ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Assessing Relative Linguistic Impairment With Model-Based Item Selection

Grant M. Walker,[a] Julius Fridriksson,[b] and Gregory Hickok[a,c]

[a] Department of Cognitive Sciences, University of California, Irvine [b] Department of Communication Sciences and Disorders, University of South Carolina, Columbia [c] Department of Language Science, University of California, Irvine

ARTICLE INFO

Article History:
Received July 22, 2023
Revision received November 30, 2023
Accepted May 3, 2024

Editor-in-Chief: Julie A. Washington
Editor: Stephen M. Wilson

https://doi.org/10.1044/2024_JSLHR-23-00439

ABSTRACT

Purpose: A picture naming test is presented that reveals impairment to specific mechanisms involved in the naming process, using accuracy scores on curated item sets. A series of psychometric validation experiments are reported.
Method: Using a computational model that enables estimation of item difficulty at the lexical and sublexical stages of word retrieval, two complimentary sets of items were constructed that challenge the respective psycholinguistic levels of representation. The difference in accuracy between these item sets yields the relative linguistic impairment (RLI) score. In a cohort of 91 people with chronic left-hemisphere stroke who enrolled in a clinical trial for anomia, we assessed psychometric properties of the RLI score and then used the new scale to make predictions about other language behaviors, lesion distributions, and functional activation during naming.
Results: RLI scores had adequate psychometric properties for clinical significance. RLI scores contained predictive information about spontaneous speech fluency, over and above accuracy. A dissociation was observed between performance on the RLI item sets and performance on the subtests of an independent language battery. Sublexical RLI was significantly associated with apraxia of speech and with lesions encompassing perisylvian regions, while lexical RLI was associated with lesions to deep white matter. The RLI construct was reflected in functional brain activity during naming, independent of overall accuracy, with a respective shift of activation between dorsal and ventral networks responsible for different aspects of word retrieval.
Conclusion: The RLI assessment satisfies the psychometric requirements to serve as a useful clinical measure.

Naming deficits are by far the most common symptom of aphasia (Azhar et al., 2017), and accuracy scores on picture naming tests provide a good proxy for overall severity of language impairment (Walker, Fridriksson, et al., 2022). For this reason, picture naming is commonly used as a clinical tool to assess language function for stroke aphasia recovery prognosis (Kristinsson et al., 2023; Osa García et al., 2020), treatment outcome studies (ALHarbi et al., 2017; Cotelli et al., 2020; Fridriksson et al., 2018; Pagnoni et al., 2021), disease progression in primary progressive aphasia (Hillis et al., 2004; Hurley et al., 2012), and presurgical functional brain mapping (Cervenka et al., 2011; Sinai et al., 2005).

Picture naming, however, is a deceptively complex, multistage task (Dell et al., 1997; Levelt, 2001; Matti et al., 1998). This is reflected in different error types on picture naming tasks, such as producing phonological, semantic, or mixed errors ("can," "dog," or "rat," respectively, for the target "cat"), among others. There have been at least three approaches to dealing with the multidimensional nature of the picture naming task in a clinical setting. The first approach has been to accept it as a nuisance confound and just evaluate accuracy scores anyway. The benefit of this approach is a unidimensional score that can be simply interpreted or further analyzed with the many statistical tools available for such measures. One

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.*

such example is the Severity-Calibrated Aphasia Naming Test (SCANT; Walker, Fridriksson, et al., 2022). This naming test consists of a set of 20 items that were selected to maximize the correlation between accuracy scores and the Western Aphasia Battery–Aphasia Quotient (WAB-AQ; Kertesz, 2007), an overall measure of aphasia severity. The SCANT is a fixed set of items that must be administered in full to obtain a valid score, which still inherits many of the drawbacks of the WAB-AQ score. Another example of unidimensional naming accuracy scores involves item response models, which enable the ranking of test items from least to most difficult based on the performance of a calibration cohort (Fergadiotis et al., 2015). To the extent that these rankings are valid, they enable the construction of comparable item sets, so that different test takers can see different items but produce comparable scores on the same measurement scale (Hula et al., 2019). If an examiner is only concerned about the overall severity of a deficit and not its source, then this may be an appropriate measure; however, the degree to which the unidimensionality assumption is violated by the actual data will increase the chances of obtaining a misleading result.

To evaluate the source of naming impairments, a second approach to data analysis has been to consider the multidimensional nature of the task by analyzing unidimensional response type rates separately, for example, semantic errors or nonword errors, assuming a one-to-one mapping between a response type and a mental subprocess. These specific error types are thought to reflect relatively clear cognitive failures during the naming process (i.e., lexical or sublexical processing, respectively). Researchers have had some success in mapping the frequency of different error types onto different neural regions, supporting the hypothesis that different levels of processing during naming rely on different neural networks (Dell et al., 2013; Halai et al., 2017; Schwartz et al., 2009). However, error type scoring introduces complications both for making inferences about the underlying components of the language system and for clinical assessment. For example, categorization of errors as one type or another can be the following:

- Ambiguous: For example, is "rat" for "cat" a phonological, semantic, or jointly determined error?
- Arbitrarily defined: For example, how many phonemes must the error and target share to count as phonologically related versus unrelated?
- Subjective: Interrater reliability can vary.
- Functionally underdetermined: There is no one-to-one relation between error type and level of processing in that disruption of different components of the naming system (or their interaction) can lead to the same error type.

Furthermore, some error types may occur infrequently, requiring many trials to reliably observe them. While previous investigations have found that naming accuracy can be reliably measured with short naming tests, naming error patterns were only reliable over a large set of items (e.g., 175 items) and were unreliable in set sizes typical of clinical naming tests (e.g., 30 or 60 items; Swiderski et al., 2023; Walker & Schwartz, 2012). These drawbacks have motivated the use of multivariate methods for naming data analysis.

The third approach to handling the complex nature of the naming task utilizes cognitive modeling. In this paradigm, like in the error type analysis approach, the multivariate distribution of naming response types is assumed to reflect success or failure of latent cognitive processes, such as word retrieval or form construction; however, rather than assuming a one-to-one correspondence between error types and mental processes, the entire distribution of errors is considered simultaneously to infer the most likely cognitive impairments. A cognitive model can consider all response types, including ones that may be multiply determined, to estimate the chance of failure for specific cognitive processes. This allows the model to overcome several of the previously noted disadvantages of error type analysis, including ambiguity, functional underdetermination, and relative paucity of certain response types (although definitions of response categories and interrater reliability remain important considerations for any behavioral measurement endeavor). One example of this approach is a spreading activation model that specifies a three-layer network of interconnected units representing the semantic, lexical, and phonological features of a lexicon (Foygel & Dell, 2000). By adjusting the amount of activation that spreads between semantic and lexical units (S-weight) or between lexical and phonological units (P-weight), different error patterns emerge, and the settings that lead to the most similar responses to those produced by an individual with a naming impairment are assumed to reflect the individual's underlying strengths and deficits. The spreading activation model has been used to investigate interactivity between hierarchical psycholinguistic representations (Rapp & Goldrick, 2000), picture naming's relation to other language test scores (Dell et al., 2007), lesion-deficit relationships (Dell et al., 2013; Hula et al., 2020), clinical diagnosis (Abel et al., 2009), and anomia treatment response (Simic et al., 2020). On the other hand, the spreading activation model must make several simplifying assumptions to make the connection strength estimation problem tractable, limiting its generalizability. Perhaps the most notable simplification is the assumption that all words in English have the same propensities to elicit different error types.

Another example of the cognitive modeling approach is the multinomial processing tree (MPT) model (Walker et al., 2018). Rather than simulating a naming attempt with a dynamic model, the MPT model directly estimates the probabilities of selecting different latent mental representations at the word and phoneme levels that lead to different response types. Using data from over 350 people with poststroke aphasia, we developed and validated an MPT model of naming that incorporates item difficulty and allows one to estimate the degree of disruption to six different subabilities in the naming process. We demonstrated that the information contained in the MPT model's item difficulty estimates improved predictions of future naming outcomes and other behavioral test scores over models that lacked consideration of item variability, including the spreading activation model (Walker et al., 2021). In principle, such a tool could be useful clinically, enabling therapists to pinpoint the source(s) of difficulty and target them for treatment (Walker, 2021) and to provide a more sensitive metric for recovery (Walker, Basilakos, et al., 2022). However, the MPT model is clinically cumbersome, requiring, in its current form, categorization of each response into one of eight types and then entering each response type for each item into a text-based software program that enables Monte Carlo Markov Chain estimation and requires a moderate level of computer programming skill to use flexibly.

The MPT model, however, also suggests a fourth approach to extracting meaningful information from picture naming scores. The suggested approach emerges out of an attempt to validate the model's item difficulty estimates. The MPT model assumes that items vary in difficulty for different cognitive processes required for production. Of course, all the supposedly dissociable processes would still be required for each naming attempt, but there may be a difference in the relative degree of reliance upon them based on the requirements of the target item. If it is truly the case that there are multiple cognitive processes required for picture naming, then individuals should exist who are relatively more impaired in one of those processes. When such an individual is presented with items that specifically challenge the impaired process, naming accuracy should be reduced relative to items that challenge the stronger processes. By examining the difference in accuracy scores from groups of items selected to specifically challenge different cognitive processes, it should be possible to estimate the relative difference in impairments to these different processes. Rather than informing about the overall severity of the impairment (which is still expected to yield highly correlated accuracy scores between different sets of items), the modulations in accuracy scores between model-selected sets of items can inform about the relative contributions of different cognitive sources to the impairment, without requiring error type scoring. By contrast, unidimensional models of the naming task predict that any differences between performance on different item sets should be meaningless noise after accounting for overall difficulty.

The goals of the present study were twofold: (a) We sought to test the validity of the MPT model's item difficulty estimates by discovering meaningful performance differences across selected item sets, and (b) we sought to characterize the psychometric properties of our measurement instrument for this novel construct, which we refer to as relative linguistic impairment (RLI). Our ultimate goal is to derive a clinically useful picture naming tool that is quick to administer, is easy to score (i.e., accuracy only), and yet will provide more detailed information than standard naming tasks regarding the source of the impairment in individual patients (as opposed to the severity of the impairment). Here, we present our first attempt at developing such a tool from archived data focusing on two subprocesses, lexical versus sublexical processing, and we believe the resulting product may serve as a useful starting point for investigating clinical applications as well as for future test development.

