PNT (p < .07)	Unimproved	41	2
	Improved	5	4
		SCANT (3 items)	
		Unimproved	Improved
PNT (p < .17)	Unimproved	40	4
	Improved	6	6
		SCANT (2 items)	
		Unimproved	Improved
PNT (p < .40)	Unimproved	36	7
	Improved	9	11

Note. Participants were randomly assigned to semantically oriented or phonologically oriented therapy; the test items were not treated for naming. The inferences based on SCANT scores relied on a fixed threshold of change out of 20 items. The inferences based on PNT scores relied on a Fisher's exact test comparing frequencies of correct responses out of 174 items, with alpha criteria for significance matched to the corresponding SCANT threshold's test–retest percentile score to approximate the same purported Type I error rate. SCANT = Severity-Calibrated Aphasia Naming Test; PNT = Philadelphia Naming Test.

participants identified by each method were .53, .55, and .58 for each of the threshold pairs, respectively. We were unable to determine whether disagreements reflected noisy measurement leading to classification errors or reflect legitimate differences in measuring dissociable constructs (i.e., change in a specific picture naming test ability vs. change in a general language ability). All participants with SCANT change scores of 4 or greater also improved their PNT scores; two participants (out of 10) with SCANT change scores of 3 or greater did not improve their PNT score; and three participants (out of 18) with SCANT change scores of 2 or greater did not improve their PNT score. Given these results, there is clearly a reduction in sensitivity to therapy-induced change when using the SCANT compared with using a data set with about 9 times as many items, as expected; however, in participants with initial scores below the potential change threshold, the ability to detect improvement has reasonable agreement with the more powerful method, particularly considering the highly reduced administration time.

In the group of 13 participants with test–retest data available for WAB AQ and the SCANT, there was a strong correlation between the measures both at test and retest (r = .95 for both). However, because the scores did not significantly improve for the majority of this small group, the correlation between change scores was not expected to be significant (r = .28, p = .36). For most of the group, differences in scores mostly represented measurement noise. In the literature, a change of 5 points in WAB AQ has been identified as a clinically meaningful benchmark (Gilmore et al., 2019) and represents approximately 1 *SEM* (Hula et al., 2010). This means that, statistically speaking, we would expect this threshold to have a 32% false-positive rate. An approximately corresponding threshold for the SCANT would be two items, yielding a 40% false-positive rate. There were three participants (23%) with WAB AQ improvements of 5 or more points. There were four participants (31%) with SCANT improvements of two or more items. Two of the three participants with significantly improved WAB AQ also had significantly improved SCANT scores. There was one participant with a WAB AQ decrease of 5 or more points, and there were no participants with a SCANT score decrease of two or more items. In this preliminary investigation, the set of participants with significant change in SCANT scores and the set of participants with significant change in WAB AQ scores showed reasonable agreement (dice similarity coefficient = .50).

Discussion

We presented a 20-item naming test that has favorable psychometric properties for measuring the severity of speech impairment in aphasia and can serve as a proxy measure for a more rigorous evaluation of general speech impairment severity. The test may be particularly useful in the acute care setting, where clinicians must divide their time between assessment of aphasia and dysphagia, and assessment is focused on diagnosis, triage, and recommendations for progression along the continuum of care. For researchers or clinicians who are interested in the WAB AQ as an outcome measure, the SCANT score may provide an efficient alternative or supplement.

The coefficients of determination (R^2) between different measures of aphasia severity and WAB AQ are presented in Table 5, along with the sample sizes N used to estimate the coefficients. Multidomain assessments with different formulas for calculating the aphasia severity metric, such as QAB (Wilson et al., 2018) and Rasch WAB (Hula et al., 2010), have coefficients of determination that are only slightly below that of the WAB AQ retest itself (.88 and .90 compared to .95, respectively). The coefficient of determination between SCANT and WAB AQ (.83) is much closer to the coefficients of determination between WAB AQ and these alternative measures of aphasia severity than it is to the coefficient of determination between WAB AQ and another popular short-form naming test, the 15-item Boston Naming Test (Kaplan et al., 2001). This coefficient of determination (.65) was estimated (J. Richardson, personal communication, July 26, 2021) using data from AphasiaBank (MacWhinney et al., 2011) that were originally curated for a study of confrontation naming and story gist production (Richardson et al., 2018). Given its extremely high interrater reliability and short administration time, the SCANT may be a useful clinical tool for measurement of aphasia severity, instead of (or in addition to) multidomain assessments of language impairment profiles.

