

Outcomes from a pilot dose comparison study of naming therapy in aphasia

Sam Harvey^{a,b}, Marcella Carragher^{a,b}, Michael Walsh Dickey^{b,c,d}, and Miranda L. Rose^{a,b}

^a*School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Australia;*

^b*Centre of Research Excellence in Aphasia Recovery and Rehabilitation, Australia;* ^c*Geriatric Research Education and Clinical Center, VA Pittsburgh Healthcare System, USA;* ^d*University of Pittsburgh, USA*

This is the Accepted Version of the manuscript reproduced under a CC BY-NC-ND license. Please visit <https://doi.org/10.1080/02687038.2022.2144112> to access the Publisher's Version of Record.

Abstract

Background: People with aphasia vary considerably in response to aphasia treatments. Treatment dose is likely an important factor in understanding treatment response variability and optimising aphasia recovery; however, there is limited empirical evidence to guide dose prescription in post-stroke aphasia rehabilitation. In the present study, we used a novel approach to personalise dose prescription and explored the effect of dose on treatment response in chronic post-stroke aphasia.

Aims: Examine the effect of providing personalised doses of a cued picture naming treatment (Kendall et al., 2014) on acquisition and maintenance of picture naming outcomes.

Method: This pilot study used a multiple-baselines design with follow-up at 4- and 12-weeks with replication across four people with chronic post-stroke anomia. Prior to treatment, a comprehensive battery of cognitive and language tests was completed. Participants then undertook a period of cued picture naming treatment (45-minute sessions, five days per week for three weeks) totalling 15 sessions (11.25 hours). Participants were allocated four picture sets – one each for three treated conditions (low dose, moderate dose, and high dose) and one for an untreated control set. The number of naming opportunities provided per dose condition was calibrated against individuals' pre-treatment picture naming accuracy and speed. Generalised linear mixed effects models were used to evaluate learning effects during treatment, maintenance of these effects, and dose-response relationships.

Outcomes: Participants received 99% of prescribed treatment doses (i.e., number of naming opportunities provided over the course of treatment). As anticipated, individual treatment responses varied substantially. Three participants demonstrated significantly improved picture naming accuracy on probed items during treatment, with varying response profiles by participant and dose. All participants were able to name more pictures accurately on a bespoke 298-item object picture naming test following treatment. However, no participant demonstrated significant pre-post treatment gains relative to untreated items, although one person demonstrated improved naming 4-weeks after treatment for items treated under the high dose condition. Dose-response relationships amongst these participants exhibited a greater number of significant results on naming probes in the high dose condition, possibly suggesting superiority of the high dose condition over lower doses of cued picture naming treatment.

Conclusion: Modest treatment effects and variable dose-response relationships were observed. We explore the role of dose, cognitive factors such as self-monitoring abilities, and linguistic factors

such as underlying lexical-semantic and phonological processing that may have influenced treatment response in these participants. Avenues for future research are identified.

Keywords

Aphasia, rehabilitation, treatment, personalisation, dose, single-case experimental design

Abbreviations

CDC	Conservative Dual Criterion
PNT	Philadelphia Naming Test
SMD	Standardized mean difference
TEA	Test of Everyday Attention
WAB-R	Western Aphasia Battery-Revised

Aphasia is a language impairment that causes communication disability for an estimated 4.5 million stroke survivors world-wide (Johnson et al., 2019). People with aphasia experience difficulties producing and understanding spoken and written language due to damage to language processing networks in the brain. High level evidence supports the effectiveness of aphasia treatments (Brady et al., 2016); however, individual response to treatment is highly variable (Menahemi-Falkov et al., 2021). Treatment dose is likely an important factor in understanding treatment response variability and optimising aphasia recovery (Brady et al., 2021).

Evaluation of dose-response relationships in aphasia rehabilitation has focused on the role of time in treatment, with little focus on the actions performed over the course of treatment (Brady et al., 2021; Brady et al., 2016; Harvey et al., 2021). However, in the context of complex behavioural interventions targeting post-stroke language impairment, dose is a multidimensional construct (Baker, 2012; Togher, 2012) which has typically been underspecified in aphasia rehabilitation research (Harvey et al., 2021). Multiple frameworks for conceptualizing treatment dose (Warren et al., 2007; Baker, 2012; Hayward, et al., 2021) have converged on the number of *episodes* of treatment-related activity over the course of intervention as key to measuring and understanding the effect of dose of outcomes. Episodes are periods of time within treatment sessions that contain the therapeutic activities presumed to affect brain and behaviour change (Hayward et al., 2021). Treatment schedules that provide a high number of episodes each session have been described as “saturated practice” (Harnish et al., 2013, p. s287). Harnish and colleagues reported a case series (n=8) exploring the effect of saturated practice of a cued picture naming treatment on the acquisition, maintenance, and generalisation of picture naming abilities in post-stroke aphasia (Harnish et al., 2013). The study implemented a protocolised cued picture naming treatment (Kendall et al., 2014) that focuses on practicing the phonological wordform and includes elements of semantic processing and verbal working memory to facilitate restoration of lexical retrieval abilities. This treatment has been shown to produce modest improvements in picture naming abilities in people with chronic post-stroke aphasia (Harnish et al., 2013; Kendall et al., 2014). The treatment procedure used by Harnish et al. (2013) provided eight opportunities to produce the name of each picture in response to a variety of cues. In each 60-minute session 50 pictures were practiced once per session, resulting in 400 opportunities to produce picture names per session. Six participants achieved significant gains in picture naming accuracy after just one treatment session and the remaining two participants achieved significant gains after three sessions. Six of the seven participants with follow up measures maintained these gains on trained items, and two of seven on untrained items, at approximately 60-days follow up.

Building on these preliminary findings, Off and colleagues (2016) compared the effects of lower and higher number of naming attempts on confrontation naming for people with chronic aphasia (n = 7). Pictures in the low-dose condition (n = 20) were presented once per session, whereas pictures in the high-dose condition (n = 20) were shown four times. Each picture presentation involved two naming attempts, one cued and one uncued, resulting in 40 naming attempts per low-dose condition and 160 per high-dose condition per session. The high-dose condition resulted in large effect sizes for two participants and a small effect size for one whereas the low-dose condition resulted in a medium effect size for one participant, relative to lexical retrieval benchmarks. We previously meta-analysed findings from the Harnish et al. and Off et al. studies (Harvey et al., 2022) and found no significant differences in outcomes between doses provided in these two studies.

There is limited empirical evidence to guide dose prescription in aphasia (Doogan et al., 2018; Harvey et al., 2021). Harnish and colleagues (2013) justified the 400 naming opportunities provided per session referring to findings from the motor learning and neuroplasticity literature that a skilled reaching task delivered 400 times per day elicited increases in the number of synapses in the rodent motor cortex (Kleim et al., 2002), whereas the same task delivered 60 times per day did not (Luke et al., 2004). One approach to dose exploration would be to determine a *theoretical* maximum dose (i.e., the maximum amount of practice that a person could complete in a given amount of time), calibrate different doses relative to this theoretical maximum, and compare the different doses in a single study. An advantage of this approach is that it allows personalisation of dose; in the present study, we used a novel approach to determine a theoretical maximum dose of a cued picture naming treatment (Kendall et al., 2014) and calibrate different doses relative to an individual's pre-treatment picture naming skills to explore the effect of dose on treatment response in post-stroke aphasia.

Aims and hypotheses

This Phase I pilot study was designed to experimentally examine dose-response relationships in cued picture naming treatment for people with chronic post-stroke aphasia. There were two research questions: (RQ1) Is there a significant effect of treatment over no treatment when treatment is provided at different doses? (RQ2) Is there significant difference in maintenance of treatment response between doses? We had three primary hypotheses: (H1) Naming accuracy would improve with treatment; (H2) There would be differences in the magnitude of treatment-related gains in naming accuracy across dose conditions; (H3) Maintenance of treatment effects may be associated with the magnitude of change in naming accuracy during treatment and would be expected to diminish over time.

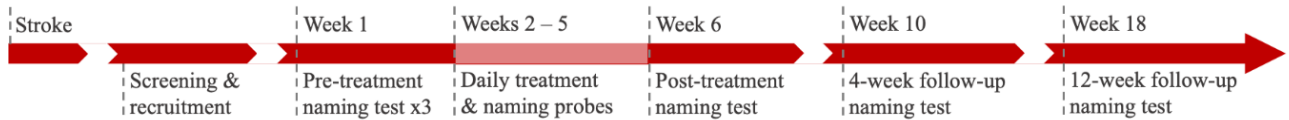
Regarding dose-response relationships, there is mounting evidence that more time in treatment produces better treatment outcomes (Brady et al., 2021). Therefore, we anticipated that improvements in naming accuracy would accrue with each treatment session and would peak towards the end of the treatment phase (H4). The current study was duration-controlled (i.e., each experimental condition was treated for the same amount of time) while the number of *episodes* provided each session was manipulated (see Method below). There is equivocal evidence regarding treatment outcomes in response to different numbers of episodes within picture naming treatments (Harvey et al., 2022); our tentative *a priori* hypothesis was that acquisition of picture naming skills would favour schedules providing a higher number of episodes (H5), in line with empirical evidence supporting massed practice schedules in aphasia treatment (Crosson et al., 2019; Kiran & Thompson, 2019).

There is limited evidence to guide hypotheses regarding the effect of dose on maintenance of picture naming treatment outcomes. Maintenance of gains following lexical retrieval treatments have been associated both with acquisition effect sizes (e.g., Menahemi-Falkov et al., 2021), potentially favouring higher doses in the current study, and with distributed practice schedules (Dignam et al., 2016), potentially favouring lower doses in the current study. Therefore, we had no *a priori* hypothesis regarding the relative superiority of doses on maintenance of treatment effects.

Method

This pilot study used a multiple-baselines design with follow-up at 4- and 12-weeks post intervention (Figure 1). Approval was granted by the La Trobe University Human Research Ethics Committee (HEC20414). Detailed description of the study method is provided in the supplemental material and will be summarised here.

Figure 1 Study schedule illustrating the four assessment time points around the multiple-baselines experiment



Participants

Four adults with chronic post-stroke aphasia were recruited from the community in a convenience sample. Participants primarily spoke English in the home; two participants (MR and TS) spoke other languages on a weekly basis (e.g., phone call to family), and one participant (LR) spoke another language on rare occasions (i.e., less than monthly). Participant demographic characteristics are shown in Table 1. Participants did not present with severe apraxia of speech, uncorrected vision or hearing problems, self-reported history of diffuse neurological injury or disease, or psychological disorder, and none of the participants received impairment-based intervention outside of the study for the duration of the research. Eligible participants provided informed written consent to participate. A detailed description of each participant is provided Appendix A.