Participants and Data

We analyzed archived behavioral, lesion, and task-based functional neuroimaging data collected during a clinical trial investigating predictors of outcomes for aphasia rehabilitation (Kristinsson et al., 2023), although, in the current study, we were interested in the cross-sectional relationships among different measures of aphasia rather than the longitudinal effects of therapy. The clinical trial was approved by the institutional review board of the University of South Carolina (Pro00053559), and all participants provided informed consent before enrolling. A total of 127 participants with chronic left-hemisphere stroke consented to participate. Participants underwent comprehensive speech and language evaluations as well as structural and functional neuroimaging studies upon enrollment. After the initial assessments, participants received 3 weeks each of semantically oriented and phonologically oriented speech therapy in a counterbalanced order and then 1-month and 6-month follow-up assessments to evaluate maintenance of treatment gains. A primary outcome measure for the clinical trial was accuracy on the Philadelphia Naming Test (PNT; Roach et al., 1996), which includes 175 black-and-white line drawings of common objects. A

subset of the participants was included in the current study based on the completeness of the participants' naming data. Participants were administered the naming test twice before undergoing any therapy. Complete data were available for 74 first administrations and 17 second administrations of the PNT, yielding 91 participants who were included in this study. Complete naming data were a requirement to ensure that an RLI score could be calculated. Table 1 presents descriptive statistics for demographic and clinical variables of the included participants. Of note, 12 participants were above the Western Aphasia Battery (WAB) cutoff for aphasia (> 93.8 aphasia quotient); however, this cutoff has been acknowledged to be conservative (Kertesz, 2022), and linguistic impairments relative to healthy controls have been detected in this population (Fromm et al., 2017; Gordon, 2020; Richardson et al., 2018). Therefore, these participants were included to determine if they also exhibited RLI.

Constructing a Scale to Measure RLI

The Walker et al. (2018) MPT model can explain naming response type rates in terms of sequential combinations of probabilities for successful or unsuccessful

Table 1. Descriptive statistics for demographic and clinical variables of the included participants.

Variable	Frequency			
Participants	91			
Sex				
Female	40			
Male	51			
Race				
White	70			
African American	18			
Asian	1			
Aphasia type				
Anomia	25			
Broca's	37			
Conduction	12			
Wernicke's	4			
Transcortical motor	1			
None	12			
Apraxia of speech				
Present	46			
Absent	45			
	M	***SD***	**Min.**	**Max.**
Age	60.54	11.01	29	80
Education	15.55	2.31	12	20
Months after onset	55.93	54.33	10	241
WAB scores				
Information Content	7.17	2.60	0	10
Fluency	5.79	2.85	1	10
Yes/No Questions	55.58	5.48	36	60
Auditory Word Recognition	53.18	9.34	20	60
Sequential Commands	55.47	19.64	10	80
Object Naming	42.52	18.62	0	60
Word Fluency	7.68	6.02	0	20
Sentence Completion	7.47	2.78	0	10
Responsive Speech	7.04	3.58	0	10
Repetition	6.01	2.97	0.10	10
Word Finding	6.47	2.88	0.10	10
Aphasia Quotient	67.23	23.15	22.80	100

Note. Min. = minimum; Max. = maximum; WAB = Western Aphasia Battery (Kertesz, 2007).

latent cognitive operations during word production at the lexical (i.e., word selection) and sublexical (i.e., phonological sequencing and speech motor planning) stages of production. Although we believe that phonological sequencing and speech motor planning are separable cognitive processes within the sublexical domain, the scoring rubric for the naming test did not enable these distinctions to be made within the model. We did not apply the lenient scoring rubric that allows for single-phoneme errors in the context of a persistent speech motor impairment; any distortion that crossed a phoneme boundary was counted as an error.

Although the MPT model posits six separate abilities to avoid different types of errors, these abilities can be merged within a representational level to summarize the probability of avoiding any error type at that level. Specifically, we can calculate the probability of a person with average abilities avoiding a semantic neighbor, a formal neighbor, and a mixed neighbor during lexical selection for a given item and multiply these probabilities to obtain the probability of avoiding any lexical selection error (unrelated errors are excluded from the calculation because they likely arise from cognitive processing errors prior to lexical access, such as perseveration errors). Given the scoring rubric limitations, only a single ability-difficulty pairing governs the probability of a sublexical error. The MPT model entails that test items should exist that will selectively challenge the lexical or sublexical abilities for word production, because the required cognitive operations function at least partially independently. For example, for a person with average abilities, a lexically challenging item such as "van" would create a 49% chance of a lexical error and only a 9% chance of a sublexical error, while a sublexically challenging item such as "stethoscope" would create a 31% chance of a lexical error and a 72% chance of a sublexical error. Item difficulties and person abilities were reported by Walker et al. (2018); Supplemental Material S1 lists the probabilities of selection errors at the lexical and sublexical levels for each item.

The RLI assessment consists of two 20-item sets of pictured objects to be named that specifically challenge the lexical or sublexical production processes, respectively. The items with the largest differences between a probability of an error at the lexical or sublexical level for a participant with average abilities were selected (i.e., the top 20 items and bottom 20 items listed in Supplemental Material S1). We chose the assessment length based on the SCANT, which indicated that a 20-item set size optimizes predictive validity of naming accuracy scores (Walker, Fridriksson, et al., 2022). Although we did not explicitly control for overall difficulty of each set, we expected that the balanced manipulation of difficulty types between the sets and the general assumption of highly correlated accuracy scores between random item sets would result in item sets that were approximately equally difficult overall for the general population of people with aphasia. The RLI score is simply calculated as the difference in accuracy scores between the two item sets (RLI score = sublexical accuracy score − lexical accuracy score). The item "Eskimo" was excluded for cultural sensitivity reasons and because it is not a common noun. Importantly, data from the participants who were included in our validation experiments were not used to select the items for the RLI assessment. Although the full set of 175 items was presented during naming assessments, responses to the RLI assessment items were extracted and analyzed separately.

Does Cognitive Model-Based Item Selection Yield Appropriate Item Sets?

Before examining how people perform on the assessment, it is worth considering lexical properties of the items to ensure that the construction procedure yields item sets that are significantly different from each other in ways that are known to affect naming performance. We tested differences between the item sets in average log lexical frequency and average phonological length, which have different association strengths with different types of item difficulty (Walker et al., 2018). We also tested differences in average unidimensional difficulty estimates based on an IRT model fit to a large cohort of people with aphasia (Fergadiotis et al., 2015). We expected item sets to be significantly different with respect to these lexical properties that affect the rate of different naming errors, and we expected overall difficulty of the items estimated by the IRT model to be balanced. Average log lexical frequency was significantly higher, unpaired $t(38) = 2.74$, $p = 9.32 \times 10^{-3}$, for lexical items ($M = 1.29$, $SD = 0.52$) versus sublexical items ($M = 0.85$, $SD = 0.48$); average phonological length was significantly lower, unpaired $t(38) = 6.47$, $p = 1.28 \times 10^{-7}$, for lexical items ($M = 4.10$, $SD = 1.07$) versus sublexical items ($M = 7.1$, $SD = 1.77$), while average overall difficulty was not significantly different, unpaired $t(38) = 0.52$, $p = .60$, for lexical items ($M = 0.50$, $SD = 0.61$) versus sublexical items ($M = 0.61$, $SD = 0.78$). Thus, inferences about items based on difficulty estimates from a multidimensional MPT model (i.e., that these item sets should be significantly different regarding important lexical properties) were valid; meanwhile, inferences about items based on difficulty estimates from a unidimensional model would miss these important differences, because that model is only concerned with the overall difficulty rather than the source of the difficulty.

Does the RLI Assessment Yield Reliable Scores?

The RLI assessment must yield reliable scores to support the claim that the RLI construct is measurable. Assessments of the same individual by different examiners or by the same examiner at different times within a stable period should yield comparable results.

Internal Consistency

Across the 91 test administrations, Cronbach's α was .96 for the sublexical test and .94 for the lexical test; however, the correlation between the sublexical test score and the lexical test score was very high (r = .96). Applying the formula for the reliability of a difference score (Tisak & Smith, 1994), we obtain a reliability coefficient of −.066. Essentially, the shared variance between the two test scores implies that the RLI score will not be internally consistent for the average person with aphasia; if no true RLI exists, which may be the case for the population majority, the difference score will represent pure noise without any pattern, as intended. Nevertheless, if even a few individual participants exhibit reliable and meaningful RLI scores, this measure could still be clinically relevant (Rogosa et al., 1982).

Interrater and Intrarater Agreement

Five speech-language pathology master's students scored and rescored nine different PNTs that were selected to represent the full range of naming ability for interrater and intrarater agreement (Walker, Basilakos, et al., 2022). The items for the lexical and sublexical tests were extracted from the full PNT, and the scores from different raters (or the same rater at different times) were compared. We report the degree of absolute agreement for the intraclass correlation, also known as criterion-referenced reliability.

On the lexical test, the intraclass correlation coefficient between scores from different raters (interrater reliability) was 9.95×10^{-1}; six pairs of raters produced the same score, two pairs of raters disagreed by 1 point, and one pair of raters disagreed by 2 points. On the sublexical test, the intraclass correlation coefficient was 9.85×10^{-1}; four pairs of raters produced the same score, three pairs of raters disagreed by 1 point, one pair of raters disagreed by 2 points, and one pair of raters disagreed by 3 points. Notably, interrater agreement was slightly lower for the sublexical test compared to the lexical test, although the average decrease in agreement was not statistically significant, paired-sample $t(8) = -1.08$, p = .31. For the RLI score, the intraclass correlation coefficient between different raters was .83, with three pairs of raters producing the same score, five pairs of raters disagreeing by 1 point, and one pair of raters disagreeing by 3 points. Cicchetti (1994) characterizes reliability coefficients greater than .75 as an excellent level of clinical significance. Notably, the most extreme RLI score in this group (−4, sublexical RLI) was rated with perfect agreement by different raters. Intraclass correlation coefficients for intrarater reliability were near perfect for the lexical test score (9.98×10^{-1}), the sublexical test score (9.94×10^{-1}), and the RLI score (9.85×10^{-1}). On the lexical test, a single item for a single participant was scored differently; on the sublexical test, a single item was scored differently for two participants.

Test–Retest Reliability

There were 65 participants with complete data available for both naming assessments prior to treatment. Assessment dates were available for 60 participants: 39 participants (65%) had one intervening day between test and retest, 16 participants (27%) had two to seven intervening days, and five participants (8%) had seven to 21 intervening days. No change in naming abilities was expected during these intervals. We report descriptive statistics for test–retest differences in the RLI score. We also tested the significance of the intraclass correlation between the first and second RLI scores. It is worth noting that if most participants truly had no RLI, then a strong correlation between the first and second scores is not expected; the reliability contributed by the minority of participants with consistent RLI over time can be dwarfed by the measurement noise of those without a true RLI. It is also worth noting that participants received feedback about the correct response after each trial to prevent perseveration, which may have impacted RLI measurements on subsequent testing. Finally, it is worth noting that several different raters scored the tests, meaning that test–retest variability was confounded with interrater and intrarater variability in our experimental design, likely inflating our estimates of the true test–retest variability.