How does SCANT compare to other naming tests for aphasia in the literature, such as an adaptive naming test (Fergadiotis et al., 2015; Hula et al., 2019, 2015)? The creation of an adaptive test is based on Item Response Theory (Lord & Novick, 1968) and is a well-established

Table 5. Coefficients of determination (R^2) between different aphasia severity measures and the WAB AQ.

Measure	WAB AQ (R^2)	N	Source
WAB AQ (retest)	.95	39	Current data
Rasch WAB	.90	101	Hula et al. (2010)
QAB	.88	16	Wilson et al. (2018)
SCANT	.83	360	Current data
BNT-15	.65	258	Richardson et al. (2018)

Note. WAB = Western Aphasia Battery; AQ = Aphasia Quotient; QAB = Quick Aphasia Battery; SCANT = Severity-Calibrated Aphasia Naming Test; BNT-15 = 15-item Boston Naming Test.

approach to measuring mental abilities. An advantage of this approach is that, along with an estimate of naming ability, it provides a severity-dependent measure of uncertainty around that estimate. Due to measurement constraints near the floor or ceiling of the scale, the uncertainty around a latent ability (or a true score or a true change) near the extrema should be different than the uncertainty near the middle of the scale. The adaptive approach also permits the use of different test items (or different numbers of test items) during different evaluations that yield comparable quantitative estimates on the same scale.

An adaptive test must be calibrated as well, in order to select items for presentation based on a modeled latent trait. It is worth noting that the validity of the person ability estimates and item difficulty estimates critically depends on the correctness of the model's assumptions about the latent trait (i.e., naming ability). For example, there is convincing evidence that picture-naming ability in aphasia is not monolithic; people may be impaired on this task for different reasons (e.g., semantic vs. phonological impairment) and therefore may find different items to be difficult (Lambon Ralph et al., 2002; Walker et al., 2018, 2021). If the calibration cohort is biased toward a particular type of impairment, then the difficulty values of items will not be valid for people who have a different type of impairment, thereby skewing the ability estimates. While the calibration cohort is a critical consideration for the validity of SCANT scores as well, the use of cross-validation during item selection guided by a holistic comparison with WAB AQ mitigates these concerns. The Fergadiotis/Hula work calibrates the short form to approximate the score on the full set of PNT items. To the extent that this is successful, we would then expect it to perform as well as the full set of items when predicting WAB scores, which is, to say, quite well in general, but perhaps not quite as well as the SCANT in a large group comparison.

One major difference between the SCANT and an adaptive test would be the motivation for the stopping criterion. While an arbitrary limit of 20 items could be imposed on an adaptive test, typically the stopping criterion is motivated by a desired precision in the estimate of latent picture-naming ability. While this may provide a rationale for administering fewer test items in some situations, the desired level of precision is arbitrary, and its estimation is heavily based on modeling assumptions. Note that when we use the term *arbitrary*, we do not mean *uninformed*; we mean that a subjective standard is applied, which may be completely reasonable but subjective, nonetheless. These standards are arbitrary in the same way that a threshold for a significant *p* value is arbitrary. The SCANT's motivation for the stopping criterion is maximizing precision in the prediction of WAB AQ scores based entirely on observed score distributions. All of the theoretical work is done by assuming a shared measurement construct between picture naming and the WAB.

Similar to the previous point, the motivations behind identifying meaningful change with an adaptive test would be different from the SCANT. The focus on the precision of estimates of latent ability yields a more flexible system that can maintain a purported level of confidence while adapting to effects of measurement noise near the extrema of the scale (i.e., floor and ceiling). It can also potentially avoid practice effects by presenting different items. However, again, these inferences depend heavily on modeling assumptions. While the SCANT-based inferences about changes in WAB scores are also based on parametric modeling assumptions, they are supplemented by empirical distributions of percentile change scores for the SCANT itself. Put differently, the adaptive test compares observed changes to theoretically expected changes, whereas the SCANT compares observed changes to other observed changes that are associated with theoretically expected changes. In general, we would expect extremely high concurrence between measurements or inferences based on the SCANT and an adaptive naming test. While the adaptive test may be more appropriate for people with very high or very low abilities, the SCANT offers an objective standard for test length, optimizes concurrent validity with WAB AQ, and can be administered without a computer.