Table 1 Participant demographic data

	MR	IP	LR	TS
Age	54	68	72	78
Sex	Male	Female	Male	Male
Years of education	15 years	10 years	15 years	19 years
Ethnicity	Asian	Caucasian	Caucasian	Asian
Language(s) spoken at home	English, Urdu	English	English	English, Cantonese
Occupation prior to stroke	Web developer	Cleaner	Retired, engineer	Retired, dentist
Premorbid handedness	Right	Right	Right	Right
Stroke type	Infarct	Infarct	Unknown	Infarct
Months since stroke	119	23	26	27
Residual hemiparesis	Right	Right	None	Dense right

Treatment

This study implemented the cued picture naming treatment protocol described by Kendall and colleagues (2014). This treatment is ideal for early-stage exploration of dose-response relationships because it has a simple structure with discretely identifiable *episodes* and active ingredients. Furthermore, reported effect sizes are relatively modest (Kendall et al., 2014) meaning the effect of modulating certain aspects of the treatment – such as its dose – may be more easily observed in this treatment than in a treatment with a consistently large treatment effect.

One treatment session was provided each weekday over a three-week period (one week/five sessions per dose condition). Sessions varied in length but included exactly 45-minutes of active therapy (total 3.75 hours per dose condition) with variable time spent inactive due to breaks. Treatment was conducted online with participants attending from their homes. A purpose-built web-based software programme delivered all aspects of each treatment session. The software was constructed to ensure protocol adherence across participants, sessions, and dose conditions. Additional description of the software is provided in the supplemental material.

In the current study, a predetermined number of episodes of cued picture naming treatment was delivered each treatment session. Each episode was identical in structure. First, a picture was presented followed by a series of eight cues (confrontation, orthographic, spoken repetition, delayed recall, semantic, phonological and phonemic, spoken repetition, delayed recall) presented at a predetermined rate. The order of cues was identical in each episode and each subsequent cue was provided regardless of picture naming accuracy. Each cue was followed by an opportunity to produce the name of the picture. The amount of time provided for each naming opportunity was prescribed to participants based on their individual pre-treatment naming speed (see *Dose calibration* below). Participants were instructed to make one naming attempt per opportunity. The clinician provided general encouragement to facilitate engagement. Feedback on performance was not provided. Participants were encouraged to rest for 5-seconds between each episode but were able to take breaks at any point and for any duration between episodes.

Outcome measures

The primary outcome used to evaluate dose-response relationships was item-level picture naming accuracy on daily picture naming probes during the multiple-baselines study. The secondary outcome used to examine maintenance of treatment effects was picture naming accuracy on a purpose-built picture naming test conducted immediately pre- and post-treatment, and at 4-week and 12-week follow up. A speech-language pathologist (first author) with over 10 years' experience working with people who have aphasia conducted the assessments and monitored all treatment sessions remotely using videoconferencing software. Language and cognition skills that are known to explain variation in picture naming treatment outcomes were assessed during screening and pre-treatment (see Figure 1) with modifications to allow remote assessment (see supplemental material for a description of these modifications).

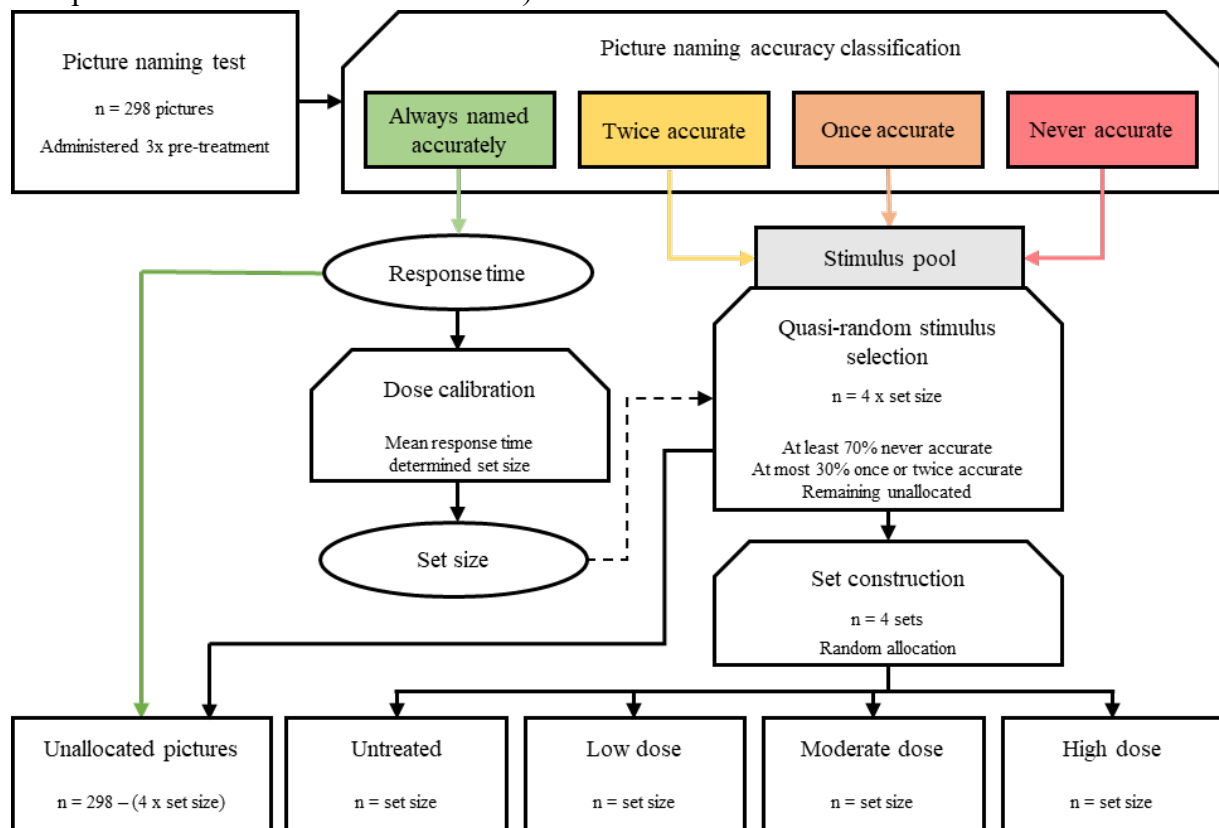
Picture naming test

A purpose-built test comprising 298 colour photos depicting objects was developed to assess picture naming accuracy across time and to select treatment stimuli. Pictures were obtained from the Bank of Standardized Stimuli (Brodeur et al., 2010) along with stimulus-specific name agreement and word length data (number of phonemes: mean 5, range 1-11). Word frequencies (mean 2.23, range 0.30-3.81) were obtained from the Log10CD measure of the SUBTLEXus database (Brysbaert & New, 2009). The picture naming test was administered three times across three days at the pre-treatment time point and once at each of the other assessment time points. Pictures were presented in random order each administration.

Stimulus pictures used in the multiple-baselines study were selected for each participant from this 298-item test based on their picture naming accuracy and speed (Figure 2). Following

three administrations of the picture naming test, pictures were classified based on accuracy across the three tests. Mean response time for pictures that were always named accurately was calculated and those pictures were set aside (unallocated). The dose calibration procedure (see *Dose calibration*) used mean response time to determine how many episodes of treatment would be provided each dose condition and, subsequently, how many pictures would be allocated to each of the four stimulus sets. A quasi-random selection of pictures that were not always named accurately during pre-treatment testing was made and these pictures were divided into four sets, balanced for naming accuracy classification and lexical properties (word frequency, word length in phonemes). To enhance participant motivation, some pictures that were named accurately at least once were included to ensure participants experienced successful naming attempts during treatment. The four stimulus sets were then randomly allocated to one of the four treatment conditions.

Figure 2 Series of operations used to allocate pictures to stimulus sets used in treatment (Harvey, 2022 reproduced under CC-BY 4.0 license)



Picture naming probes

At the beginning of each session during the multiple-baselines study, picture naming accuracy was probed using all the allocated stimulus pictures that were never named accurately during pre-treatment testing. During probes, a randomly selected picture was presented for 12 seconds without cues. Probe naming accuracy was the dependent variable in this study and was rated according to Philadelphia Naming Test (PNT) scoring criteria (Roach et al., 1996).

Reliability of picture naming ratings

Picture naming accuracy was rated online by SH. It was not possible to blind study participants nor the clinician conducting the training and real-time picture naming accuracy. A random selection (20%) of the naming probe data and picture naming test data for each participant was double-rated from video recordings by a trained assessor blinded to assessment time point and treatment condition (probe data only). Reliability calculations (Cohen's kappa; Landis & Koch, 1977) demonstrated almost perfect inter-rater agreement across the four participants' naming data (MR: 0.97; IP: 0.94; LR: 0.87; TS: 0.90). The first author rescored a random selection of 10% of picture naming data to determine intra-rater reliability which demonstrated almost perfect agreement (Cohen's kappa = .99).

Additional language and cognitive tests

The *Western Aphasia Battery-Revised* (WAB-R; Kertesz, 2007) was used to determine aphasia severity and classification. The WAB-R has been validated with modifications for remote administration (Dekhtyar et al., 2020). The *Philadelphia Naming Test* (PNT; Roach et al., 1996) was used to determine the severity and nature of anomia. The *Test of Everyday Attention* (Robertson et al., 1994) Elevator Counting subtest was used to test sustained attention. Immediate verbal memory was assessed using the *Picture span forward test* (DeDe et al., 2014). A modified version of the *Corsi Block Tapping Test* (Kessels et al., 2000) was used to assess non-verbal memory. Pre-treatment language and cognitive test results are shown in Table 2.

Table 2 Pre-treatment test results

Test (possible range)	MR	IP	LR	TS
WAB-R classification	Broca's	Wernicke's	Conduction	Broca's
WAB-R aphasia quotient (0-100)	66.8	53.2	63.2	64.8
Aphasia severity	Mild	Moderate	Moderate	Moderate
Apraxia Severity Rating Scale				
<i>Score >8 suggestive of apraxia of speech (Strand et al., 2014)</i>	3	0	0	0
TEA Elevator Counting (0-7)	3	2	6	7
<i>Score <6 indicative of reduced sustained attention (Robertson et al., 1994)</i>				
Picture span forward (0-100)	32	10	14	16
<i>Score <29 indicative of reduced verbal working memory (DeDe et al., 2014)</i>				
Modified Corsi Block Tapping Test (0-9)	3	3	4	3
<i>Score <5 indicative of impaired visuo-spatial working memory (Kessels et al., 2000)</i>				
Philadelphia Naming Test (0-175)	119	36	95	113
s-weight	0.021	0.001	0.021	0.019
p-weight	0.032	0.018	0.032	0.022
Mean (range) correct responses on naming battery (0-298)	161 (156-165)	45 (37-51)	144 (128-162)	153 (140-160)
Mean response time (seconds) for accurately named items	3.16	2.24	3.58	2.81

Notes: WAB-R=Western Aphasia Battery-Revised, TEA=Test of Everyday Attention, n.r.=no result

Aphasia severity rating based on the following WAB-R aphasia quotient score ranges: mild (66-93.7), moderate (33-65), severe (0-32).

s-weight and p-weight calculated using the Webfit algorithm available at <http://langprod.cogsci.illinois.edu/cgi-bin/webfit.cgi>

Dose calibration

The dose dimension being manipulated was the *episode* (Hayward et al., 2021). Each dose condition was operationalised as ‘the number of episodes provided in a session’. There were three dose conditions and one untreated condition. Under the high dose condition, each picture was shown three times per session. Under the moderate dose condition, each picture was shown twice per session. Under the low dose condition, each picture was shown once per session.