The average test–retest difference for the RLI score was not significantly different from zero ($\mu = -0.28$ points, SD = 2.60 points), $t(64) = 0.86$, p = .39, while the average absolute test–retest difference was 1.94 (SD = 1.74) points. The maximum absolute test–retest difference was 7 points; 80% of absolute test–retest differences were 3 points or less. The intraclass correlation coefficient between the first and second RLI scores was significant (r_I = .43 [.20, .61], F = 2.48, $p = 1.83 \times 10^{-4}$). Cicchetti (1994) characterizes reliability coefficients between .40 and .59 as having a fair level of clinical significance, although others in the field might reasonably consider this to be a poor level of clinical significance. We do not wish to understate this point. In the assessment's current form, a

single RLI score from a randomly selected person with aphasia cannot be considered a reliable indicator of future performance. There is a considerable noise component to each RLI measurement; however, when a set of RLI scores in a cross-sectional sample of people with aphasia or in a longitudinal sample from an individual person with aphasia is examined, a meaningful signal emerges. Furthermore, convergent evidence can be used to contextualize an individual RLI score to aid in its interpretation. We provide further examples to illustrate these points in the remaining sections.

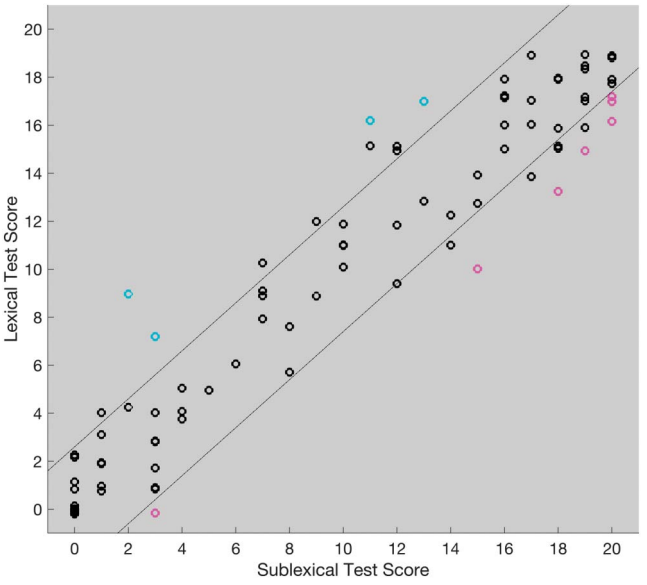
Do Individual Participants Exhibit Significant RLI?

We examined the difference in the proportion of accurate responses between the lexical and sublexical tests using a Barnard exact test for each participant. RLI scores for all participants and all administrations are provided in Supplemental Material S2, and corresponding p values from a Barnard's test of each RLI score are provided in Supplemental Material S3. For reference, Barnard's test p values are provided for any pair of lexical and sublexical test scores in Supplemental Material S4. We examined results in our cross-sectional sample of initial test administrations at two levels of significance ($p < .05$ and $p < .1$), without correcting for multiple comparisons.

At the $p < .05$ level, there was one participant with a significant sublexical RLI and five participants with a significant lexical RLI. At the $p < .1$ level, there were four participants with a significant sublexical RLI and seven participants with a significant lexical RLI. Figure 1 shows the scatter plot of scores on the lexical and sublexical tests, with significant lexical RLI highlighted in magenta and significant sublexical RLI highlighted in cyan ($p < .1$). Dotted lines indicate 1 *SD* of test–retest score differences (2.60 points).

Given the modest internal consistency and test–retest reliability of the RLI score, there is a legitimate concern whether the classifications of significant RLI scores can be reliable over time within an individual participant. If the item sets were truly equally difficult for a participant, significantly discrepant scores between the item sets could still emerge by chance; however, over multiple administrations, significant discrepancies in both directions would be expected. We examined repeated administrations of the RLI assessment to determine if any patients exhibited significant RLI in one direction too frequently to conclude that a single underlying rate could explain performance on both item sets. There were 80 patients with more than two RLI administrations in our data set, ranging from four to 12 administrations per person. We

Figure 1. Lexical and sublexical test scores for 91 participants with left-hemisphere stroke. The left panel shows the scatter plot illustrating the relationship between lexical test scores and sublexical test scores. A random Gaussian jitter (*SD* = 0.2) has been applied to the lexical test score (*y*-axis) for display purposes. Dashed lines represent one 1 *SD* of test–retest differences (2.60 points). Black circles indicate participants with no relative linguistic impairment (RLI), magenta circles indicate participants with significant lexical RLI, and cyan circles indicate participants with significant sublexical RLI. Significance was determined with Barnard's exact test ($p < .1$).



simulated performances on the RLI assessment for 1,000 instances of a simulated participant, each performing the assessment from four to 12 times, with a latent accuracy rate of 50% for all items, thereby maximizing the variance and potential to observe randomly discrepant scores between item sets. We then calculated the absolute value of the difference in the number of significant positive and significant negative RLI scores according to a Barnard's test ($p < .1$), obtaining the count of significant RLI scores in one direction versus the other. Finally, we obtained the 99th percentile of this absolute difference statistic over the 1,000 simulated instances for each number of test administrations to identify a threshold for the expected number of significant RLI scores in one direction when a single rate governs performance on both item sets. These thresholds are listed in Table 2.

There were two participants (Participants 49 and 113) with repeatedly significant lexical RLI scores exceeding the identified thresholds. Participant 49 was not included in our cross-sectional sample of participants with significant lexical RLI, because the RLI score only approached significance in the initial test administrations, becoming reliably significant after treatment later in the

Table 2. For each number of test administrations, the expected number of significant relative linguistic impairment (RLI) scores in one direction (i.e., subtracting the number of significant scores in the other direction) when a single underlying rate governs accuracy on the lexical and sublexical tests.

Number of test administrations	Threshold for unidirectional sig. RLI
4	2
5	2
6	3
7	3
8	3
9	3
10	3
11	4
12	4

Note. sig. = significant.

study. Two participants in our cross-sectional sample of participants with significant lexical RLI were not reliably significant over time (Participants 9 and 91), and four participants did not have repeated testing to determine reliability (Participants 76, 78, 88, and 119). There were six participants (12, 15, 29, 45, 82, 100) with repeatedly significant sublexical RLI scores exceeding the identified thresholds. One participant (104) in our cross-sectional sample of participants with significant sublexical RLI was not reliably significant over time, although the reliability was equal to (but not greater than) the threshold (i.e., three significant scores out of six administrations). Three participants (12, 29, 45) were not included in our cross-sectional sample of participants with significant sublexical RLI, because their initial RLI scores were not significant, only becoming significant in later administrations. In total, we identified eight out of 80 participants (10%) who exhibited significant RLI scores in one direction too reliably to be explained by a single underlying accuracy rate; however, none of these participants exhibited a significant RLI score on all test administrations, reinforcing the point that there is a substantial noise component to the RLI score. Nevertheless, it is clearly possible to observe enough significant RLI scores within an individual participant over time to reject the hypothesis that these item sets are equally challenging for that individual.

Are RLI Scores Meaningful?

Ideally, we would like to know if RLI scores can inform clinical decisions to improve patient outcomes; however, to investigate this would require a large, prospective study comparing clinical decisions made with and without knowledge of RLI scores. In lieu of these data, we retrospectively examined archived data for evidence that inferences based on the intended meaning of RLI scores about people with aphasia lead to valid conclusions. Essentially, we investigated whether RLI scores provide useful information about neurologically acquired language impairments, beyond overall naming accuracy scores. The following experiments were all intended to provide evidence of construct validity.

Process Dissociation

Jacoby (1991) introduced a methodological framework called process dissociation to investigate psychologically complex tasks. Rather than associating an individual mental process with an individual task that is supposedly process pure, Jacoby assumed that experimental manipulations could specifically modulate component processes of a complex task, thereby revealing multiple factors that contribute to behavioral performance. Crucially, different tasks are assumed to rely on some shared mental processes, and experimental conditions can be formed based on the presumed cognitive components underpinning the task of interest. If the multicomponent model holds, then it should be possible to find people who are relatively impaired on one component of a task (and thus impaired on other tasks that rely on that component) while being relatively unimpaired on the other component (and thus unimpaired on other tasks that rely on that component). It should also be possible to find people with the reverse pattern of task performances. Because tasks are assumed not to be process pure, typically only relative dissociations in performances are expected to result from experimental manipulations.

Method

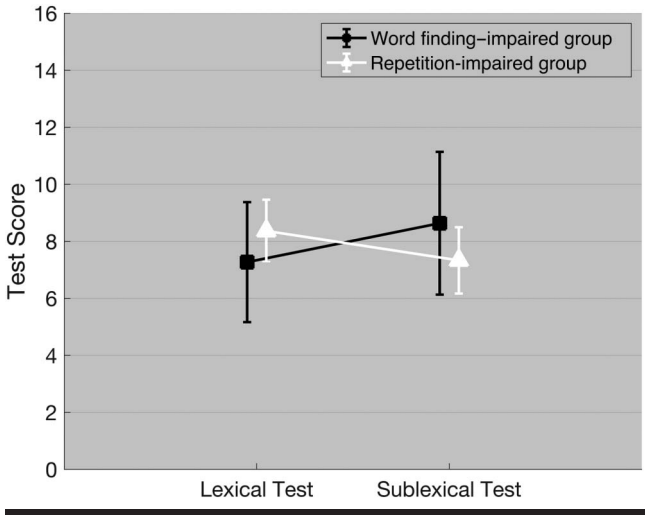
Our central claim is that behavioral performance on a picture naming task is not monolithic but instead relies on multiple cognitive abilities, with a primary distinction between lexical and sublexical processing. Following the logic of Jacoby (1991), we assumed there should be individuals who are more impaired in lexical processing than sublexical processing (and vice versa), and this relative impairment should be reflected in different tasks (creating theoretically relevant groups) as well as in modulations of naming accuracy on specific item sets (creating experimental conditions). That is, a dissociation in the performance on the lexical and sublexical naming tests was predicted based on a dissociation in the performance on WAB subtests. In particular, the Word Finding subscore—a composite of Object Naming, Sentence Completion, and Responsive Speech task scores—served as a proxy measure for lexical processing ability, and the Repetition subscore served as a proxy measure for sublexical processing ability. The repetition task has previously been used as a proxy measure for the sublexical stages of naming,

because it requires articulation but does not require access to the meanings of words (Dell et al., 2007). The Word Finding subscore was selected as a proxy for lexical processing based on face validity (Kertesz & Poole, 1974). Using a 10% or greater difference between the selected WAB subscores as a criterion, participants were classified as *repetition impaired* or *word finding impaired*. We used a repeated-measures analysis of variance model with naming test type (lexical/sublexical) as a within-subjects factor and WAB-based subgroup as a between-subjects factor to test the interaction effect on naming accuracy. We used paired *t* tests to examine if word finding–impaired participants performed significantly worse on average on the lexical naming test compared to the sublexical naming test and if repetition-impaired participants performed significantly worse on average on the sublexical naming test compared to the lexical naming test.