The WAB subtests provide more fine-grained information than just the AQ, with notable variance in subscores for comprehension and fluency. The SCANT was designed to predict AQ, and so it is not expected, on its own, to provide a detailed evaluation of the particular pattern of language subabilities that underlie a given patient's overall aphasia severity. However, given that the WAB AQ is derived from the various subtest scores and given that the SCANT is a good predictor of the WAB AQ, it is likely that SCANT (and naming tasks more generally) is picking up information measured by the WAB subtests. Indeed, Table 6 gives the coefficients of determination between each aphasia severity measure (AQ or SCANT) and each WAB subtest, estimated from the 177 participants in the LESMAP and POLAR studies (these data were not available for the MAPPD participants). Because the AQ is calculated directly from the subtest scores, we expect strong correlations between these variables. However, the SCANT also exhibits reasonably high correlations with WAB subtests, particularly in the expressive domain. Three subtest scores (Object Naming, Speech Repetition, and Spontaneous Speech Information Content) shared between 72% and 77% of their variance with SCANT scores. Four more subtest scores (Responsive Speech, Word Fluency, Sentence Completion, and Spontaneous Speech Fluency) shared between 55% and 68% of their variance with SCANT scores. Like the AQ, SCANT

Table 6. Coefficients of determination (R^2) between each aphasia severity measure (AQ or SCANT) and each WAB subtest, estimated from the 177 participants in the LESMAP and POLAR studies.

Measure	IC	FLU	Yes/no questions	Aud. word recognition	Sequential commands	Repetition	Object naming	Word fluency	Sentence completion	Responsive speech
AQ	.87	.72	.39	.57	.59	.91	.86	.65	.73	.81
SCANT	.72	.55	.30	.41	.40	.74	.77	.62	.61	.68

Note. AQ = Aphasia Quotient; SCANT = Severity-Calibrated Aphasia Naming Test; WAB = Western Aphasia Battery; LESMAP = lesion mapping in aphasia; POLAR = predictors of language outcomes after rehabilitation; IC = spontaneous speech information content; FLU = spontaneous speech fluency.

scores shared the least amount of variance with comprehension subtest scores; nevertheless, these subtest scores (Auditory Word Recognition, Sequential Commands, and Yes/No Questions) still shared between 30% and 41% of their variance with SCANT scores. This, in turn, raises the possibility that a richer analysis of performance on the SCANT (e.g., error type analysis) or a naming test with a different set of items might be capable of predicting finer grained features of aphasia. Ideally, outcome measures should be sensitive to the type of change that is targeted by therapy, but there can be practical constraints. When selecting outcome measures, the practitioner's goals must be considered in trying to strike a balance between the diversity of assessments and the cost and efficiency of testing participants or clients.

A change of 5 points in WAB AQ has been suggested in the literature as a benchmark for clinically significant change. There have been at least three justifications provided for this benchmark: (a) A reasonable clinician would agree that a change of 5 points is meaningful for evaluating the course of the impairment (Elman & Bernstein-Ellis, 1999; Katz & Wertz, 1997). (b) A meta-analysis of published treatment studies assessing within-group effect size (i.e., change in pretreatment to posttreatment WAB AQ scores) and between-groups effect size (i.e., the difference in WAB AQ scores between a treatment and a control group) both indicated an average difference of 5 points (Dekhtyar et al., 2020; Gilmore et al., 2019). (c) The *SEM* for WAB AQ (i.e., the standard deviation of repeatedly observed scores around an individual's true score) has been estimated to be 5 points (Hula et al., 2010; Shewan & Kertesz, 1980). How does this compare with the SCANT benchmarks? Importantly, we found that the SCANT score and the WAB AQ score had nearly identical test–retest reliabilities as measured by Pearson correlation, the primary statistic driving the calculation of *SEM*. So, to the extent that a clinician feels confident in the reliability of WAB AQ scores and their observed changes, they can also interpret SCANT scores and their changes similarly. For example, if a change of 5 points in WAB AQ is considered clinically meaningful, the estimated *SEM* suggests that, statistically speaking, there is a 32% chance of observing this large of a change when no change in the true score occurred. A corresponding benchmark for the SCANT would be two items, indicating a 40% chance of observing this large of a change when no meaningful change is expected. If the cost of ignoring a real change is greater than the cost of treating a stable person, then this may be a perfectly acceptable false-positive rate in a clinical setting. In our preliminary investigation, we found reasonable agreement between these two benchmarks for classifying meaningful improvement. A change of four items on the SCANT, however, provides a high degree of confidence (93%) that a true change occurred in an individual, corresponding to approximately 2 *SEM*, for both SCANT scores and WAB AQ. An individual gain of this size is infrequent (about 10% of participants in the POLAR study) but clearly attainable.