The number of pictures allocated to each participant was calibrated against their individual response time for accurately named pictures in pre-treatment testing (Table 3); people who were faster to name pictures accurately were allocated larger picture sets than people who were slower to respond accurately. Each picture in the high dose condition was treated three times per session which provided 120 naming opportunities per picture per week. Pictures in the moderate dose condition were treated twice per session (80 opportunities per picture per week). Pictures in the low dose condition were treated once per session (40 opportunities per picture per week). Each session consisted of 45-minutes of cued picture naming treatment (time-on-task), with additional time for breaks as required. A comprehensive description of the treatment dose prescribed and received is provided in Appendix B.

Table 3 Summary of each participant's response times, allocated items per set, and prescribed number of episodes and naming opportunities per session and per week

Participant	RT	Set size (probe items)	Episodes per session (week)			Naming opportunities per session (week)		
			Low dose	Mod. dose	High dose	Low dose	Mod. dose	High dose
MR	3.16	15 (11)	15 (75)	30 (150)	45 (225)	120 (600)	240 (1200)	360 (1800)
IP	2.24	17 (15)	17 (85)	34 (170)	51 (255)	136 (680)	272 (1360)	408 (2040)
LR	3.58	15 (15)	15 (75)	30 (150)	45 (225)	120 (600)	240 (1200)	360 (1800)
TS	2.81	16 (11)	16 (80)	32 (160)	48 (240)	128 (640)	256 (1280)	384 (1920)

NB: RT = average response time for accurately named items on the pre-treatment picture naming test; Mod. = Moderate

Items in the low dose condition were treated once per session, items in the moderate dose condition were treated twice per session, and items in the high dose condition were treated three times per session.

Example dose calibration

Formula to calculate the estimated theoretical maximum number of episodes for a session:

$$\begin{aligned} \text{Episodes per session} &= (60/\text{episode duration in seconds}) \times \text{session duration in minutes} \\ \text{Episode duration} &= \text{total cue duration in seconds per episode} + (\text{number of} \\ &\quad \text{naming opportunities per episode} \times \text{response time}) + \text{buffer} \end{aligned}$$

The total cue duration (28 seconds) and buffer (five seconds) were the same for each episode.

Hypothetical mean pre-treatment response time for accurately named pictures = 2.0 seconds.

$$\begin{aligned} \text{Episodes per session} &= (60/(28 + (8 \times 2.0) + 5)) \times 45 \\ &= 55 \end{aligned}$$

The estimated theoretical maximum number of episodes of cued picture naming treatment in a 45-minute session for a person with average pre-treatment response time of 2.0 seconds is 55. To calibrate low, moderate, and high doses relative to the theoretical maximum, take the highest number less than this maximum that is evenly divisible by three (i.e., the number of treatment conditions). In this case, 54.

$$\begin{aligned} \text{Set size} &= 54/3 \\ &= 18 \end{aligned}$$

Low dose: 18 episodes, each picture shown once per session, 18.8 seconds per naming opportunity.

Moderate dose: 36 episodes, each picture shown twice, 9.4 seconds per naming opportunity.
High dose: 54 episodes, each picture shown three times, 6.25 seconds per naming opportunity.

Stimulus sets

Each participant had four sets of pictures: one for each dose condition (low, medium, high) and one untreated (Figure 2). The number of pictures in the allocated and unallocated groups was different for each participant; IP had 68 allocated and 230 unallocated pictures, MR and TS had 64 allocated and 234 unallocated pictures, and LR had 60 allocated and 238 unallocated pictures. Following allocation of pictures to stimulus sets, the difference in mean word frequency and mean word length of each set was tested; there was no significant difference between sets within participants. Finally, the order that dose conditions were treated was randomised. Each participant's stimulus sets are listed in the supplemental material.

Analyses

Following visual inspection of picture naming data, dose-response relationships were examined via two separate mixed-effects modelling approaches using naming probe data (Analysis 1, primary analysis) and picture naming test data (Analysis 2, secondary analysis). Statistical analyses were conducted using the lme4 package (Bates, Maechler, et al., 2015) in R (R Core Team, 2013) implemented in RStudio Version 3.6.1. Model specification was data-driven following the method of parsimonious mixed models (Bates, Kliegl, et al., 2015), chosen to limit the possibility of Type I error. We chose not to apply corrections for multiple comparisons. A detailed summary of the analysis workflow is available via the supplemental material.

Analysis 1 – Learning effects throughout treatment

Analysis 1 modelled treatment response and dose effects (RQ1) by examining day-by-day learning effects throughout treatment, estimating item-level naming accuracy after each subsequent treatment day relative to baseline across the three dose conditions, including comparisons with the untreated condition. This approach is analogous to analyses based on visual inspection such as Conservative Dual-Criterion analysis (Fisher et al., 2003) and non-parametric statistical approaches such as Tau-U or Non-Overlap of Pairs (Parker et al., 2011) that attempt to determine whether performance during the treatment phase significantly exceeds performance during the baseline phase. We used generalised logistic mixed effects models including an ordinal *times treated* variable (which increments with each treatment day). Using an ordinal variable avoids the assumption of linearity during treatment, allowing estimation of the difference in the likelihood of a correct response in a naming probe after each treatment day, compared to baseline. The reference level for these models was the untreated condition at baseline (i.e., *times treated* = 0). We selected the untreated condition as the reference level in order to control for possible practice effects associated with repeated naming probes (Nickels, 2002). We included fixed-effects for item-level (word frequency, word length in phonemes) characteristics. The random-effects structure included random intercepts *by-item*. Models including random slopes for fixed effects and interaction terms failed to converge. Models of identical structure were fitted to each participant's data separately

allowing within-subject estimation of the effect of treatment versus no treatment for each dose condition and comparison of the effect of each dose condition.

Analysis 2 – Acquisition and maintenance of treatment effects

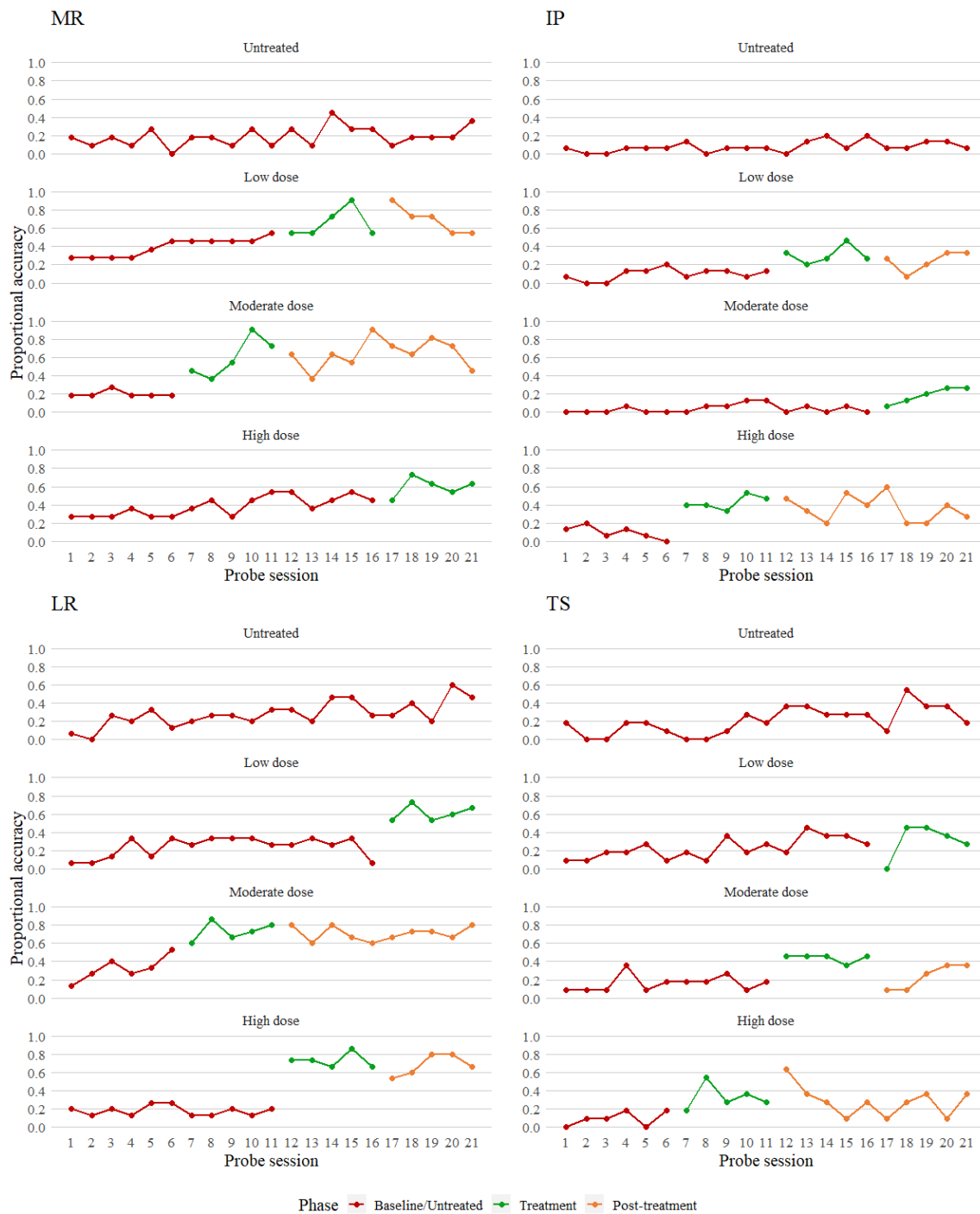
Analysis 2 modelled treatment response, dose effects, and maintenance of these effects (RQ2) using secondary outcome data from the 298-item object picture naming test conducted pre-treatment, immediately post-treatment, and at two follow-up time points. A generalised linear mixed effects model was constructed to test for main effects of *dose* and *time* and their interactions, with random intercepts *by-item*. These models included fixed effects for item-level lexical properties (word frequency, phoneme length). Models with identical fixed-effects structure were fitted to each participant's data separately allowing within-subject estimation of the effect of treatment versus no treatment for each dose condition and comparison of the effect of each dose condition. Picture naming data for items that were not allocated to stimulus sets were excluded from these analyses.