Results

Thirty participants were classified as repetition impaired, and 11 participants were classified as word finding impaired. In terms of group average performance, a significant crossover interaction was revealed ($F = 10.64$, $p = 2.3 \times 10^{-3}$), shown in Figure 2. The word finding–impaired participants performed significantly worse on average on the lexical naming test compared to the sublexical naming test ($\mu_{lexical} = 7.33$, $\mu_{sublexical} = 8.37$, $SD_{pooled} = 2.17$), $t(29) = 2.60$, $p = .014$, while the repetition-impaired participants performed significantly worse on average on the sublexical naming test compared to the lexical naming test ($\mu_{lexical} = 8.63$, $\mu_{sublexical} = 7.27$, $SD_{pooled} = 1.80$), $t(10) = 2.51$, $p = .031$. These findings would be impossible to predict with a unidimensional item response model, which asserts that between-persons performance differences can be explained by a single variable (i.e., naming ability). The observed interaction demonstrates that accuracy depends on at least two variables: lexical and sublexical abilities. While the unidimensional naming model has accumulated substantial evidence supporting the validity of inferences in most cases, particularly when lexical and sublexical impairments are comparable, the RLI assessment can predict when the unidimensional model will fail, illuminating its blind spots.

Figure 2. A significant crossover interaction emerged between average performance on the lexical test versus the sublexical test of the relative linguistic impairment assessment for the 11 word finding -impaired participants and the 30 repetition-impaired participants who were classified based on Western Aphasia Battery (Kertesz, 2007) subtest scores. Whiskers represent 1 standard error of the mean.



Speech Fluency

Historically, speech fluency has been among the most important factors used by clinicians to distinguish between aphasia types (Kertesz & Poole, 1974). A prominent criticism of using naming accuracy as a proxy for aphasia severity has been that it can sometimes diverge from assessments of connected speech in a naturalistic context (Conroy et al., 2009; Fergadiotis & Wright, 2016; Mason & Nickels, 2022; Mayer & Murray, 2003). Given that speech fluency is strongly related to the characterization of aphasia and can vary at least partially independently of severity (e.g., patients with moderate or severe aphasia may still be highly fluent), we expected that the RLI score would contain predictive information about speech fluency scores, beyond overall naming accuracy scores.

The modern perspective of speech fluency views it, like naming, as a multicomponent construct (Casilio et al., 2019; Gordon, 2020). Although the WAB Fluency scale incorporates grammatical, motor, and paraphasic aspects of speech into a single rating, the holistic qualitative judgment of speech fluency continues to be a prevalent clinical indicator that guides assessment and treatment (Casilio et al., 2019). Importantly, these linguistic processes are closely aligned with the MPT model's sublexical processes (i.e., paraphasia and motor speech, while grammatical impairments may or may not be related to lexical concept retrieval). We therefore investigated whether the RLI scores contained predictive information about WAB Fluency scores, beyond naming accuracy scores.

Method

Scores for each subtest of the WAB are included in Supplemental Material S5. We used leave-one-out cross-validation to test how well the following models could predict WAB Fluency scores (range: 1–10): (a) the average WAB Fluency score (a null model), (b) a linear model with the accuracy score from the SCANT as the independent variable, and (c) a multiple linear model with the SCANT accuracy score and the RLI score as independent

variables. Paired t tests were used to compare the mean absolute prediction errors (MAPEs) between the models. Notably, the cross-validation procedure is not biased by the different number of predictors in each model and can provide evidence of predictive validity (Yarkoni & Westfall, 2017). We examined the regression coefficient β for the RLI score in the multiple linear model fit to the full participant sample to interpret the direction of the effect.

Results

The model that included both the SCANT and RLI scores resulted in the best predictions for WAB Fluency (MAPE = 1.50, SD = 1.08), and the partial correlation between the RLI score and the WAB Fluency score, after accounting for the SCANT accuracy score, was significant (β = 0.35), $t(88) = 3.83$, $p = 9.10 \times 10^{-3}$. While prediction error for WAB Fluency score was significantly lower, $t(90) = 4.65$, $p = 1.14 \times 10^{-5}$, for the linear model that included the SCANT accuracy score (MAPE = 1.68, SD = 1.04) than for the null model that included only the average WAB Fluency score (MAPE = 2.53, SD = 1.34), prediction error for WAB Fluency score was significantly lower still for the multiple linear model that included the SCANT accuracy score and the RLI score compared to the linear model that included only the SCANT accuracy score, $t(90) = 2.67$, $p = 9.10 \times 10^{-3}$. According to the multiple linear model fit to the full participant sample, after accounting for SCANT accuracy, the coefficient β for the RLI score was positive, indicating that people with positive (lexical) RLI scores tended to have higher WAB Fluency scores than expected while people with negative (sublexical) RLI scores tended to have lower WAB Fluency scores than expected. The WAB Fluency scores for the four participants with significant sublexical RLI were (in participant order) [2, 4, 2, 6], while the scores for the seven participants with significant lexical RLI were (in participant order) [4, 10, 10, 10, 9, 9, 10]. These findings confirm that RLI scores contain predictive information about speech fluency beyond unidimensional naming accuracy scores, and they accord with the predicted direction of the effect based on the intended meaning of the RLI scores.

Apraxia of Speech

As indicated previously, speech fluency and motor speech deficits (i.e., apraxia of speech) are separate but related constructs (Strand et al., 2014). The contemporary view of apraxia of speech considers it to be a continuous, multidimensional syndrome (Haley & Jacks, 2023), but the important point for our purpose is that the disorder is ontologically unrelated to lexical retrieval deficits, falling squarely in the domain of motor planning (although these distinct impairments can and do co-occur). This means that, like speech fluency, we expected motor speech deficits to be specifically related to sublexical RLI.

Method

Presence or absence of apraxia of speech was determined for each participant based on the Apraxia of Speech Rating Scale (Strand et al., 2014); these data can be found in Supplemental Material S5. We compared the incidence rates of apraxia of speech among the four participants with significant sublexical RLI and among the seven participants with significant lexical RLI using a Barnard exact test. For comparison, we also examined the incidence rates of apraxia of speech among the four participants with the lowest SCANT accuracy scores and among the seven participants with the highest SCANT accuracy scores.

Results

All four participants with significant sublexical RLI (100%) presented with apraxia of speech, while only one participant with significant lexical RLI (14%) presented with apraxia of speech. The difference in the incidence rates between the two groups was significant (Wald = 2.75, $p = 4.21 \times 10^{-3}$). Two participants with the lowest SCANT accuracy scores (50%) presented with apraxia of speech, and two participants with the highest SCANT accuracy scores (28%) presented with apraxia of speech. The difference in incidence rates was not significant (Wald = 0.71, $p = .36$). These results again demonstrate that the RLI score contains information about a specific source of impairment, here apraxia of speech, beyond unidimensional naming accuracy scores.

Lesion Location

Characteristics of the lesion, especially the parts of the brain that have been affected by a stroke, have been another historically important consideration for distinguishing among types of aphasias (Geschwind, 1965; Lichtheim, 1885). Although the modern perspective on the neurobiology of language has moved beyond the simplicity of the classical model (Hickok, 2022; Hickok & Poeppel, 2004; Tremblay & Dick, 2016), it is still generally assumed that different symptom patterns result from damage to specific functional brain networks. We therefore posited that individuals with extreme RLI scores may represent syndromes that are associated with specific patterns of lesion damage.

Method

Structural magnetic resonance imaging data acquisition and preprocessing. Participants underwent neuromaging at two time points: prior to and immediately after treatment. In this study, we examined pretreatment

neuroimaging, which, for most participants, was collected on the same day as behavioral assessments. Magnetic resonance imaging (MRI) data were collected on a Siemens Prisma 3T scanner (Siemens Medical Systems) with a 20-channel head coil. For each participant, we acquired a T1-weighted image (MP-RAGE: 1-mm isotropic voxels, matrix = 256 × 256, 9° flip angle) with the following sequence parameters: 192 slices, TR = 2,250 ms, TI = 925 ms, and TE = 4.15 ms. We also obtained a T2-weighted structural image using a three-dimensional turbo spin echo scan with the following sequence parameters: 192 slices, TR = 2,800 ms, and TE = 403 ms.

Structural scans were preprocessed using MATLAB (R2017b, The MathWorks) with a publicly available custom image-processing pipeline specifically designed to work with stroke populations (Rorden et al., 2012; https://github.com/neurolabusc/nii_preprocess). The pipeline utilizes the following software to process neuroimaging data: SPM8 (Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, https://www.fil.ion.ucl.ac.uk/spm), FSL (Version 6.0.3; Jenkinson et al., 2012), ASLtbx (https://www.cnf.upenn.edu/~zewang/ASLtbx.php), and MRtrix (2012). The output is generated in standard space and can be further queried at the voxel or region-of-interest level. The output files were visually inspected to ensure quality of the data preprocessing.