Regardless of whether a naming test can or should replace a multidomain assessment for aphasia, the SCANT provides a useful tool for those interested in quantifying the severity of anomia symptoms in the context of aphasia. The principled selection of items based on the responses of hundreds of participants with aphasia yields an instrument that is properly calibrated to the range of abilities found in the aphasia population. Observed scores and changes on the SCANT scale can be interpreted in relation to meaningful, independent measurements of language ability.

Limitations

It is important to note that most of the primary limitations of the WAB are inherited by the SCANT. The WAB provides a poor measure of the interactive aspects of language use, including nonlinguistic, functional communication, and provides limited insight into how clients use language and communicate in other settings outside the clinical evaluation. The same is true of the SCANT. It is also important to emphasize that a single number, such as a WAB AQ or a SCANT score, can only provide a limited view of the complex syndromes that accompany a diagnosis of aphasia. While the SCANT can detect aphasia in most patients who receive a clinical diagnosis, it will likely miss language deficits in those who are also classified by the WAB as not having aphasia, such as those with pure agrammatism. It will also miss or underestimate severity in people with aphasia who have selective impairment in naming actions/verbs relative to objects/nouns (Miceli et al., 1988). Lastly, aphasias that present primarily with mild impairments in verbal working memory or speech comprehension (e.g., Francis et al., 2010) are unlikely to be detected by the SCANT. However, with these caveats, the SCANT is likely to be a clinically useful screening tool for aphasia. It does require further investigation in aphasic individuals across cultures, both within and outside the United States, before recommending for widespread clinical use.

The predictive accuracy reported for the model can only be expected to extend to new participants that are well represented by the cohort included in the current study. The included participants were all American English speakers with confirmed left-hemisphere stroke. Notably, the lowest possible WAB AQ scores (< 15.8) were not represented in our cohort. Participants exhibiting ceiling or floor effects warrant further evaluation; evaluation of these participants with the SCANT may not be appropriate.

The current results for the SCANT are based on responses that were extracted from a larger set. It is currently unknown how the surrounding items may have influenced the accuracy of responses (e.g., through cumulative semantic interference or fatigue), although it is expected that these influences are negligible in the general aphasia population. Previous "simulation" studies that extracted short-form naming tests in a similar fashion have yielded comparable results in validation studies (Hula et al., 2019, 2015; Walker & Schwartz, 2012).

While we have demonstrated that the SCANT items work well for predictions of aphasia severity within our study cohort and we expect the items to be broadly recognizable across the English-speaking world, this particular set of items may not be optimal for predictions within all cultures at all times. Reassuringly, partial correlations between race and SCANT scores while controlling for WAB AQ did not reveal any significant associations within the three study cohorts (p = .84, .42, and .99 for MAPPD, LESMAP, and POLAR, respectively). To be used in countries other than the United States, alternative responses would need to be scored as correct (e.g., "torch" for *flashlight* and "waistcoat" for *vest*); however, these morphological changes may render these items less useful for prediction. Aside from dialectal variation, an item such as *football* may not be as recognizable across international boundaries, whereas an item such as *typewriter* may not be as recognizable across intergenerational boundaries. While this certainly presents a challenge for cross-cultural language assessments, it is worth considering that the WAB may also include culturally specific material, thereby justifying the inclusion of these potentially problematic items as an accurate reflection of the standard measure. While the specific utility that each pictured item has for tapping into an individual's language competency is expected to vary across cultures, the SCANT demonstrates the feasibility of constructing calibrated naming tests from large collections of data from people with aphasia.

Conclusions

If the clinical or scientific goal is to measure aphasia severity and if there are practical constraints limiting the feasibility of a comprehensive aphasia assessment, the SCANT provides an alternative, enabling an aphasia severity estimate that approximates the WAB AQ within about 5 min. For those interested in measuring anomia symptoms in the context of stroke aphasia, the SCANT provides a calibrated scale with favorable psychometric properties.

Data Availability Statement

All data analyzed in this work are public and archived.

Data from the MAPPD study are available at https://osf.io/8xzdc/?view_only=7a8db6248e53453d84d3f1e55ca06868

Data from the LESMAP study are available at https://www.cogsci.uci.edu/~alns/MPTfit.php

Data from the POLAR study are available at https://asha.figshare.com/articles/journal_contribution/Measuring_change_in_picture_naming_ability_Walker_et_al_2021_/17019515

Acknowledgments

This work was supported by National Institute on Deafness and Communication Disorders Grant P50DC014664-06A1 awarded to Julius Fridriksson.