Results

Learning effects throughout treatment (Analysis 1)

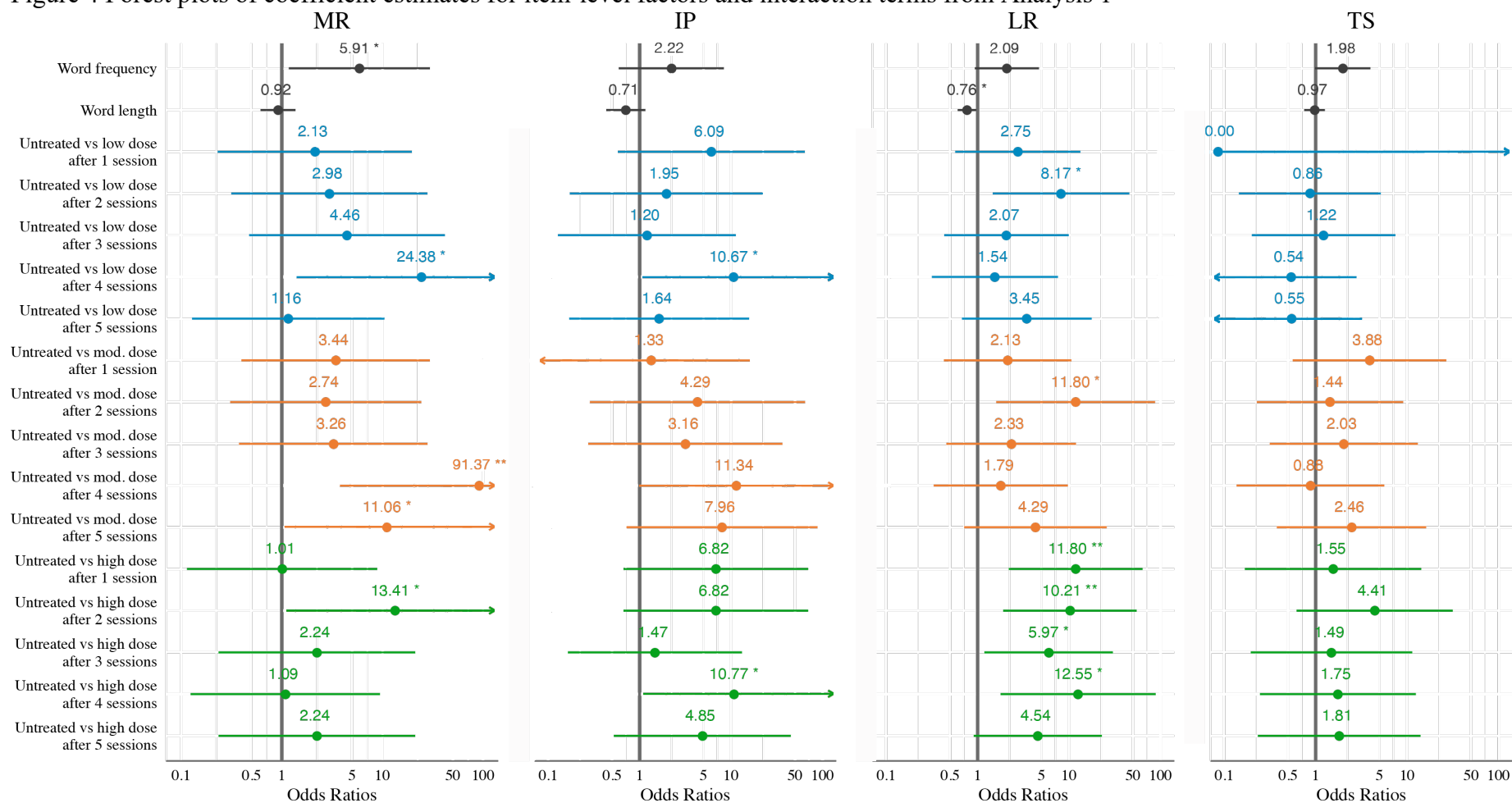
Figure 3 shows the proportion of pictures named accurately each probe session, by dose condition for each participant.

Figure 3 Multiple-baselines case charts



Analysis 1 used these data to model learning effects throughout treatment by estimating naming accuracy after each subsequent treatment day relative to baseline across dose conditions (RQ1). The reference level for these models was the untreated condition at baseline (i.e., *times treated* = 0). Observations from the post-treatment phase were not included in these analyses. Figure 4 shows Analysis 1 coefficient estimates with 95% confidence intervals for item-level factors (*word frequency*, *word length in phonemes*) and interaction terms (*times treated* \times *dose*). A significant interaction coefficient indicates the magnitude of change in naming accuracy from baseline for treated items was significantly greater than the magnitude of change in the untreated condition over the same period. A summary of Analysis 1 models is presented in supplemental Table I.

Figure 4 Forest plots of coefficient estimates for item-level factors and interaction terms from Analysis 1

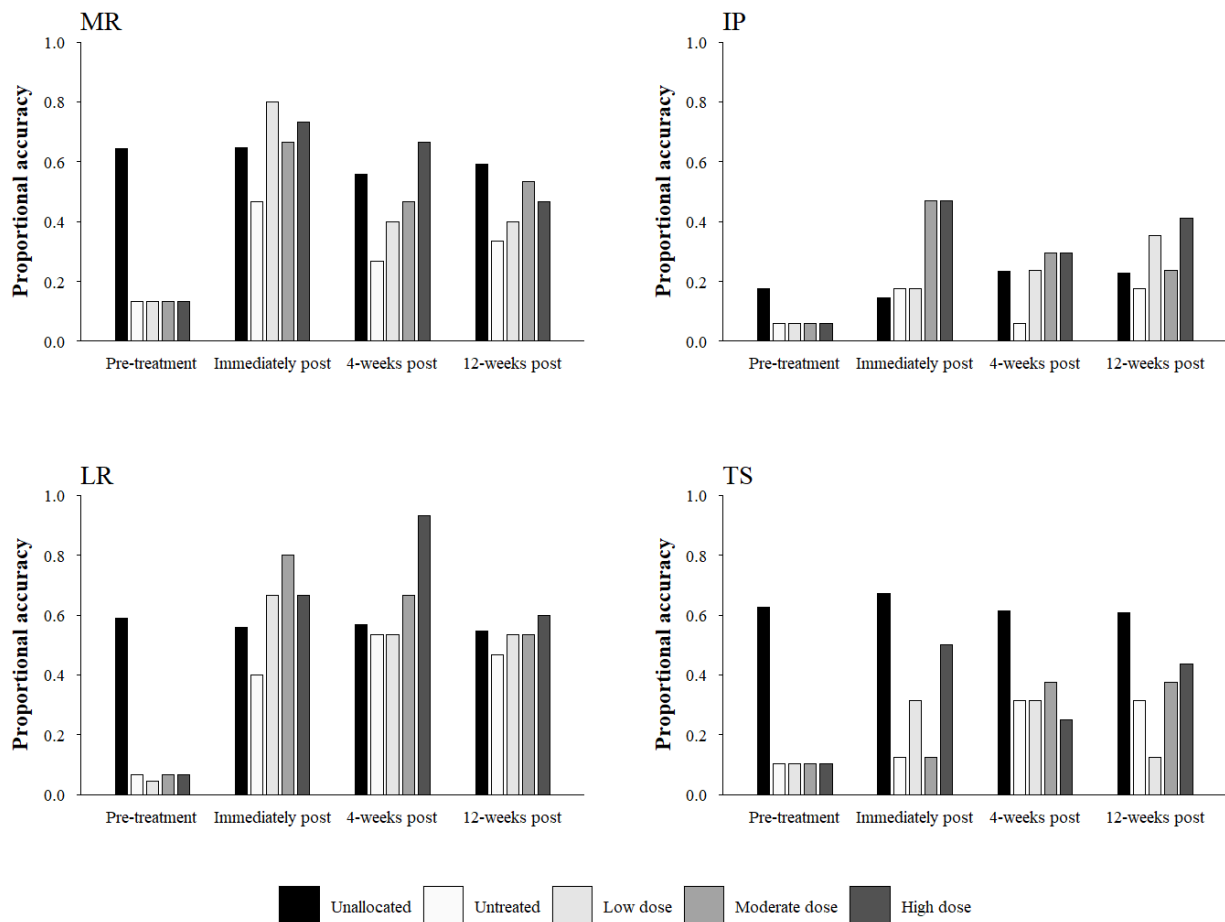


For MR, there was a main effect of *word frequency* (OR 5.91, 95% CI [1.17, 29.78], $p=.031$), with more accurate naming for higher-frequency words, but not *word length* ($p=.690$). The interaction terms indicate peak performance for the low dose condition after four treatment sessions (OR 24.38, 95% CI [1.61, 368.35], $p=.021$), the moderate dose condition after four sessions (OR 91.37, 95% CI [3.35, 2489.54], $p=.007$), and the high dose condition after two sessions (OR 13.41, 95% CI [1.12, 159.98], $p=.040$). These findings are consistent with visual inspection of the multiple-baselines chart (Figure 3). IP demonstrated statistically significant improvement in naming accuracy relative to the untreated condition after four treatment sessions for both the low dose (OR 10.67, 95% CI [1.07, 106.62], $p=.044$) and high dose (OR 10.77, 95% CI [1.08, 107.75], $p=.043$) conditions, consistent with visual inspection of probe data. LR demonstrated a main effect of *word length* (OR 0.76, 95% CI [0.60, 0.97], $p=.025$) indicating longer words were associated with less accurate naming. There were many significant interactions with greatest magnitude of change in naming for the low dose condition after two treatment sessions (OR 8.17, 95% CI [1.46, 45.63], $p=.017$), the moderate dose condition after two sessions (OR 11.80, 95% CI [1.60, 87.26], $p=.016$), and the high dose condition after four sessions (OR 12.55, 95% CI [1.78, 88.51], $p=.011$). These findings are consistent with visual inspection of the multiple-baselines case chart (Figure 3). TS's results indicate that improved naming accuracy of treated items did not exceed improvement in untreated item naming to a significant degree for any dose condition after any number of treatment sessions.

Maintenance of treatment effects (Analysis 2)

Figure 5 shows proportional naming accuracy on the 298-item picture naming test at four time points (pre-treatment, immediately post-treatment, and 4- and 12-week follow-up) for each participant.