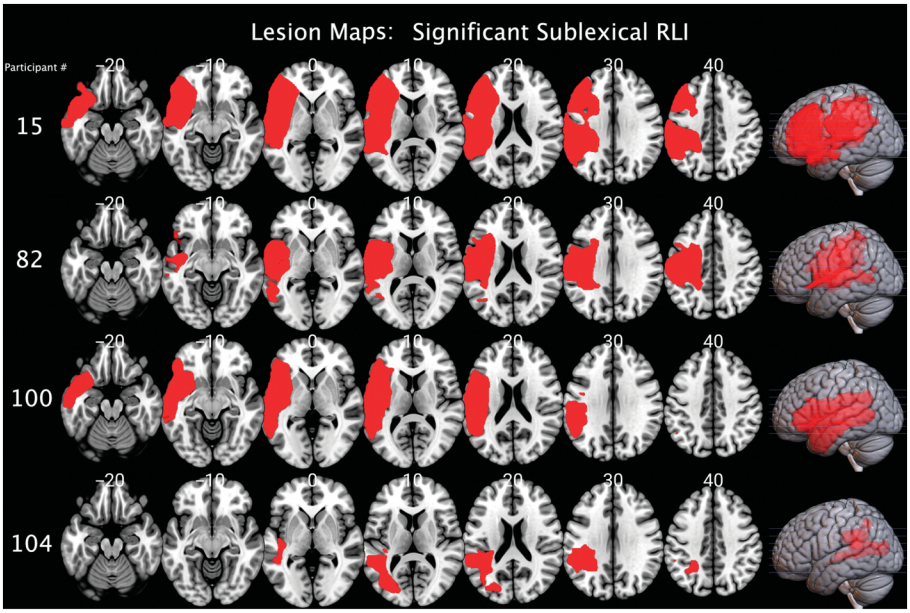
Generation of lesion maps. Lesions were manually demarcated on individual T2-weighted images by a licensed neurologist using the MRIcron software (Rorden et al., 2012). For each participant, the lesion was drawn on the T2-weighted image, which was then coregistered to the T1-weighted image. The resulting transformation matrix was leveraged to reslice the lesion to native T1 space. A 3-mm full-width half-maximum Gaussian kernel was used to smooth the lesion maps and remove jagged edges due to manual demarcation. We then performed enantiomorphic (Nachev et al., 2008) segmentation–normalization (Ashburner & Friston, 2005) to convert the images to Montréal Neurological Institute (MNI) standard space (https://github.com/rordenlab/spmScripts/blob/master/nii_enat_norm.m). This process relies on the fact that the brain is left–right symmetrical; healthy tissue from the right hemisphere is used to replace the lesioned voxels of the T1 image prior to normalization, which may not perform well in damaged brains. Next, the lesion image was resliced into standard MNI space (1 × 1 × 1 mm isotropic voxels) by way of linear interpolation. The tissue segmentation maps generated by the enantiomorphic normalization–segmentation routine were used to create brain-extracted examples of each individual's T1 and T2 scans, which were then used to normalize the functional MRI (fMRI) scans as discussed below.

Data analysis. Lesion maps for the four participants with significant sublexical RLI and the seven participants with significant lexical RLI were grouped and informally examined for similarities and differences. We were particularly interested in damage to areas that are known to be important for speech and language, particularly perisylvian regions including the superior temporal gyrus, the inferior frontal gyrus, and the insula.

Results

Figure 3 shows the lesion maps for the four participants with significant sublexical RLI, and Figure 4 shows the lesion maps for the seven participants with significant lexical RLI. The lesions associated with sublexical RLI tend to encompass the lateral portions of the perisylvian region, whereas the lesions associated with lexical RLI tend to spare the temporal and frontal operculum as well as the insula, instead mainly impacting the medial white matter. There are exceptions to both patterns: Participant 104 with a sublexical RLI had damage confined to the most posterior aspects of the perisylvian regions, while Participant 9 with a lexical RLI had extensive damage to perisylvian regions (although the inferior frontal gyrus was spared). These exceptional cases did not exhibit reliably significant RLI scores over time. The number of voxels damaged in each region of a combined gray-and-white matter atlas (Catani & Thiebaut de Schotten, 2008; Tzourio-Mazoyer et al., 2002) is presented in Supplemental Material S6. All four participants with significant sublexical RLI scores had damage to the Rolandic operculum; insula; middle occipital gyrus; postcentral gyrus; inferior parietal lobe; supramarginal gyrus; angular gyrus; Heschl's gyrus; superior and middle temporal gyri; arcuate fasciculus including the anterior, posterior, and long segments; cortico-ponto-cerebellar tract; corticospinal tract; inferior longitudinal fasciculus; inferior fronto-occipital fasciculus; internal capsule; and optic radiations. All seven participants with significant lexical RLI scores had damage to the arcuate fasciculus, corpus callosum, corticospinal tract, and internal capsule. Damage to the corpus callosum is notable, particularly as it distinguishes the lexical RLI group from the sublexical RLI group. This white matter bridge between the cerebral hemispheres is not typically associated with aphasia, although it has been associated with unilateral visual anomia (Lausberg et al., 1999). Interpretation of these findings relative to the extant literature is challenging. The lesion characteristics alone are not sufficient to predict the presence of an RLI, because most patients in our sample with similar lesion characteristics to either group's prototypical lesions do not exhibit RLI; however, when an RLI is present, the type of RLI implicates different lesion patterns. The different patterns of lesions associated with different types of RLI lend support to the biological meaningfulness of the RLI construct.

Figure 3. Lesion maps for the four participants with significant sublexical relative linguistic impairment (RLI). Each row contains axial slices from a single participant and a final image showing a sagittal view of a three-dimensional brain rendering and horizontal lines corresponding to the slices. Numbers on slices indicate the axial coordinate in Montréal Neurological Institute space.



Functional Activation During Naming

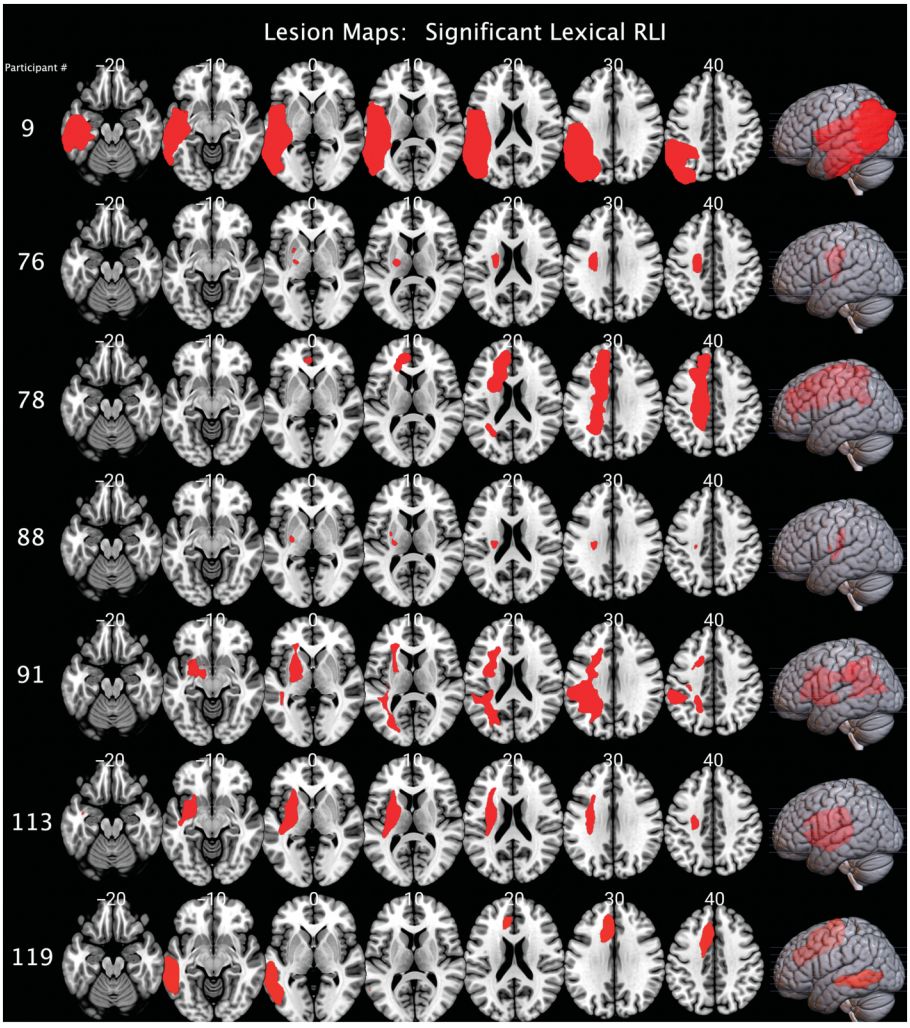
The picture naming task in the fMRI scanner is known to reliably activate a set of bilateral perisylvian regions including superior temporal gyrus, inferior frontal gyrus, Rolandic operculum, and insula, as well as nonperisylvian regions including occipital lobe, fusiform gyrus, middle temporal gyrus, temporal pole, angular gyrus, and anterior cingulate gyrus in both healthy controls and people with aphasia (Abel et al., 2015; Fridriksson et al., 2009; Sebastian & Kiran, 2011; Skipper-Kallal et al., 2017; Stefaniak et al., 2021). Furthermore, increased task difficulty (and increased error rate) is associated with increased activation in these same regions, presumably reflecting increased effort or inefficient processing (Fridriksson et al., 2010; Postman-Caucheteux et al., 2010; Stefaniak et al., 2021). The dual-stream model of language representation in the brain (Hickok & Poeppel, 2007) posits that lexical processing during naming occurs in a ventral stream, primarily in the middle and anterior temporal lobe, while sublexical processing occurs in a dorsal stream, primarily in a superior temporal, temporoparietal, and inferior frontal circuit underlaid by the arcuate fasciculus. It stands to reason that if the lexical and sublexical naming tests are appropriately modulating the difficulty of specific processes during naming, this should be reflected in the activity within the relevant functional networks. Likewise, if the RLI score is appropriately reflecting each participant's relative challenges during naming, and each participant exerts different relative levels of effort in processing to obtain an observed accuracy score, the modulations in brain activity should be detectable when naming any set of items. We did not collect neuroimaging of participants naming the RLI items; however, the intended meaning of the RLI score was expected to generalize to any set of named items.

Method

Functional MRI data acquisition. Participants underwent task-based fMRI sessions before and after treatment. Task-based fMRI data were acquired using T2* MRI echo-planar imaging with the following sequence parameters: 60 full brain volumes, 90° flip angle, TR = 10 seconds, TA = 2 s, TE = 30 ms, matrix = 64 × 64, in-plane resolution = 3.25 × 3.25 mm, slice thickness = 3.2 mm (no gap), and 33 axial slices collected in planes aligned parallel to the anterior commissure–posterior commissure line.

The fMRI task utilized a simple picture-naming paradigm where participants were asked to attempt to name 40 colored high-frequency noun pictures (see Supplemental Material S7) and to stay silent during the presentation of 20 colored abstract images. The pictured objects named in the scanner were not the same items that were used to evaluate naming accuracy or RLI outside the scanner (except one item: "piano"). The fMRI scanning session lasted 10 min, and abstract images were presented at random among the real picture presentations. Pictures were back-projected on an MRI-compatible screen, and

Figure 4. Lesion maps for the seven participants with significant lexical relative linguistic impairment (RLI). Each row contains axial slices from a single participant and a final image showing a sagittal view of a three-dimensional brain rendering and horizontal lines corresponding to the slices. Numbers on slices indicate the axial coordinate in Montréal Neurological Institute space.



participants observed the pictures via a mirror mounted on the scanner head coil. Each picture was presented for 2 s. Naming attempts were recorded through a nonferrous microphone and were subsequently scored offline. The fMRI task was designed to allow us to isolate activation associated with naming. The same task has been used effectively in our prior research (Fridriksson et al., 2010, 2012; Kristinsson et al., 2021).