References

Armstrong, B. G. (1996). Optimizing power in allocating resources to exposure assessment in an epidemiologic study. *American Journal of Epidemiology, 144*(2), 192–197. https://doi.org/10.1093/oxfordjournals.aje.a008908

Bernstein-Ellis, E., Higby, E., & Gravier, M. (2021, May 4). Responding to culturally insensitive test items. *The ASHA LeaderLive*. https://leader.pubs.asha.org/do/10.1044/leader.AE.26052021.26/full/

Bloom, L. (1976). One word at a time. In *One word at a time*. De Gruyter Mouton. https://doi.org/10.1515/9783110819090

Bond, B. (2019). *The test–retest reliability of the Western Aphasia Battery–Revised*. University of Kansas. http://hdl.handle.net/1808/30075

Bruce, C., & Edmundson, A. (2009). Letting the CAT out of the bag: A review of the Comprehensive Aphasia Test. Commentary on Howard, Swinburn, and Porter. *Aphasiology, 24*(1), 79–93. https://doi.org/10.1080/02687030802453335

Calero, M. D., Arnedo, M. L., Navarro, E., Ruiz-Pedrosa, M., & Carnero, C. (2002). Usefulness of a 15-item version of the Boston Naming Test in neuropsychological assessment of low-educational elders with dementia. *The Journals of Gerontology: Series B, 57*(2), P187–P191. https://doi.org/10.1093/GERONB/57.2.P187

Caramazza, A., & Hillis, A. E. (1989). The disruption of sentence production: Some dissociations. *Brain and Language, 36*(4), 625–650. https://doi.org/10.1016/0093-934X(89)90091-6

Conroy, P., Sotiropoulou Drosopoulou, C., Humphreys, G. F., Halai, A. D., & Lambon Ralph, M. A. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain, 141*(6), 1815–1827. https://doi.org/10.1093/BRAIN/AWY087

Corina, D. P., Loudermilk, B. C., Detwiler, L., Martin, R. F., Brinkley, J. F., & Ojemann, G. (2010). Analysis of naming errors during cortical stimulation mapping: Implications for models of language representation. *Brain and Language, 115*(2), 101–112. https://doi.org/10.1016/J.BANDL.2010.04.001

Cunningham, K. T., & Haley, K. L. (2020). Measuring lexical diversity for discourse analysis in aphasia: Moving-average type–token ratio and word information measure. *Journal of Speech, Language, and Hearing Research, 63*(3), 710–721. https://doi.org/10.1044/2019_JSLHR-19-00226

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

Dede, G., & Salis, C. (2019). Temporal and episodic analyses of the story of Cinderella in latent aphasia. *American Journal of Speech-Language Pathology, 29*(1S), 449–462. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0210

Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., & Kiran, S. (2020). Videoconference administration of the Western Aphasia Battery–Revised: Feasibility and validity. *American Journal of Speech-Language Pathology, 29*(2), 673–687. https://doi.org/10.1044/2019_AJSLP-19-00023

DeLeon, J., Gottesman, R. F., Kleinman, J. T., Newhart, M., Davis, C., Heidler-Gary, J., Lee, A., & Hillis, A. E. (2007). Neural regions essential for distinct cognitive processes underlying picture naming. *Brain, 130*(5), 1408–1422. https://doi.org/10.1093/brain/awm011

Dore, J. (1975). Holophrases, speech acts and language universals. *Journal of Child Language, 2*(1), 21–40. https://doi.org/10.1017/S0305000900000878

El Hachioui, H., Visch-Brink, E. G., de Lau, L. M. L., van de Sandt-Koenderman, M. W. M. E., Nouwens, F., Koudstaal, P. J., & Dippel, D. W. J. (2017). Screening tests for aphasia in patients with stroke: A systematic review. *Journal of Neurology, 264*(2), 211–220. https://doi.org/10.1007/s00415-016-8170-8

Elman, R. J., & Bernstein-Ellis, E. (1999). The efficacy of group communication treatment in adults with chronic aphasia. *Journal of Speech, Language, and Hearing Research, 42*(2), 411–419. https://doi.org/10.1044/JSLHR.4202.411

Enderby, P. (1997). *Therapy outcome measures*. Singular.