Figure 5 Proportional naming accuracy on the picture naming test

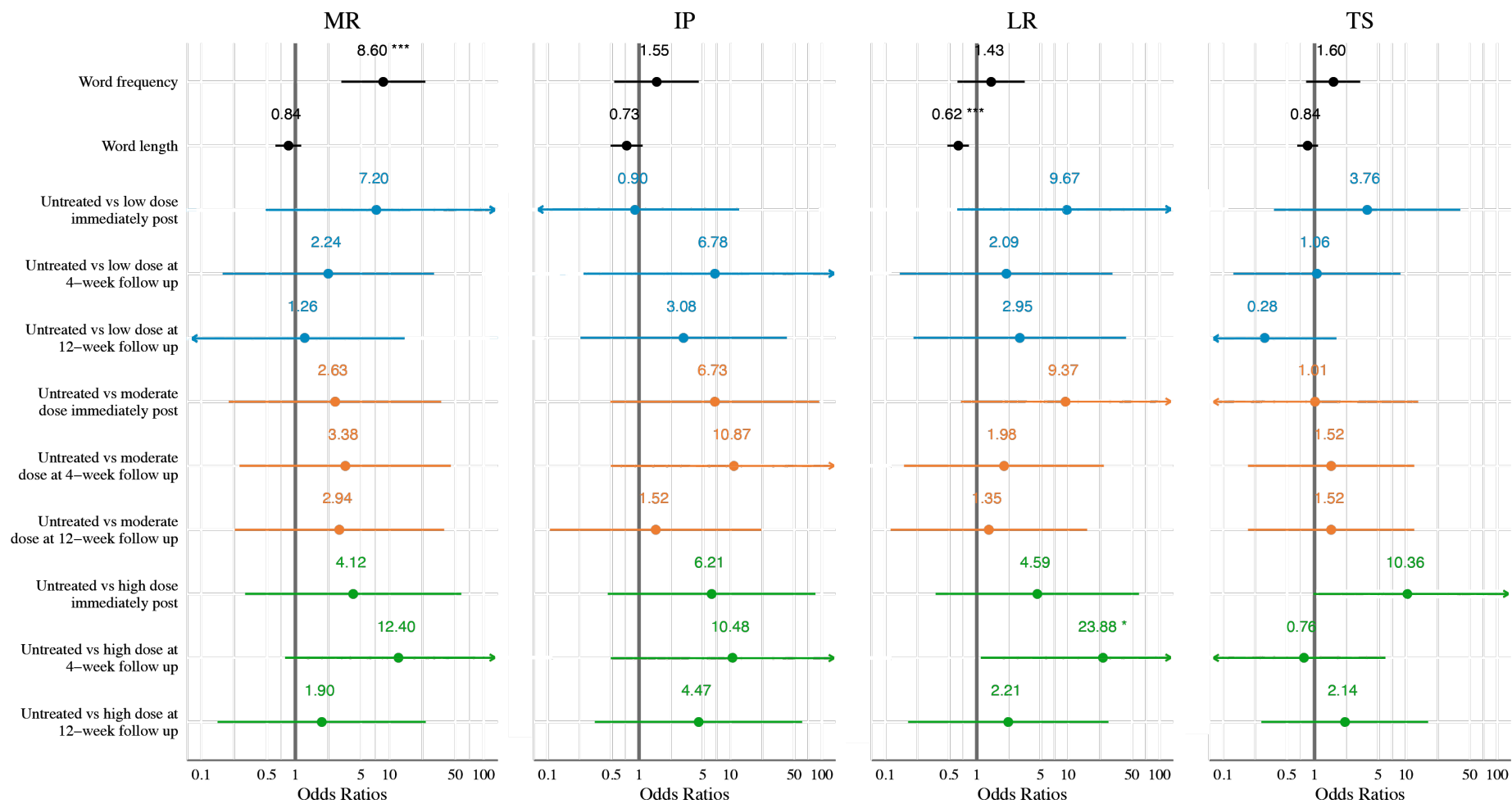


Analysis 2 used responses for allocated items (untreated, low dose, moderate dose, and high dose conditions) on the picture naming test to model acquisition (RQ1) and maintenance (RQ2) of treatment effects across dose conditions. The reference level for the interaction comparisons was the untreated condition at the pre-treatment time point. Figure 6 shows the coefficient estimates with 95% confidence intervals for item-level factors (*word frequency*, *word length in phonemes*) and interaction terms (*dose x time*) from these models. A significant result for the low, moderate, or high dose conditions indicated naming accuracy improved from pre-treatment and the magnitude of this improvement was significantly greater than any change in naming accuracy of untreated items over the same period, after accounting for word-level lexical effects. A summary of the outputs of this model can be found in supplemental Table II.

For participant MR, there was a main effect of *word frequency* (OR 8.6, 95% CI [3.08, 24.01], $p < .001$), but not *word length* ($p = .283$). As expected, there was no main effect of *dose* because each set was matched for accuracy based on pre-treatment performance on this test. There was a main effect of *time* immediately post-treatment (OR 21.73, 95% CI [3.12, 151.44], $p = .002$) and at 12-week follow up (OR 7.89, 95% CI [1.16, 53.53], $p = .035$) demonstrating inconsistent naming performance for *untreated items* over time. There was no significant interaction between *dose* and *time* (Figure 6). For IP, there were no main effects or interactions despite increased naming accuracy on the picture naming test for treated relative to untreated items (Figure 5). LR again demonstrated a main effect of *word length* (OR 0.62, 95% CI [1.78, 88.51], $p = .001$)

indicating that untreated items with longer phoneme length were less likely to be named accurately during pre-treatment assessment. There was a main effect of *time* immediately post-treatment (OR 15.83, 95% CI [2.59, 96.85], $p=.003$), at 4-week follow up (OR 31.7, 95% CI [5.07, 198.22], $p<.001$), and at 12-week follow up (OR 22.48, 95% CI [3.66, 138.17], $p=.001$) indicating improved naming performance for untreated items over time. As shown in Figure 6, the *dose x time* interaction was significant for the high dose condition at 4-week follow up (OR 23.88, 95% CI [1.03, 555.57], $p=.048$). TS demonstrated a main effect of *time* indicating that naming accuracy of untreated items improved to a significant degree by 4-weeks and 12-weeks following cessation of treatment relative to pre-treatment. Despite increased naming accuracy for treated relative to untreated items at multiple time points, these differences were estimated not to be statistically significant for any dose condition at any time point.

Figure 6 Forest plots of coefficient estimates for item-level factors and interaction terms from Analysis 2



Discussion

We tested three primary hypotheses relating to the acquisition (RQ1) and maintenance (RQ2) of picture naming skills following personalised doses of cued picture naming treatment. We found that picture naming accuracy improved with treatment for some participants (H1 partially accepted), but at different rates and to different degrees across dose conditions (H2 accepted). There was no maintenance of treatment effects (H3 rejected). Regarding the relative superiority of higher or lower doses of treatment, we anticipated learning effects to accrue throughout treatment (H4 accepted) and that items treated under the high dose condition might demonstrate relatively greater gains in picture naming accuracy (H5 partially accepted). We found that each participant responded differently to treatment. Of the four participants, LR demonstrated greatest learning effects during treatment (Analysis 1) and was the only participant to demonstrate maintenance of treatment versus no-treatment gains four weeks following treatment (Analysis 2). Both these patterns were clearest in the high-dose condition, consistent with H5. MR and IP demonstrated significant learning effects during treatment but no significant improvement or maintenance following treatment. TS did not demonstrate learning effects. We will now examine factors that may help understand this treatment response variability.

Exploration of factors associated with treatment response

The role of dose and schedule

The lack of significantly improved picture naming following treatment cessation is at odds with findings from studies by Harnish et al. (2013) and Kendall et al. (2014) in which most participants demonstrated maintenance of cued picture naming treatment effects for at least two months. As with our study, participants in the Harnish and Kendall studies represent a small but typically heterogeneous sample in terms of demographic, stroke, language, and cognition characteristics. While participants in our study were prescribed lots of opportunity to practice the targeted behaviour in each session, in traditional terms the dose was low: 3.75 hours per treatment condition (daily 45-minute sessions for five days totalling 11.25 hours across treatment conditions). Indeed, participants in the Harnish study received approximately eight hours of treatment on a single stimulus set of 50 items (approximately one hour per day, four days per week for two weeks) and participants in the Kendall study received 20 hours of treatment (one hour per day, three days per week for six to seven weeks; 45 trained items). Thus, the Harnish study provided more hours of treatment, and the Kendall study provided more hours delivered over a distributed schedule. A recent network meta-analysis of individual participant data (n=959) from 25 RCTs investigated associations between dose (i.e., total hours of treatment), schedule (i.e., hours and days of treatment per week), and language outcomes for people with aphasia (Brady et al., 2021). Optimal picture naming outcomes (as measured on the Boston Naming Test; Goodglass et al., 1983) were obtained in studies that delivered treatment for one to four days per week, for two to four hours per week, for more than 10 weeks, and for a total of up to five hours *or* between 20 and 50 hours of treatment. The current study provided 11.25 total hours of treatment over three weeks of daily treatment. Given the RELEASE findings, it is possible that the current study was either outside the range of effective doses or too intensive to engender longer-term treatment gains. Taken together with the findings from the studies by Harnish and colleagues (2013) and Off and colleagues (2016), there is

yet no strong evidence for the superiority of saturated practice over other less intensive session-level practice schedules in picture naming treatment.

Self-monitoring and integrity of the language processing network

There is growing evidence supporting the contribution of cognitive skills to the potential for recovery from post-stroke aphasia (Brownsett et al., 2014; Dutta et al., 2022; Fillingham et al., 2005; Geranmayeh et al., 2014; Geranmayeh et al., 2017; Gilmore et al., 2019; Lambon Ralph et al., 2010). In a series of studies examining response to picture naming treatment in aphasia, Fillingham and colleagues (2005) demonstrated that executive functions – specifically, problem solving, self-monitoring, recognition memory, and attention – were essential in understanding treatment response. The participants in the current study had disparate cognitive profiles (see Table 2 above). However, a possible connection between treatment outcomes and self-monitoring behaviours did emerge. Self-monitoring is important for error detection and error repair (Fillingham et al., 2005; Schwartz et al., 2016). Reduced self-monitoring, as measured by error awareness or repair, has been associated with compromised *production*, as described by Dell's interactive two-step model of naming (Foygel & Dell, 2000; see Schwartz et al., 2016 for a detailed discussion). In this model, semantic errors arise due to heightened conflict between the target and conceptually related lexical nodes and phonological errors arise due to heightened conflict among phonemes. Conflict at these two levels (lexical and phonological) is heightened when the connections between levels are weakened; the strength of these connections is estimated by the s- and p-weights, respectively, with lower weights indicating weaker connection and therefore heightened conflict. In a study of naming error detection in aphasia, Nozari and colleagues (2011) showed detection of semantic errors correlated significantly with the strength of the s-weights ($r = .59$, $p = .001$) and the detection of phonological errors with the strength of the p-weights ($r = .43$, $p = .02$).