We used a sparse imaging sequence where a single full brain volume was collected every 10 s to improve clarity of the audio recordings and to minimize speech-related head movements. Acquisition of each volume lasted 2 s, which allowed for 8 s of scanner silence until the next volume was acquired. This 8-s window was utilized for presentation of stimulus pictures (2 s) and a naming attempt. The interval between picture presentation was jittered (i.e., sampling at different time points following each picture presentation) to better model the hemodynamic response in the fMRI data analysis. The interval between picture presentations varied between 6 and 8 s. In order to minimize the chance that participants would name pictures during acquisition of fMRI data, pictures appeared at least 3 s prior to the acquisition of the subsequent scan.

Preprocessing of fMRI data. fMRI data were corrected for motion using SPM12's default *realign and unwarp* procedure, and the output images were spatially realigned with the brain-extracted T2-weighted image due to the similar contrast across the two T2-weighted scans. Stimulus onsets were convolved with the canonical hemodynamic response function and its temporal derivative following slice time correction. A mean fMRI image was derived by averaging all 60 volumes acquired during the

fMRI session for each participant, regardless of accuracy. Each participant's mean image was scalp stripped using FSL's Brain Extraction Tool normalized to the scalp-stripped T2 scan (note that the T2-weighted image had previously been mended using enantiomorphic unified segmentation–normalization and was therefore in standard space as described above). Of note, the T2* fMRI image and the high-resolution, low-distortion T2-weighted image have very similar image contrast, including at the location of the lesion. The resulting normalization deformation was applied to the original (i.e., not scalp-stripped) fMRI series. Nonbrain tissue was ignored for the final normalization. All fMRI data were then smoothed with a Gaussian kernel with full-width half-maximum of 6 mm. Voxel-wise data were detrended using mean signal from the white matter, and subject and independent component analysis was used to automatically identify and remove lesion-driven artifacts in the data (Yourganov et al., 2017). Finally, we estimated the main effects of the two task conditions of interest (overt naming of high-frequency nouns regardless of accuracy and silent viewing of abstract images) using SPM12's general linear model to generate naming-related activation maps in standard space. These difference maps represented areas of greater signal during picture naming than viewing of abstract pictures (Ashburner et al., 2012). Subsequently, for each participant, the average activation in each region of the AICHA atlas was calculated (Joliot et al., 2015).

Data analysis. There were missing data from some of the observed fMRI variables. Three participants were missing fMRI data entirely. Furthermore, signal dropout from region to region caused different group sizes, ranging from 54 to 88 participants. However, there were at least 24 participants with fMRI data and with less than 5% of the volume damaged in each region. We used an arbitrary 5% volume damage criterion for identifying potentially disrupted regions to account for individual variability introduced by warping lesion maps to a standard template. This coverage enabled reliable estimation of the linear effects on fMRI activation within each region, comorbid with damage elsewhere in the hemisphere.

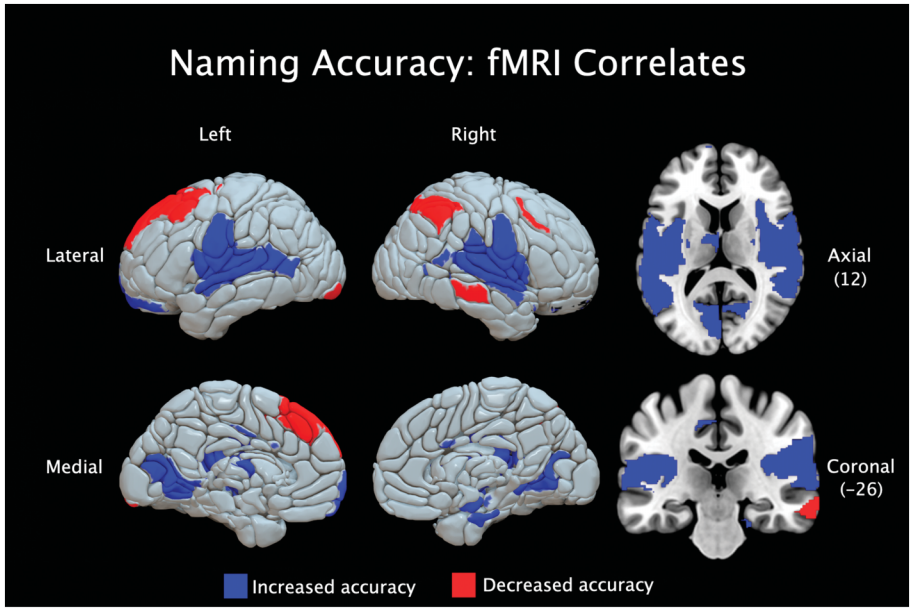
For each brain region, we fit a linear least-squares regression model predicting activation from (a) the PNT accuracy rate (see Supplemental Material S5), (b) the RLI score, and, if two or more participants had more than 5% damage to the region, (c) a dummy variable indicating which participants had lesions in that region. If the dummy variable were included in models for regions that are not damaged in any participants, it would simply act as a constant and have no impact on the estimation of the other coefficients. For reference, the lesion distribution for all 91 participants is shown in Supplemental Material S8, covering nearly the entirety of the left cerebral hemisphere with highest frequencies near the insula and in dorsal white matter. We used permutation tests to examine the significance of the regression coefficients for accuracy rates and RLI scores. Holding the other independent variables fixed, the scores from the variable of interest (i.e., the accuracy rate or the RLI score) were randomly assigned to different participants without replacement, and then the regression model was fit to these permuted data. This process was repeated 2,000 times to generate a random distribution of coefficients to compare against the coefficient that was obtained from the observed data. The permutation p value was calculated as the proportion of random coefficients that were greater than the observed coefficient (or less than the observed coefficient if it was negative); this was a one-tailed test. Permutation p values less than .05 were interpreted as significant. For reference, parametric two-tailed p values based on partial correlations of each independent variable were also reported.

Results

Contrast values for each participant in each region of the AICHA atlas are presented in Supplemental Material S9. Beta coefficients, partial correlations, parametric p values (two-tailed), and permutation p values (one-tailed) for the independent variables (i.e., accuracy scores and RLI scores) predicting activation in each region of the AICHA atlas are provided in Supplemental Material S10. Figure 5 shows brain regions where activity during naming was significantly modulated in association with overall accuracy scores (obtained outside the scanner). The regions with positive associations, where increased accuracy was associated with increased activation, are shown in blue; many of these regions are among those that are the most activated during naming or speech production in healthy individuals. These regions included bilateral midline cingulate cortex, putamen, occipital and occipitoparietal cortex, Rolandic sulcus and operculum, anterior and posterior insula, and superior temporal gyrus and sulcus; left postcentral gyrus, frontal orbitalis, middle temporal gyrus, and thalamus; and right inferior temporal gyrus and parahippocampal gyrus. Regions with a negative association, where increased accuracy was associated with reduced activation, are shown in red. These regions included bilateral middle frontal gyrus; left superior frontal gyrus and sulcus, superior precentral sulcus, occipital pole, and supplementary motor area; and right supramarginal gyrus, angular gyrus, inferior parietal and intraparietal sulcus, and middle temporal gyrus.

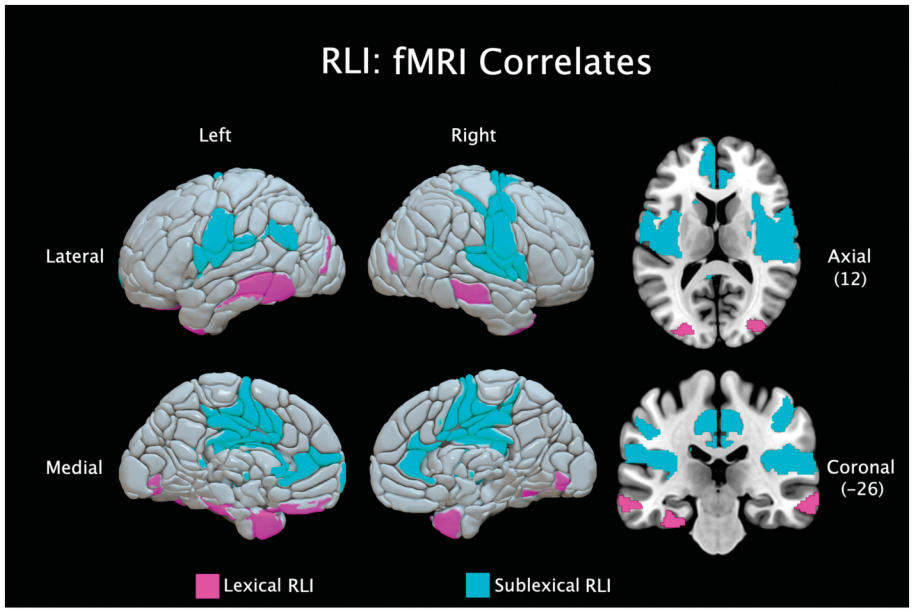
Figure 6 shows brain regions where activity during naming was significantly modulated in association with the RLI score, independent of the accuracy score. Participants with lexical RLI (i.e., positive RLI scores) tended to have greater activity than expected and participants with sublexical RLI (i.e., negative RLI scores) tended to have

Figure 5. Regions highlighted in blue or red have activation that is positively or negatively associated, respectively, with individual naming accuracy outside of the scanner, after controlling for relative linguistic impairment and presence or absence of lesion (greater than 5% of the region's volume). Permutation, $p < .05$. The axial slice MNI coordinate is 12, while the coronal slice MNI coordinate is –26. fMRI = functional magnetic resonance imaging; MNI = Montréal Neurological Institute.



less activity than expected in ventral regions highlighted in magenta. These regions included bilateral occipital cortex, middle temporal gyrus, middle temporal pole, lingual gyrus, and fusiform gyrus as well as left olfactory, inferior temporal gyrus, and superior temporal pole. Participants with sublexical RLI (i.e., negative RLI scores) tended to have greater activity than expected and participants with lexical RLI (i.e., positive RLI scores) tended to have less

Figure 6. Regions highlighted in magenta or cyan have increased activation that is associated with increased lexical RLI or sublexical RLI, respectively, after controlling for naming accuracy outside the scanner and presence or absence of lesion (greater than 5% of the region's volume). Permutation, $p < .05$. The axial slice MNI coordinate is 12, while the coronal slice MNI coordinate is –26. fMRI = functional magnetic resonance imaging; MNI = Montréal Neurological Institute; RLI = relative linguistic impairment.



activity than expected in dorsal regions highlighted in cyan. These regions included bilateral precentral and postcentral gyrus (right dominant), Rolandic sulcus and operculum, supramarginal gyrus (left dominant), anterior and posterior insula, midline cingulate cortex, and caudate (left dominant); left superior temporal sulcus; and right superior temporal gyrus.