Evans, W. S., Cavanaugh, R., Gravier, M. L., Autenreith, A. M., Doyle, P. J., Hula, W. D., & Dickey, M. W. (2021). Effects of semantic feature type, diversity, and quantity on semantic feature analysis treatment outcomes in aphasia. *American Journal of Speech-Language Pathology, 30*(1S), 344–358. https://doi.org/10.1044/2020_AJSLP-19-00112

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2018). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology, 33,* 544–560. https://doi.org/10.1080/02687038.2018.1482404

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58*(3), 865–877. https://doi.org/10.1044/2015_JSLHR-L-14-0249

Francis, D., Clark, N., & Humphreys, G. (2010). The treatment of an auditory working memory deficit and the implications for sentence comprehension abilities in mild "receptive" aphasia. *Aphasiology, 17,* 723–750. https://doi.org/10.1080/02687030344000201

Fridriksson, J., Rorden, C., Elm, J., Sen, S., George, M. S., & Bonilha, L. (2018). Transcranial direct current stimulation vs sham stimulation to treat aphasia after stroke: A randomized clinical trial. *JAMA Neurology, 75*(12), 1470–1476. https://doi.org/10.1001/jamaneurol.2018.2287

Fromm, D., Forbes, M., Holland, A., Dalton, S. G., Richardson, J., & MacWhinney, B. (2017). Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff. *American Journal of Speech-Language Pathology, 26*(3), 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071

Gilmore, N., Dwyer, M., & Kiran, S. (2019). Benchmarks of significant change after aphasia rehabilitation. *Archives of Physical Medicine and Rehabilitation, 100*(6), 1131–1139.e87. https://doi.org/10.1016/J.APMR.2018.08.177

Howard, D., Patterson, K., Franklin, S., Orchard-lisle, V., & Morton, J. (1985). Treatment of word retrieval deficits in aphasia: A comparison of two therapy method. *Brain, 108*(4), 817–829. https://doi.org/10.1093/brain/108.4.817

Howard, D., Swinburn, K., & Porter, G. (2009). Putting the CAT out: What the Comprehensive Aphasia Test has to offer. *Aphasiology, 24*(1), 56–74. https://doi.org/10.1080/02687030802453202

Hula, W. D., Donovan, N. J., Kendall, D. L., & Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology, 24*(11), 1326–1341. https://doi.org/10.1080/02687030903422502

Hula, W. D., Fergadiotis, G., Swiderski, A. M., Silkes, J. P., & Kellough, S. (2019). Empirical evaluation of computer-adaptive alternate short forms for the assessment of anomia severity. *Journal of Speech, Language, and Hearing Research, 63*(1), 163–172. https://doi.org/10.1044/2019_JSLHR-L-19-0213

Hula, W. D., Kellough, S., & Fergadiotis, G. (2015). Development and simulation testing of a computerized adaptive version of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58*(3), 878–890. https://doi.org/10.1044/2015_JSLHR-L-14-0297

Joe, S., & Kuo, F. Y. (2003). Remark on algorithm 659. *ACM Transactions on Mathematical Software, 29*(1), 49–57. https://doi.org/10.1145/641876.641879

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test*. Lea & Febiger.

Katz, R. C., & Wertz, R. T. (1997). The efficacy of computer-provided reading treatment for chronic aphasic adults. *Journal of Speech, Language, and Hearing Research, 40*(3), 493–507. https://doi.org/10.1044/JSLHR.4003.493

Kertesz, A. (1979). *Aphasia and associated disorders: Taxonomy, localization, and recovery*. Holt Rinehart & Winston.

Kertesz, A. (1982). *The Western Aphasia Battery test manual*. The Psychological Corporation.

Kertesz, A. (2007). *Western Aphasia Battery–Revised examiner's manual*. Pearson Education.

Kertesz, A. (2022). The Western Aphasia Battery: A systematic review of research and clinical applications. *Aphasiology, 36*(1), 21–50. https://doi.org/10.1080/02687038.2020.1852002

Kertesz, A., & Phipps, J. B. (1977). Numerical taxonomy of aphasia. *Brain and Language, 4*(1), 1–10. https://doi.org/10.1016/0093-934X(77)90001-3

Kertesz, A., & Poole, E. (1974). The aphasia quotient: The taxonomic approach to measurement of aphasic disability. *Canadian Journal of Neurological Sciences, 1*(1), 7–16. https://doi.org/10.1017/S031716710001951X

Kiran, S., Cherney, L. R., Kagan, A., Haley, K. L., Antonucci, S. M., Schwartz, M., Holland, A. L., & Simmons-Mackie, N. (2018). Aphasia assessments: A survey of clinical and research settings, *32*(Suppl. 1), 47–49. https://doi.org/10.1080/02687038.2018.1487923