In the current study, we noted anecdotal evidence of self-monitoring via observations of error awareness and attempts to self-correct responses during assessment, treatment, and conversational exchanges. During assessment and treatment, MR and LR were aware of errors to the point of frustration. Conversely, TS's error awareness fluctuated with fatigue whereas IP was consistently unaware of errors and was rarely observed to attempt self-correction. Compared to the other participants, MR and LR had relatively high s- and p-weights and demonstrated better learning effects (Analysis 1, Figure 3). Of note, poorer error awareness (key for self-monitoring) is commonly reported for individuals with diagnostic profiles consistent with Wernicke's aphasia, whereas individuals with Broca's and conduction-type profiles are commonly reported to have better error awareness. This was true for the participants in this study. In summary, self-monitoring behaviours might give an indication of the underlying strength of lexical representations which influences word learning (Schwartz et al., 2016). Future studies of cued picture naming treatment could explore associations between aphasia type and error awareness/self-monitoring, given the potential importance of self-monitoring to cued picture naming treatment response. Critically, the implementation of cued picture naming treatment in this study did not include feedback or shaping of production, so it is reasonable to assume that relatively good self-monitoring is a requirement for responsivity to this treatment. Furthermore, self-monitoring and underlying lexical-semantic processing may have contributed to specific aspects of the treatment responses observed in this study, particularly the propensity to benefit from repetition priming.

Repetition priming

Some people with aphasia benefit from exposure to naming opportunities in the absence of treatment (Creet et al., 2019; Nickels, 2002; Sage et al., 2011). In a lexical-retrieval treatment where generalisation to untreated items is uncommon such as cued picture naming treatment (Kendall et al., 2014), improved naming accuracy of untreated items may indicate a priming effect. Howard and colleagues (2015) suggest that to determine whether repeated probes affect behaviour, it is necessary to compare performance on treated items against untreated probe items *and* a second untreated set that is only probed before and after the treatment phase. In this study, of the 298 items named pre-treatment, some were allocated to stimulus sets which were repeatedly probed during the study (21 probes across four weeks) and the remaining unallocated pictures were tested prior to the treatment phase and following treatment. Comparison of naming performance across these different groupings of picture items provides evidence of priming effects. Three participants (MR, LR, TS) demonstrated increased naming accuracy of *untreated items* following the treatment period (Figure 5; supplemental Table II, *Time* coefficient estimates), but none of the participants showed improved naming of unallocated items on pre/post naming test. Improvement in naming of probed items but not unallocated items suggests the influence of a priming effect rather than generalisation of treatment gains to untreated items.

Priming appears to benefit people with intact self-monitoring and less severe anomia and hinder people with reduced self-monitoring and more severe anomia (Creet et al., 2019). As described above, MR and LR consistently demonstrated awareness of errors and successful naming and TS demonstrated awareness of his naming performance unless he was fatigued. Of these three, LR appeared to benefit most from the priming effect; he named untreated items significantly more accurately at each time point following the treatment period. Interestingly, this priming effect may have enhanced the overall effect of treatment; LR's naming accuracy for treated items improved and peaked quickly during treatment (Figure 4) and he was the only participant to maintain any treatment gains in the follow up period (Figure 6). MR and TS also exhibited improvement in untreated items, but to a lesser extent. Conversely, IP did not demonstrate awareness of errors or a priming effect; naming accuracy of untreated items did not improve significantly at any time point following the treatment period. The characteristics of these participants are consistent with those expected to benefit (or not) from repeated exposure to picture stimuli (Creet et al., 2019). Furthermore, the cognitive and linguistic characteristics (i.e., intact self-monitoring and relatively spared lexical-semantic processing) that may predispose people to priming effects may also make them good candidates for this type of repetition-based naming therapy.

Other factors

A number of key predictors of response to aphasia therapy have been identified (e.g., Watila & Balarabe, 2015) including neuroanatomical features such as lesion site and size (e.g., Kristinsson et al., 2022; Plowman et al., 2012; Price et al., 2010) and cognitive status (e.g., Lambon Ralph et al., 2010; Simic et al., 2020). Furthermore, there is emerging evidence of distinct recovery trajectories for people with bilingual or multilingual aphasia which may be linked to differential executive functioning in this population (e.g., Radman et al., 2016). Examining the role of these predictors was beyond the scope of this study however it is possible that these factors may have influenced participants' response to treatment in this study.

Limitations

Method of determining treatment effect

Comparison of treatment effects across studies requires homogeneity of effect size measures; however, there is no gold standard effect size measure in aphasiology. We implemented mixed effects modelling whereas previous studies of cued picture naming treatment used Standardised Mean Difference (SMD; Kendall et al., 2014) or the non-parametric Conservative Dual Criterion (CDC) method (Harnish et al., 2013). The characterisation of what constitutes a significant treatment effect will be different across these studies. Mixed effects modelling approaches have several advantages over SMD and CDC when evaluating time series data from a multiple-baselines study. For example, mixed effects models use all observations at every time point to estimate variance rather than averaging across phases which under-represents true variability in the data (Pustejovsky et al., 2014). Importantly, mixed effects models can adjust for autocorrelation (Wiley & Rapp, 2019), and can account for known and unknown factors that may contribute to changes in the dependent variable over time (such as item-, person-, and treatment-related covariates) and the relative magnitude of these moderating effects (Wiley & Rapp, 2019). In addition to treatment-related variables, our models accounted for changes in naming accuracy attributable to lexical properties of individual items. For two participants, these variables demonstrated statistically significant effects (MR, word frequency; LR, word length). This highlights another strength of this modelling approach over other statistical approaches (such as simple effect-size measures, that aggregate across items) which do not account for lexical properties when estimating treatment effects. Notably, including these lexical factors may have led to a different picture of the effects of cued picture naming treatment from that seen in previous studies. Lastly, a significant limitation of SMD and CDC is that these methods do not use statistical inferencing to account for practice effects for treated items, particularly over and above practice effects that may be observed for untreated items. This limitation may lead to over-estimation of treatment effect size (Nickels et al., 2015). Future work might aim to synthesise findings by meta-analysing raw naming probe data from these studies.

Understanding patterns of naming accuracy throughout treatment

Participants who demonstrated learning effects (MR, IP, LR) exhibited peaks in naming accuracy that were not always sustained. For example, participants demonstrated rising trends in accuracy which dropped off after five treatment sessions: the low dose condition (MR), the moderate dose condition (MR, IP), and the high dose condition (LR). One factor that may have contributed to this response profile relates to the study structure. The study ran for four weeks, Monday to Friday. Naming probes were conducted at the beginning of each session. The first six probes were baseline (Monday – Monday). Treatment phase probes commenced Tuesday following the first treatment session (*times treated* = 1) and finished the following Monday (*times treated* = 5). Therefore, the fifth treatment phase probe occurred after the weekend. This two-day lag may have contributed to decay in naming performance seen after five treatment sessions (Figure 4). This pilot study was under-powered to examine other factors such as cognitive functions that may be important in understanding why performance peaked and then decayed earlier in the treatment phase for some participants/dose conditions (e.g., MR high dose, LR low and moderate dose).

Another limitation of the multiple-baselines design is that with sequential treatment phases, the period between treatment cessation and the post-treatment assessment timepoint is different

across dose conditions. A possible alternative approach to estimating pre-post treatment effects would be to use probe data from the end of the treatment phase as the ‘post treatment’ measure (e.g., Wambaugh et al., 2017). However, we did not do this in the current study because the treatment phase included only five probes and pre-post acquisition was a secondary analysis. It is acknowledged that this may have affected estimates of pre-post acquisition (Analysis 2) due to decay of naming performance for earlier treated items.

Sample size

As is common for exploratory pilot studies in aphasia research, this study with four participants was under-powered. Furthermore, the complexity of mixed-effects models is constrained by the sample size. There is a distinction between sample sizes at different levels of mixed-effects models (Maas & Hox, 2005; Wiley & Rapp, 2019). The first-level sample size refers to the number of items and the second-level sample size is the number of participants. The minimum number of observations required for a given model structure is the first-level sample size multiplied by the number of random-effects (Wiley & Rapp, 2019). Literature suggests that an adequate second-level sample size is as few as five to estimate beta coefficients of the fixed effects and that random-effects estimates with small participant n will be susceptible to inflated Type 1 error, that is, reporting an effect as significant when it is not (Wiley & Rapp, 2019). Therefore, we elected not to examine group-level effects by analysing aggregated participant data.

Generalisation

This Phase I pilot study examining dose-response relationships in cued picture naming treatment did not examine generalisation of treatment effects. As mentioned previously, this treatment has limited evidence for stimulus and response generalisation (Harnish et al., 2013; Kendall et al., 2014). A number of mechanisms and factors that favour generalisation in picture naming treatments have been proposed including typicality of stimuli (e.g., Kiran, 2008), semantic relatedness of stimuli (e.g., Quique et al., 2019), and lexical-semantic processing (e.g., Best et al., 2013). For example, using a cued naming treatment, Best et al. (2013) demonstrated people with relatively intact lexical-semantic processing and relatively impaired phonological processing demonstrated generalisation to untreated items. A limitation of the current study is that we did not control for factors that might promote generalisation such as typicality or by treating items from the same semantic category. While it is possible that response generalisation may have occurred in the present study, another equally plausible explanation of improved naming on untreated items is *regression to the mean*. In this study, items were allocated to the four treatment conditions on the basis of having been named inaccurately prior to treatment. Regression to the mean – a ubiquitous statistical phenomenon in repeated measures and time series data (Barnett et al., 2005) – dictates that there will be some drift towards improved naming accuracy due to random measurement error. Future work should aim to explore relationships between treatment dose and generalisation, particularly for treatment approaches or individuals whose characteristics favour generalisation.

Clinical implications and future directions

In the context of the current and previous evidence (Harnish et al., 2013; Kendall et al., 2014), good candidates for cued picture naming therapy may have intact self-monitoring and

relatively mild lexical-semantic and phonological processing deficits. Current evidence suggests a person with this profile might be expected to make rapid gains in picture naming within a small number of treatment sessions and that significant changes in naming accuracy accrue with time. Improvements in naming accuracy are likely to be stimulus specific and may not be maintained in the longer-term.

Further research is required to test whether these person-level factors can predict a person's outcome after cued picture naming treatment delivered in a saturated practice schedule. The current study provided a relatively low number of hours of treatment outside the range recently identified to be associated with optimal naming treatment outcomes (Brady et al., 2021). Given the modest effects observed in the current study, further research might use the same methods to investigate dose effects after longer treatment duration (i.e., 20-50 hours of cued picture naming treatment). For this investment of time, people with aphasia, clinicians, and researchers might reasonably expect treatment gains to generalise to novel contexts with increasing cognitive, linguistic, and social demands. Ultimately, investigation of generalisation to dialogue will be needed. Further research is also required to determine if adjusting dose and schedule parameters could enhance long-term maintenance of treatment gains (e.g., a period of intensive, saturated practice followed by a period of distributed, random practice) and to investigate the potential benefits of self-managed practice. Computer-delivered treatment programs that are easily customisable, such as the software developed for this pilot study, may be especially well suited to self-management strategies (Nichol et al., 2022). Research could also investigate the effect of coupling cued picture naming treatment with treatment approaches that aim to improve sentence- and discourse-level outcomes to promote generalisation of treatment effects into situated language use. Finally, we developed and tested a novel method to investigate the effects of personalising dose in cued picture naming treatment. Future research might explore the applicability of this method to the examination of dose-response relationships in other aphasia treatments.