To summarize, activation associated with overall accuracy was concentrated in bilateral perisylvian areas, consistent with previous investigations of activation related to speech production. Simultaneously, an independent effect of RLI was found in dorsal and ventral language networks. Increased activation in these networks was hypothesized to reflect increased effort or inefficient retrieval processes. Given that the participants were naming different items in the scanner than were named for the RLI assessment, these findings provide strong evidence for the validity and generalizability of the RLI score. The RLI construct was reflected in brain activity during naming, regardless of which items were being named or what overt responses were observed.

Discussion

The claim that the picture naming task relies on dissociable mental processes is not new (Lichtheim, 1885); however, the methods used to support this claim in the past, such as latent decomposition of a redundant test battery (Lambon Ralph et al., 2002) or computational modeling of a naming task with many trials and response types (Foygel & Dell, 2000), have been too cumbersome to translate into widely adopted and clinically useful tools. Here, we presented a methodological framework that enables clinicians to obtain relevant information about the source of impairments from simple naming accuracy scores by selecting items that challenge specific mental processes based on a cognitive model. We believe this procedure can improve efficiency and reliability over transcribing and scoring error types.

Potential Clinical Applications of the RLI Assessment

It is important to note that, at present, the RLI has only been evaluated in a research context, with the RLI items extracted from the larger set of PNT items that were administered in full (along with transcription and scoring of error types). While further clinical research to evaluate the efficacy and validity of the RLI assessment is warranted and ongoing, we can already envision potential applications in a clinical setting.

Picture naming tasks have many clinical applications, including assessment, diagnosis, prognosis, and monitoring of language abilities in the context of pathological conditions. The SCANT, a 20-item naming test, can be used to detect the presence of aphasia or detect changes in the severity of aphasia (Walker, Fridriksson, et al., 2022). In combination with the SCANT, the RLI scale enables each patient to be placed in a two-dimensional space characterizing both the severity of the impairment (SCANT) and the nature of impairment (RLI). The opposing ends of the RLI scale represent extremely specific damage to different mechanisms, as reflected in behavioral scores and neuroimaging of lesion damage and task-based activation, so it is reasonable to suspect that the effectiveness of specific treatment interventions might reflect these distinctions as well. Previous studies have not found reliable impairment-based predictors of responses to semantic versus phonologically oriented treatment (Abel et al., 2005; Kristinsson et al., 2023; Wambaugh et al., 2001), perhaps due to difficulty with identifying candidate recipients or due to difficulty with designing targeted treatments. This may be an area where the information provided by the RLI assessment can guide treatment research.

The RLI assessment is constructed based on the Walker et al. (2018) MPT model of latent error opportunities during word production, which itself is based on the spreading activation model (Foygel & Dell, 2000) that posits a relationship between overt error types and latent processing errors at different levels of psycholinguistic representation. The spreading activation model has been used to characterize the effects of anomia treatment, relying on response type scores (i.e., correct responses and five other error types) to estimate network connection strengths at the semantic and phonological levels of word retrieval. For example, Simic et al. (2020) found that treatment with phonological components analysis was effective and led to significantly increased estimates of lexical–phonological connection strength. Similarly, Bruehl et al. (2021) found that effective anomia therapy and generalization of treatment effects to untrained items corresponded with increased estimates of lexical–semantic connection strength. Exploring optimal cutoff times for naming in aphasia, Evans et al. (2020) found that people with higher estimates of lexical–semantic connection strength may benefit more from additional time to accumulate information and retrieve a word. In all these cases, information about the relative impact on lexical versus sublexical processing was helpful for explaining the effects of clinical interventions, and that is precisely the information that the RLI assessment provides, without the need for transcription, error type scoring, or a computational model. (Of course, the MPT model was essential for constructing the RLI assessment and can provide more detailed information at the expense of more data

collection and computational analysis.) The RLI assessment, therefore, may provide a simple way to validate these reported effects of treatment, by identifying patients with the most to gain from specific interventions or by serving as an outcome measure to detect specific effects.

Future Directions

The RLI assessment can be evaluated and enhanced in several ways moving forward. It is important to understand the effects of administering the items in a targeted set, rather than interleaved among many other items, as semantic interference effects in aphasia are well documented (Schnur, 2014; Schnur et al., 2006). It is unknown how this might impact RLI scores. Although the test was developed using data from patients with stroke aphasia, the conceptual model should extend to any disruption of the speech production system, whether developmental or degenerative, in individuals who are expected to be proficient with the core vocabulary of the language. Reliability is a major concern for clinical use, and the RLI assessment can, in theory, be augmented with more items to achieve better estimates. This would require evaluating these items in a large cohort of people with speech impairments to assess their difficulties empirically. This search process could be aided by identifying lexical properties that are associated with specific dimensions of difficulty, but ultimately, these items would need to be verified experimentally. Finally, clinical trials and feedback from stakeholders will be important for understanding its potential uses. This work is all currently underway.

Conclusions

Picture naming accuracy in aphasia is complex and multidimensional. The current study provides a novel way to detect RLI for word production in aphasia. This RLI is reflected in behavioral scores, lesion damage locations, and functional activations during naming. Although more research is warranted into the reliability, validity, and clinical applicability of the RLI assessment, our psychometric evaluations and validation studies endorse the RLI assessment as a potentially useful scale for identifying the nature of expressive language impairments in clinical or research settings.

Data Availability Statement

All data required to replicate the analyses are included in the article or the supplemental materials. Item-level picture naming data and preprocessed neuroimaging data are available from the authors upon request.

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders Grant P50 DC014664 (awarded to Julius Fridriksson) and the National Institute on Aging Grant 2RF1NS050915-18A1 (awarded to Maria Luisa Gorno Tempini and Gregory Hickok).

References

Abel, S., Grande, M., Huber, W., Willmes, K., & Dell, G. S. (2005). Using a connectionist model in aphasia therapy for naming disorders. *Brain and Language, 95*(1), 102–104. https://doi.org/10.1016/j.bandl.2005.07.056

Abel, S., Huber, W., & Dell, G. S. (2009). Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology, 23*(11), 1353–1378. https://doi.org/10.1080/02687030903022203

Abel, S., Weiller, C., Huber, W., Willmes, K., & Specht, K. (2015). Therapy-induced brain reorganization patterns in aphasia. *Brain, 138*(Pt. 4), 1097–1112. https://doi.org/10.1093/BRAIN/AWV022

ALHarbi, M. F., Armijo-Olivo, S., & Kim, E. S. (2017). Transcranial direct current stimulation (tDCS) to improve naming ability in post-stroke aphasia: A critical review. *Behavioural Brain Research, 332,* 7–15. https://doi.org/10.1016/J.BBR.2017.05.050

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., & Phillips, C. (2012). *SPM8 manual*. Functional Imaging Laboratory, Institute of Neurology.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage, 26*(3), 839–851. https://doi.org/10.1016/J.NEUROIMAGE.2005.02.018

Azhar, A., Maqbool, S., Butt, G. A., Iftikhar, S., & Iftikhar, G. (2017). Frequency of aphasia and its symptoms in stroke patients. *Journal of Speech Pathology & Therapy, 2*(1), 1–3. https://doi.org/10.4172/2472-5005.1000121

Bruehl, S., Willmes, K., & Binkofski, F. (2021). Interfered-Naming Therapy for Aphasia (INTA): Behavioural and computational effects of a novel linguistic-executive approach. *Aphasiology, 37*(2), 227–248. https://doi.org/10.1080/02687038.2021.1995841

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology, 28*(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Catani, M., & Thiebaut de Schotten, M. (2008). A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex, 44*(8), 1105–1132. https://doi.org/10.1016/J.CORTEX.2008.05.004

Cervenka, M. C., Boatman-Reich, D. F., Ward, J., Franaszczuk, P. J., & Crone, N. E. (2011). Language mapping in multilingual patients: Electrocorticography and cortical stimulation during naming. *Frontiers in Human Neuroscience, 5,* Article 13. https://doi.org/10.3389/fnhum.2011.00013

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

Conroy, P., Sage, K., & Ralph, M. L. (2009). Improved vocabulary production after naming therapy in aphasia: Can gains in picture naming generalize to connected speech? *International*

Journal of Language & Communication Disorders, 44(6), 1036–1062. https://doi.org/10.1080/13682820802585975

Cotelli, M., Manenti, R., Ferrari, C., Gobbi, E., Macis, A., & Cappa, S. F. (2020). Effectiveness of language training and non-invasive brain stimulation on oral and written naming performance in primary progressive aphasia: A meta-analysis and systematic review. *Neuroscience & Biobehavioral Reviews, 108,* 498–525. https://doi.org/10.1016/J.NEUBIOREV.2019.12.003

Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language, 56*(4), 490–520. https://doi.org/10.1016/j.jml.2006.05.007

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Branch Coslett, H. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition, 128*(3), 380–396. https://doi.org/10.1016/j.cognition.2013.05.007

Evans, W. S., Hula, W. D., Quique, Y., & Starns, J. J. (2020). How much time do people with aphasia need to respond during picture naming? Estimating optimal response time cutoffs using a multinomial ex-Gaussian approach. *Journal of Speech, Language, and Hearing Research, 63*(2), 599–614. https://doi.org/10.1044/2019_JSLHR-19-00255

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

Fergadiotis, G., & Wright, H. H. (2016). Modelling confrontation naming and discourse performance in aphasia. *Aphasiology, 30*(4), 364–380. https://doi.org/10.1080/02687038.2015.1067288

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language, 43*(2), 182–216. https://doi.org/10.1006/JMLA.2000.2716

Fridriksson, J., Baker, J. M., & Moser, D. (2009). Cortical mapping of naming errors in aphasia. *Human Brain Mapping, 30*(8), 2487–2498. https://doi.org/10.1002/HBM.20683

Fridriksson, J., Bonilha, L., Baker, J. M., Moser, D., & Rorden, C. (2010). Activity in preserved left hemisphere regions predicts anomia severity in aphasia. *Cerebral Cortex, 20*(5), 1013–1019. https://doi.org/10.1093/cercor/bhp160