Kohn, S. E., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and Language, 24*(2), 266–283. https://doi.org/10.1016/0093-934X(85)90135-X

Lambon Ralph, M. A., Moriarty, L., & Sage, K. (2002). Anomia is simply a reflection of semantic and phonological impairments: Evidence from a case-series study. *Aphasiology, 16*(1–2), 56–82. https://doi.org/10.1080/02687040143000448

Leonard, C., Rochon, E., & Laird, L. (2008). Treating naming impairments in aphasia: Findings from a phonological components analysis treatment. *Aphasiology, 22*(9), 923–947. https://doi.org/10.1080/02687030701831474

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

MacOir, J., Chagnon, A., Hudon, C., Lavoie, M., & Wilson, M. A. (2021). TDQ-30—A new color picture-naming test for the diagnostic of mild anomia: Validation and normative data in Quebec French adults and elderly. *Archives of Clinical Neuropsychology, 36*(2), 267–280. https://doi.org/10.1093/ARCLIN/ACZ048

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

Martin, N., Schlesinger, J., Obermeyer, J., Minkina, I., & Rosenberg, S. (2020). Treatment of verbal short-term memory abilities to improve language function in aphasia: A case series treatment study. *Neuropsychological Rehabilitation, 31*(5), 731–772. https://doi.org/10.1080/09602011.2020.1731554

McNeill, D. (1970). *The acquisition of language: The study of developmental psycholinguistics*. Harper & Row.

Miceli, G., Mazzucchi, A., Menn, L., & Goodglass, H. (1983). Contrasting cases of Italian agrammatic aphasia without comprehension disorder. *Brain and Language, 19*(1), 65–97. https://doi.org/10.1016/0093-934X(83)90056-1

Miceli, G., Silveri, M. C., Nocentini, U., & Caramazza, A. (1988). Patterns of dissociation in comprehension and production of nouns and verbs. *Aphasiology, 2*(3–4), 351–358. https://doi.org/10.1080/02687038808248937

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology, 27*(6), 495–504. https://doi.org/10.1080/02643294.2011.574112

Nespoulous, J. L., Dordain, M., Perron, C., Ska, B., Bub, D., Caplan, D., Mehler, J., & Lecours, A. R. (1988). Agrammatism in sentence production without comprehension deficits: Reduced availability of syntactic structures and/or of grammatical morphemes? A case study. *Brain and Language, 33*(2), 273–295. https://doi.org/10.1016/0093-934X(88)90069-7

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *Journal of Anatomy, 197*(3), 335–359. https://doi.org/10.1046/j.1469-7580.2000.19730335.x

Raymer, A. M., Thompson, C. K., Jacobs, B., & Le Grand, H. R. (1993). Phonological treatment of naming deficits in aphasia model-based generalization analysis. *Aphasiology, 7*(1), 27–53. https://doi.org/10.1080/02687039308249498

Richardson, J. D., Hudspeth Dalton, S. G., Fromm, D., Forbes, M., Holland, A., & MacWhinney, B. (2018). The relationship between confrontation naming and story gist production in aphasia. *American Journal of Speech-Language Pathology, 27*(1S), 406–422. https://doi.org/10.1044/2017_AJSLP-16-0211

Risser, A. H., & Spreen, O. (1985). The Western Aphasia Battery. *Journal of Clinical and Experimental Neuropsychology, 7*(4), 463–470. https://doi.org/10.1080/01688638508401277

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. R. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology, 24,* 121–133. http://eprints-prod-05.library.pitt.edu/215/1/24-09.pdf

Saravani, A. G., Forseth, K. J., Tandon, N., & Pitkow, X. (2019). Dynamic brain interactions during picture naming. *eNeuro, 6*(4). https://doi.org/10.1523/ENEURO.0472-18.2019

Shewan, C. M., & Kertesz, A. (1980). Reliability and validity characteristics of the Western Aphasia Battery (WAB). *Journal of Speech and Hearing Disorders, 45*(3), 308–324. https://doi.org/10.1044/jshd.4503.308

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics, 7*(4), 86–112. https://doi.org/10.1016/0041-5553(67)90144-9

Spell, L. A., Richardson, J. D., Basilakos, A., Stark, B. C., Teklehaimanot, A., Hillis, A. E., & Fridriksson, J. (2020). Developing, implementing, and improving assessment and treatment fidelity in clinical aphasia research. *American Journal of Speech-Language Pathology, 29*(1), 286–298. https://doi.org/10.1044/2019_AJSLP-19-00126

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive Aphasia Test (CAT)*. APA PsycTests. https://doi.org/10.1037/t13733-000

The MathWorks, Inc. (2021). *MATLAB R2021b*.

Tierney-Hendricks, C., Schliep, M. E., & Vallila-Rohter, S. (2021). Using an implementation framework to survey outcome measurement and treatment practices in aphasia. *American Journal of Speech-Language Pathology, 31*(3), 1133–1162. https://doi.org/10.1044/2021_AJSLP-21-00101

Trupe, A. E. (1984). Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification. In R. H. Brookshire (Ed.), *Clinical aphasiology* (pp. 55–69). BRK. http://aphasiology.pitt.edu/796/1/14-08.pdf

Walker, G. M., Basilakos, A., Fridriksson, J., & Hickok, G. (2021). Beyond percent correct: Measuring change in individual picture naming ability. *Journal of Speech, Language, and Hearing Research, 65*(1), 215–237. https://doi.org/10.1044/2021_JSLHR-20-00205

Walker, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological Assessment, 30*(6), 809–826. https://doi.org/10.1037/pas0000529

Walker, G. M., & Schwartz, M. F. (2012). Short-form Philadelphia Naming Test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology, 21*(2), S140–S153. https://doi.org/10.1044/1058-0360(2012/11-0089)

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

Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE, 13*(2), Article e0192773. https://doi.org/10.1371/JOURNAL.PONE.0192773

Yourganov, G., Smith, K. G., Fridriksson, J., & Rorden, C. (2015). Predicting aphasia type from brain damage measured with structural MRI. *Cortex, 73,* 203–215. https://doi.org/10.1016/J.CORTEX.2015.09.005

Appendix A

SCANT Items

1.	well
2.	banana
3.	king
4.	calendar
5.	typewriter
6.	saw
7.	bone
8.	football
9.	owl
10.	hose
11.	man
12.	foot
13.	lion
14.	nose
15.	baby
16.	hat
17.	vest
18.	helicopter
19.	rope
20.	queen

Appendix B

Potential Severity-Calibrated Aphasia Naming Test Items

The SCANT item selection procedure was repeated 5 times to assess reliability. Values for the log lexical frequency in movie and television transcripts, the number of phonological segments, and the log phonological neighborhood density (i.e., the number of words that can be created by adding, subtracting, or substituting a single phonological segment) were reported by Walker et al. (2018).

Item	Selection frequency (out of 5)	Log Lex Freq	No. of segments	Phon density
banana	5	5.78	6	0.69
bone	5	6.75	3	3.69
calendar	5	5.43	7	0.69
hat	5	7.45	3	3.89
king	5	7.43	3	3.58
man	5	8.99	3	3.87
well	5	9.01	3	3.69
foot	4	8.58	3	2.48
football	4	6.69	6	0.69
helicopter	4	6.11	9	0.00
hose	4	5.69	3	3.66
lion	4	5.79	4	2.64
monkey	3	6.66	5	2.30
queen	3	7.08	4	2.20
saw	3	8.77	2	3.47
baby	2	8.53	4	2.30
balloon	2	5.56	5	1.95
clock	2	7.40	4	2.94
comb	2	5.40	3	3.43
goat	2	5.82	3	3.61
nose	2	7.66	3	3.66
nurse	2	7.00	3	2.77
owl	2	4.92	2	3.04
pineapple	2	4.36	7	0.00
tree	2	7.39	3	3.14
typewriter	2	4.52	7	0.00
apple	1	6.54	4	1.95
cannon	1	5.54	5	1.95
church	1	7.20	3	1.79
corn	1	6.00	4	3.26
ear	1	7.78	2	3.04
fireman	1	4.65	6	0.69
flashlight	1	5.35	7	0.00
hand	1	8.62	4	3.09
house	1	8.71	3	2.71
letter	1	7.47	4	3.26
rake	1	4.47	3	3.66
ring	1	7.64	3	3.74
rope	1	6.48	3	3.47
vest	1	5.31	4	3.14

Note. Values for the log lexical frequency in movie and television transcripts, the number of phonological segments, and the log phonological neighborhood density (i.e., the number of words that can be created by adding, subtracting, or substituting a single phonological segment) were reported by Walker et al. (2018). SCANT = Severity-Calibrated Aphasia Naming Test; Lex Freq = lexical frequency.