Fridriksson, J., Richardson, J. D., Fillmore, P., & Cai, B. (2012). Left hemisphere plasticity and aphasia recovery. *NeuroImage, 60*(2), 854–863. https://doi.org/10.1016/J.NEUROIMAGE.2011.12.057

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

Geschwind, N. (1965). Disconnexion syndromes in animals and man. II. *Brain, 88*(3), 585–644. https://doi.org/10.1093/brain/88.3.585

Gordon, J. K. (2020). Factor analysis of spontaneous speech in aphasia. *Journal of Speech, Language, and Hearing Research, 63*(12), 4127–4147. https://doi.org/10.1044/2020_JSLHR-20-00340

Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2017). Triangulation of language–cognitive impairments, naming errors and their neural bases post-stroke. *NeuroImage: Clinical, 17,* 465–473. https://doi.org/10.1016/J.NICL.2017.10.037

Haley, K. L., & Jacks, A. (2023). Three-dimensional speech profiles in stroke aphasia and apraxia of speech. *American Journal of Speech-Language Pathology, 32*(4S), 1825–1834. https://doi.org/10.1044/2022_AJSLP-22-00170

Hickok, G. (2022). The dual stream model of speech and language processing. *Handbook of Clinical Neurology, 185,* 57–69. https://doi.org/10.1016/B978-0-12-823384-9.00003-7

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1–2), 67–99. https://doi.org/10.1016/J.COGNITION.2003.10.011

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hillis, A. E., Oh, S., & Ken, L. (2004). Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology, 55*(2), 268–275. https://doi.org/10.1002/ANA.10812

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., Panesar, S., Gravier, M. L., Yeh, F. C., Dresang, H. C., Dickey, M. W., & Fernandez-Miranda, J. C. (2020). Structural white matter connectometry of word production in aphasia: An observational study. *Brain, 143*(8), 2532–2544. https://doi.org/10.1093/BRAIN/AWAA193

Hurley, R. S., Paller, K. A., Rogalski, E. J., & Mesulam, M. M. (2012). Neural mechanisms of object naming and word comprehension in primary progressive aphasia. *Journal of Neuroscience, 32*(14), 4848–4855. https://doi.org/10.1523/JNEUROSCI.5984-11.2012

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage, 62*(2), 782–790. https://doi.org/10.1016/J.NEUROIMAGE.2011.09.015

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mazoyer, B., & Tzourio-Mazoyer, N. (2015). AICHA: An atlas of intrinsic connectivity of homotopic areas. *Journal of Neuroscience Methods, 254,* 46–59. https://doi.org/10.1016/j.jneumeth.2015.07.013

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

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., & 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

Kristinsson, S., Basilakos, A., den Ouden, D. B., Cassarly, C., Spell, L. A., Bonilha, L., Rorden, C., Hillis, A. E., Hickok, G., Johnson, L., Busby, N., Walker, G. M., McLain, A., & Fridriksson, J. (2023). Predicting outcomes of language rehabilitation: Prognostic factors for immediate and long-term

outcomes after aphasia therapy. *Journal of Speech, Language, and Hearing Research, 66*(3), 1068–1084. https://doi.org/10.1044/2022_JSLHR-22-00347

Kristinsson, S., Zhang, W., Rorden, C., Newman-Norlund, R., Basilakos, A., Bonilha, L., Yourganov, G., Xiao, F., Hillis, A., & Fridriksson, J. (2021). Machine learning-based multimodal prediction of language outcomes in chronic aphasia. *Human Brain Mapping, 42*(6), 1682–1698. https://doi.org/10.1002/HBM.25321

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

Lausberg, H., Göttert, R., Münssinger, U., Boegner, F., & Marx, P. (1999). Callosal disconnection syndrome in a left-handed patient due to infarction of the total length of the corpus callosum. *Neuropsychologia, 37*(3), 253–265. https://doi.org/10.1016/S0028-3932(98)00079-7

Levelt, W. J. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America, 98*(23), 13464–13471. https://doi.org/10.1073/pnas.231459498

Lichtheim, L. (1885). On aphasia. *Brain, 7*(4), 433–484. https://doi.org/10.1093/brain/7.4.433

Mason, C., & Nickels, L. (2022). Are single-word picture naming assessments a valid measure of word retrieval in connected speech? *International Journal of Speech-Language Pathology, 24*(1), 97–109. https://doi.org/10.1080/17549507.2021.1966098

Matti, L., Anneli, T., & Martti, J. (1998). Modelling anomia by the discrete two-stage word production architecture. *Journal of Neurolinguistics, 11*(3), 275–294. https://doi.org/10.1016/S0911-6044(97)00015-8

Mayer, J., & Murray, L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology, 17*(5), 481–497. https://doi.org/10.1080/02687030344000148

Nachev, P., Coulthard, E., Jäger, H. R., Kennard, C., & Husain, M. (2008). Enantiomorphic normalization of focally lesioned brains. *NeuroImage, 39*(3), 1215–1226. https://doi.org/10.1016/J.NEUROIMAGE.2007.10.002

Osa García, A., Brambati, S. M., Brisebois, A., Désilets-Barnabé, M., Houzé, B., Bedetti, C., Rochon, E., Leonard, C., Desautels, A., & Marcotte, K. (2020). Predicting early post-stroke aphasia outcome from initial aphasia severity. *Frontiers in Neurology, 11,* Article 120. https://doi.org/10.3389/fneur.2020.00120

Pagnoni, I., Gobbi, E., Premi, E., Borroni, B., Binetti, G., Cotelli, M., & Manenti, R. (2021). Language training for oral and written naming impairment in primary progressive aphasia: A review. *Translational Neurodegeneration, 10*(1), Article 24. https://doi.org/10.1186/s40035-021-00248-z

Postman-Caucheteux, W. A., Birn, R. M., Pursley, R. H., Butman, J. A., Solomon, J. M., Picchioni, D., McArdle, J., & Braun, A. R. (2010). Single-trial fMRI shows contralesional activity linked to overt naming errors in chronic aphasic patients. *Journal of Cognitive Neuroscience, 22*(6), 1299–1318. https://doi.org/10.1162/JOCN.2009.21261

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*(3), 460–499. https://doi.org/10.1037/0033-295X.107.3.460

Richardson, J. D., 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

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 [PDF]

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., & Karnath, H. O. (2012). Age-specific CT and MRI templates for spatial normalization. *NeuroImage, 61*(4), 957–965. https://doi.org/10.1016/J.NEUROIMAGE.2012.03.020

Schnur, T. T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language, 75,* 27–44. https://doi.org/10.1016/j.jml.2014.04.006

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language, 54*(2), 199–227. https://doi.org/10.1016/J.JML.2005.10.002

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from aphasia. *Brain, 132*(Pt. 12), 3411–3427. https://doi.org/10.1093/brain/awp284

Sebastian, R., & Kiran, S. (2011). Task-modulated neural activation patterns in chronic stroke patients with aphasia. *Aphasiology, 25*(8), 927–951. https://doi.org/10.1080/02687038.2011.557436

Simic, T., Chambers, C., Bitan, T., Stewart, S., Goldberg, D., Laird, L., Leonard, C., & Rochon, E. (2020). Mechanisms underlying anomia treatment outcomes. *Journal of Communication Disorders, 88,* Article 106048. https://doi.org/10.1016/J.JCOMDIS.2020.106048

Sinai, A., Bowers, C. W., Crainiceanu, C. M., Boatman, D., Gordon, B., Lesser, R. P., Lenz, F. A., & Crone, N. E. (2005). Electrocorticographic high gamma activity versus electrical cortical stimulation mapping of naming. *Brain, 128*(Pt. 7), 1556–1570. https://doi.org/10.1093/BRAIN/AWH491

Skipper-Kallal, L. M., Lacey, E. H., Xing, S., & Turkeltaub, P. E. (2017). Right hemisphere remapping of naming functions depends on lesion size and location in poststroke aphasia. *Neural Plasticity, 2017,* Article 8740353. https://doi.org/10.1155/2017/8740353

Stefaniak, J. D., Alyahya, R. S. W., & Lambon Ralph, M. A. (2021). Language networks in aphasia and health: A 1000 participant activation likelihood estimation meta-analysis. *NeuroImage, 233,* Article 117960. https://doi.org/10.1016/J.NEUROIMAGE.2021.117960

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The Apraxia of Speech Rating Scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders, 51,* 43–50. https://doi.org/10.1016/J.JCOMDIS.2014.06.008

Swiderski, A. M., Hula, W. D., & Fergadiotis, G. (2023). Accuracy of naming error profiles elicited from adaptive short forms of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 66*(4), 1351–1364. https://doi.org/10.1044/2023_JSLHR-22-00439

Tisak, J., & Smith, C. S. (1994). Defending and extending difference score methods. *Journal of Management, 20*(3), 675–682. https://doi.org/10.1177/014920639402000310

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language, 162,* 60–71. https://doi.org/10.1016/J.BANDL.2016.08.004

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI

single-subject brain. *NeuroImage, 15*(1), 273–289. https://doi.org/10.1006/NIMG.2001.0978

Walker, G. M. (2021). Disentangling the psycholinguistic loci of anomia with cognitive psychometric models. *Seminars in Speech and Language, 42*(3), 256–274. https://doi.org/10.1055/s-0041-1731367

Walker, G. M., Basilakos, A., Fridriksson, J., & Hickok, G. (2022). 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., Fridriksson, J., & Hickok, G. (2021). Connections and selections: Comparing multivariate predictions and parameter associations from latent variable models of picture naming. *Cognitive Neuropsychology, 38*(1), 50–71. https://doi.org/10.1080/02643294.2020.1837092

Walker, G. M., Fridriksson, J., Hillis, A. E., den Ouden, D. B., Bonilha, L., & Hickok, G. (2022). The Severity-Calibrated Aphasia Naming Test. *American Journal of Speech-Language Pathology, 31*(6), 2722–2740. https://doi.org/10.1044/2022_AJSLP-22-00071

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)

Wambaugh, J. L., Linebaugh, C. W., Doyle, P. J., Martinez, A. L., Kalinyak-Fliszar, M., & Spencer, K. A. (2001). Effects of two cueing treatments on lexical retrieval in aphasic speakers with different levels of deficit. *Aphasiology, 15*(10–11), 933–950. https://doi.org/10.1080/02687040143000302

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yourganov, G., Fridriksson, J., Stark, B., & Rorden, C. (2017). Removal of artifacts from resting-state fMRI data in stroke. *NeuroImage: Clinical, 17,* 297–305. https://doi.org/10.1016/J.NICL.2017.10.027