Conclusion

In this pilot study we implemented a novel method to examine the effect of personalising the dose of a cued picture naming treatment provided to four people with chronic post-stroke anomia. Using a novel analytical technique, we found modest treatment effects and variable dose-response relationships amongst the participants. We explored treatment-related factors, and cognitive and linguistic factors such as self-monitoring abilities and underlying lexical-semantic and phonological processing that may have influenced treatment response in these participants. Treatment-related factors such as treatment type, timing, and dose are predictors of aphasia recovery which, unlike biographic or stroke-related predictors, are modifiable. Greater understanding of dose-response relationships and treatment personalisation may lead to enhanced rehabilitation outcomes for people recovering from post-stroke aphasia.

References

- Baker, E. (2012). Optimal intervention intensity. *International Journal of Speech-Language Pathology*, 14(5), 401-409. <https://doi.org/10.3109/17549507.2012.700323>
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215-220. <https://doi.org/10.1093/ije/dyh299>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv:1506.04967*. <https://doi.org/10.48550/arXiv.1506.04967>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., & Grothendieck, G. (2015). Package ‘lme4’. *Convergence*, 12(1).
- Best, W., Greenwood, A., Grassly, J., Herbert, R., Hickin, J., & Howard, D. (2013). Aphasia rehabilitation: Does generalisation from anomia therapy occur and is it predictable? A case series study. *Cortex*, 49(9), 2345-2357. <https://doi.org/10.1016/j.cortex.2013.01.005>
- Brady, M. C., Ali, M., VandenBerg, K., Williams, L. J., Williams, L. R., Abo, M., Becker, F., Bowen, A., Brandenburg, C., & Breitenstein, C. (2021). Dosage, intensity, and frequency of language therapy for aphasia: A systematic review–based, individual participant data network meta-analysis. *Stroke*. <https://doi.org/10.1161/STROKEAHA.121.035216>
- Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*(6). <https://doi.org/10.1002/14651858.CD000425.pub4>
- Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS One*, 5(5), e10773. <https://doi.org/10.1371/journal.pone.0010773>
- Brownsett, S. L. E., Warren, J. E., Geranmayeh, F., Woodhead, Z., Leech, R., & Wise, R. J. S. (2014). Cognitive control and its impact on recovery from aphasic stroke. *Brain*, 137(1), 242-254. <https://doi.org/10.1093/brain/awt289>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990. <https://doi.org/10.3758/BRM.41.4.977>
- Creet, E., Morris, J., Howard, D., & Nickels, L. (2019). Name it again! investigating the effects of repeated naming attempts in aphasia. *Aphasiology*, 33(10), 1202-1226. <https://doi.org/10.1080/02687038.2019.1622352>
- Crosson, B., Rodriguez, A. D., Copland, D., Fridriksson, J., Krishnamurthy, L. C., Meinzer, M., Raymer, A. M., Krishnamurthy, V., & Leff, A. P. (2019). Neuroplasticity and aphasia treatments: New approaches for an old problem. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(10), 1147-1155. <https://doi.org/10.1136/jnnp-2018-319649>
- DeDe, G., Ricca, M., Knilans, J., & Trubl, B. (2014). Construct validity and reliability of working memory tasks for people with aphasia. *Aphasiology*, 28(6), 692-712. <https://doi.org/10.1080/02687038.2014.895973>

- Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., & Kiran, S. (2020). Videoconference administration of the Western Aphasia Battery–Revised: Feasibility and validity. *American Journal of Speech-Language Pathology*, 29(2), 673-687.
https://doi.org/https://doi.org/10.1044/2019_AJSLP-19-00023
- Dignam, J. K., Rodriguez, A. D., & Copland, D. A. (2016). Evidence for intensive aphasia therapy: Consideration of theories from neuroscience and cognitive psychology. *PM&R*, 8(3), 254-267. <https://doi.org/https://doi.org/10.1016/j.pmrj.2015.06.010>
- Doogan, C., Dignam, J., Copland, D., & Leff, A. (2018). Aphasia recovery: When, how and who to treat? *Current Neurology and Neuroscience Reports*, 18(12), 90.
<https://doi.org/10.1007/s11910-018-0891-x>
- Dutta, M., Murray, L. L., & Stark, B. C. (2022). Assessing the integrity of executive functioning in chronic aphasia. *Aphasiology*, 1-38.
<https://doi.org/https://doi.org/10.1080/02687038.2022.2049675>
- Fillingham, J., Sage, K., & Lambon Ralph, M. (2005). Further explorations and an overview of errorless and errorful therapy for aphasic word-finding difficulties: The number of naming attempts during therapy affects outcome. *Aphasiology*, 19(7), 597-614.
<https://doi.org/https://doi.org/10.1080/02687030544000272>
- Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, 36(3), 387-406. <https://doi.org/https://doi.org/10.1901/jaba.2003.36-387>
- 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/https://doi.org/10.1006/jmla.2000.2716>
- Geranmayeh, F., Brownsett, S. L. E., & Wise, R. J. S. (2014). Task-induced brain activity in aphasic stroke patients: What is driving recovery? *Brain*, 137(10), 2632-2648.
<https://doi.org/https://doi.org/10.1093/brain/awu163>
- Geranmayeh, F., Chau, T. W., Wise, R. J. S., Leech, R., & Hampshire, A. (2017). Domain-general subregions of the medial prefrontal cortex contribute to recovery of language after stroke. *Brain*, 140(7), 1947-1958. <https://doi.org/https://doi.org/10.1093/brain/awx134>
- Goodglass, H., Kaplan, E., & Weintraub, S. (1983). *Boston naming test*. Philadelphia, PA: Lea & Febiger.
- Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2019). Nonlinguistic cognitive factors predict treatment-induced recovery in chronic poststroke aphasia. *Archives of Physical Medicine and Rehabilitation*, 100(7), 1251-1258.
<https://doi.org/https://doi.org/10.1016/j.apmr.2018.12.024>
- Harnish, S. M., Morgan, J., Lundine, J. P., Bauer, A., Singletary, F., Benjamin, M. L., Gonzalez Rothi, L. J., & Crosson, B. (2013). Dosing of a cued picture-naming treatment for anomia. *American Journal of Speech-Language Pathology*, 23(2), S285-299.
https://doi.org/https://doi.org/10.1044/2014_AJSLP-13-0081
- Harvey, S. (2022). Series of operations used to allocate pictures to stimulus sets. In https://s4harvey.github.io/other_links/Figure_Series_of_operations_picture_stimuli_dose_comparison.pdf.
- Harvey, S., Carragher, M., Dickey, M. W., Pierce, J. E., & Rose, M. L. (2022). Dose effects in behavioural treatment of post-stroke aphasia: A systematic review and meta-analysis.

- Disability and Rehabilitation*, 44(12), 2548-2559.
<https://doi.org/https://doi.org/10.1080/09638288.2020.1843079>
- Harvey, S. R., Carragher, M., Dickey, M. W., Pierce, J. E., & Rose, M. L. (2021). Treatment dose in post-stroke aphasia: A systematic scoping review. *Neuropsychological Rehabilitation*, 31(10), 1629-1660. <https://doi.org/10.1080/09602011.2020.1786412>
- Hayward, K. S., Churilov, L., Dalton, E. J., Brodtmann, A., Campbell, B. C. V., Copland, D., Dancause, N., Godecke, E., Hoffmann, T. C., & Lannin, N. A. (2021). Advancing stroke recovery through improved articulation of nonpharmacological intervention dose. *Stroke*, 52(2), 761-769. <https://doi.org/https://doi.org/10.1161/STROKEAHA.120.032496>
- Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology*, 29(5), 526-562.
<https://doi.org/https://doi.org/10.1080/02687038.2014.985884>
- Johnson, C. O., Nguyen, M., Roth, G. A., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H. N., & Abu-Rmeileh, N. M. E. (2019). Global, regional, and national burden of stroke, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 18(5), 439-458.
[https://doi.org/https://doi.org/10.1016/S1474-4422\(19\)30034-1](https://doi.org/https://doi.org/10.1016/S1474-4422(19)30034-1)
- Kendall, D., Raymer, A., Rose, M., Gilbert, J., & Gonzalez Rothi, L. J. (2014). Anomia treatment platform as behavioral engine for use in research on physiological adjuvants to neurorehabilitation. *Journal of Rehabilitation Research and Development*, 51(3).
<https://doi.org/https://doi.org/10.1682/JRRD.2013.08.0172>
- Kertesz, A. (2007). *Western Aphasia Battery: Revised*. Pearson.
<https://doi.org/https://doi.org/10.1037/t15168-000>
- Kessels, R. P. C., van Zandvoort, M. J. E., Postma, A., Kappelle, L. J., & de Haan, E. H. F. (2000). The Corsi Block-Tapping Task: Standardization and normative data. *Applied Neuropsychology*, 7(4), 252-258. https://doi.org/10.1207/S15324826AN0704_8
- Kiran, S. (2008). Typicality treatment for naming deficits in aphasia: Why does it work? *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, 18(1), 6-14. <https://doi.org/https://doi.org/10.1044/nnsld18.1.6>
- Kiran, S., & Thompson, C. K. (2019). Neuroplasticity of language networks in aphasia: Advances, updates and future challenges. *Frontiers in Neurology*, 10, 295.
<https://doi.org/https://doi.org/10.3389/fneur.2019.00295>
- Kleim, J. A., Barbay, S., Cooper, N. R., Hogg, T. M., Reidel, C. N., Rempel, M. S., & Nudo, R. J. (2002). Motor learning-dependent synaptogenesis is localized to functionally reorganized motor cortex. *Neurobiology of Learning and Memory*, 77(1), 63-77.
<https://doi.org/https://doi.org/10.1006/nlme.2000.4004>
- Kristinsson, S., den Ouden, D. B., Rorden, C., Newman-Norlund, R., Neils-Strunjas, J., & Fridriksson, J. (2022). Predictors of therapy response in chronic aphasia: Building a foundation for personalized aphasia therapy. *Journal of Stroke*, 24(2), 189-206.
<https://doi.org/https://doi.org/10.5853/jos.2022.01102>
- Lambon Ralph, M. A., Snell, C., Fillingham, J. K., Conroy, P., & Sage, K. (2010). Predicting the outcome of anomia therapy for people with aphasia post CVA: Both language and cognitive status are key predictors. *Neuropsychological Rehabilitation*, 20(2), 289-305.
<https://doi.org/https://doi.org/10.1080/09602010903237875>

- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.
<https://doi.org/https://doi.org/10.2307/2529786>
- Luke, L. M., Allred, R. P., & Jones, T. A. (2004). Unilateral ischemic sensorimotor cortical damage induces contralesional synaptogenesis and enhances skilled reaching with the ipsilateral forelimb in adult male rats. *Synapse*, 54(4), 187-199.
<https://doi.org/https://doi.org/10.1002/syn.20080>
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86-92. <https://doi.org/https://doi.org/10.1027/1614-2241.1.3.86>
- Menahemi-Falkov, M., Breitenstein, C., Pierce, J. E., Hill, A. J., O'Halloran, R., & Rose, M. L. (2021). A systematic review of maintenance following intensive therapy programs in chronic post-stroke aphasia: importance of individual response analysis. *Disability and rehabilitation*, 1-16. <https://doi.org/https://doi.org/10.1080/09638288.2021.1955303>
- Nichol, L., Rodriguez, A. D., Pitt, R., Wallace, S. J., & Hill, A. J. (2022). "Self-management has to be the way of the future": Exploring the perspectives of speech-language pathologists who work with people with aphasia. *International Journal of Speech-Language Pathology*, 1-15. <https://doi.org/https://doi.org/10.1080/17549507.2022.2055144>
- Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10-11), 935-979. <https://doi.org/https://doi.org/10.1080/02687030244000563>
- Nickels, L., Best, W., & Howard, D. (2015). Optimising the ingredients for evaluation of the effects of intervention. *Aphasiology*, 29(5), 619-643.
<https://doi.org/https://doi.org/10.1080/02687038.2014.1000613>
- Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive psychology*, 63(1), 1-33. <https://doi.org/https://doi.org/10.1016/j.cogpsych.2011.05.001>
- Off, C. A., Griffin, J. R., Spencer, K. A., & Rogers, M. A. (2016). The impact of dose on naming accuracy with persons with aphasia. *Aphasiology*, 30(9), 983-1011.
<https://doi.org/https://doi.org/10.1080/02687038.2015.1100705>
- Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, 35(4), 303-322.
<https://doi.org/https://doi.org/10.1177/0145445511399147>
- Plowman, E., Hentz, B., & Ellis, C. (2012). Post-stroke aphasia prognosis: A review of patient-related and stroke-related factors. *Journal of evaluation in clinical practice*, 18(3), 689-694. <https://doi.org/https://doi.org/10.1111/j.1365-2753.2011.01650.x>
- Price, C. J., Seghier, M. L., & Leff, A. P. (2010). Predicting language outcome and recovery after stroke: the PLORAS system. *Nature Reviews Neurology*, 6(4), 202-210.
<https://doi.org/https://doi.org/10.1038/nrneurol.2010.15>
- Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, 39(5), 368-393.
<https://doi.org/https://doi.org/10.3102/1076998614547577>
- Quique, Y. M., Evans, W. S., & Dickey, M. W. (2019). Acquisition and generalization responses in aphasia naming treatment: A meta-analysis of semantic feature analysis outcomes. *American*

- Journal of Speech-Language Pathology*, 28(1), 230-246.
https://doi.org/https://doi.org/10.1044/2018_AJSLP-17-0155
- R Core Team. (2013). R: A language and environment for statistical computing.
- Radman, N., Mouthon, M., Di Pietro, M., Gaytanidis, C., Leemann, B., Abutalebi, J., & Annoni, J.-M. (2016). The role of the cognitive control system in recovery from bilingual aphasia: A multiple single-case fMRI study. *Neural Plasticity*, 2016.
<https://doi.org/https://doi.org/10.1155/2016/8797086>
- Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, 24, 121-133.
<https://doi.org/https://doi.org/10.1037/t56477-000>
- Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). The Test of Everyday Attention (TEA). *Bury St. Edmunds, UK: Thames Valley Test Company*, 197-221.
- Sage, K., Snell, C., Lambon Ralph, M. A., Sage, K., Snell, C., & Lambon Ralph, M. A. (2011). How intensive does anomia therapy for people with aphasia need to be? *Neuropsychological Rehabilitation*, 21(1), 26-41. <https://doi.org/https://doi.org/10.1080/09602011.2010.528966>
- Schwartz, M. F., Middleton, E. L., Brecher, A., Gagliardi, M., & Garvey, K. (2016). Does naming accuracy improve through self-monitoring of errors? *Neuropsychologia*, 84, 272-281.
<https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2016.01.027>
- Simic, T., Bitan, T., Turner, G., Chambers, C., Goldberg, D., Leonard, C., & Rochon, E. (2020). The role of executive control in post-stroke aphasia treatment. *Neuropsychological Rehabilitation*, 30(10), 1853-1892.
<https://doi.org/https://doi.org/10.1080/09602011.2019.1611607>
- Togher, L. (2012). Challenges inherent in optimizing speech-language pathology outcomes: It's not just about counting the hours. *International Journal of Speech-Language Pathology*, 14(5), 438-442. <https://doi.org/https://doi.org/10.3109/17549507.2012.689334>
- Wambaugh, J. L., Nessler, C., Wright, S., Mauszycki, S. C., DeLong, C., Berggren, K., & Bailey, D. J. (2017). Effects of blocked and random practice schedule on outcomes of sound production treatment for acquired apraxia of speech: Results of a group investigation. *Journal of Speech, Language, and Hearing Research*, 60(6S), 1739-1751.
https://doi.org/https://doi.org/10.1044/2017_JSLHR-S-16-0249
- Watila, M. M., & Balarabe, S. A. (2015). Factors predicting post-stroke aphasia recovery. *Journal of the Neurological Sciences*, 352(1-2), 12-18.
<https://doi.org/https://doi.org/10.1016/j.jns.2015.03.020>
- Wiley, R. W., & Rapp, B. (2019). Statistical analysis in Small-N Designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, 33(1), 1-30.
<https://doi.org/10.1080/02687038.2018.1454884>

Acknowledgements

The authors would like to thank the participants for their hard work and commitment to the study and their families for supporting participation. Thank you to Shaarang Tanpure for designing and building the online therapy platform and to Simon Oakley, DM and Maya Menahemi-Falkov for assistance testing and refining the training software. Thanks to Bianca Tidey for double-rating the outcome data.

Funding

This work was supported by an Australian Government Research Training Program Scholarship and the NHMRC funded Centre for Research Excellence in Aphasia Recovery and Rehabilitation (#1153236).

Conflicts of interest

The authors declare no potential conflict of interest.

Appendix A

Detailed participant information

Participant MR was a 54-year-old man who sustained a left hemispheric ischaemic stroke ten years prior to recruitment. Prior to his stroke, MR worked in IT and spoke both English and Urdu daily. At recruitment, MR was receiving community-based support from a speech-language pathologist focussed on life participation and was actively involved in a choir for stroke survivors. On testing, MR presented on the lower range of mild aphasia (WAB AQ 66.8), reduced sustained attention, and reduced visuo-spatial working memory (see Table 1 for specific scores). Across the three pre-treatment administrations of the picture naming test, MR was on average able to name 161 of 298 (54%) items accurately. Most errors were due to no response being provided within the 12 second time limit (59% of all errors, Table 2). Semantically related responses contributed 20% of errors. Analysis of naming errors on the PNT suggested lexical retrieval deficits primarily at the level of lexical-semantic processing (s-weight .021), and secondarily at the level of phonological processing (p-weight .032). MR frequently reported feeling frustrated by difficulties naming pictures throughout assessment but remained determined to participate.

IP was a 68-year-old woman who sustained a left hemispheric ischaemic stroke 23 months prior to recruitment. Prior to her stroke, IP worked several casual jobs, most recently as a cleaner. She was receiving fortnightly impairment-focussed speech-language therapy which was paused prior to enrolment in this study. Pre-treatment assessment showed that IP presented with moderate aphasia (WAB AQ 53.2), reduced sustained attention, and reduced verbal and visuo-spatial working memory. IP was able to name, on average, 45 from 298 (15.1%) pictures accurately across the three administrations of the picture naming test. Most errors were phonological in nature; phonologically unrelated non-words (approximately 50% of errors), phonologically related real words (5%), and phonologically related non-words (3%). In contrast, picture naming errors on the PNT were primarily semantic (s-weight 0.001) and secondarily phonological (p-weight 0.018) in nature. Throughout assessment and treatment, it was observed that IP had limited awareness of speech errors. This was also evident in conversational exchanges; IP's verbal output was fluent and jargonistic with infrequent recognition of word errors or attempts to self-correct.

LR was a 72-year-old man who sustained a stroke 26 months prior to recruitment. He was born in Italy and had lived in South Africa where he ran a large mechanical engineering firm before migrating to Australia. He spoke Italian, Afrikaans, and English before his stroke. At recruitment, LR was attending a weekly online support group for people with aphasia. On testing, LR presented with moderate aphasia (WAB-R AQ 63.2), and reduced verbal and visuo-spatial working memory. LR's sustained attention was within the normal range. LR averaged 144 accurate responses (48.3%) on the pre-treatment picture naming test. Seventy percent of errors were due to no response being provided in the 12 second time frame. Semantic errors constituted about half of the remaining errors (13% of all errors). Analysis of naming errors on the PNT suggested lexical retrieval deficits primarily at the level of lexical-semantic processing (s-weight 0.021) and secondarily at the level of phonological processing (p-weight 0.032). LR reported 'tip of the tongue' word finding difficulties and was aware of, and frequently frustrated by, naming errors.

TS was a 78-year-old man who sustained a left hemispheric ischaemic stroke 27 months prior to recruitment. TS was a retired dentist who spoke Mandarin Chinese and English prior to his stroke. At recruitment, TS was receiving weekly impairment-focussed speech-language therapy

which was paused prior to enrolment in this study. Pre-treatment assessment demonstrated that TS presented with moderate aphasia (WAB AQ 64.8), and reduced verbal and visuo-spatial working memory. TS was able to name, on average, 153 from 298 (51.3%) pictures accurately across the three administrations of the picture naming test. Errors were predominantly semantically related (approximately 28% of all errors) and phonologically unrelated and related non-word responses (approximately 20% and 8%, respectively). A substantial number of responses were unrelated (17%) or perseverative (8%). Picture naming errors on the PNT revealed similarly weighted deficits in both semantic (s-weight 0.019) and phonological (p-weight 0.021) processing. TS had a residual dense right-sided hemiplegia and required assistance to reposition for comfort and frequent breaks to rest throughout assessment and treatment sessions.

Appendix B

Comprehensive description of treatment dose provided to participants reported using the multidimensional dose articulation framework (Hayward et al., 2021)

Dose dimension	MR	IP	LR	TS
Duration	3 weeks total (1 week per dose condition)			
Days	Daily treatment (15 days, 5 days per dose condition)			
Sessions	1 session per day			
Total time-on-task	Low dose: 225 min; Moderate dose: 225 min; High dose: 225 min; Total: 675 min			
Session length including breaks (mean [SD])	55 min 4 s (3 min 47 s) Low dose: 54 min 36 s (2 min 36 s) Moderate dose: 57 min 36 s (3 min 30 s) High dose: 63 min (2 min 21 s)	55 min 28 s (4 min 59 s) Low dose: 51 min 48 s (3 min 34 s) Moderate dose: 55 min (1 min 25 s) High dose: 59 min (5 min 49 s)	58 min 40 s (6 min 47 s) Low dose: 50 min 36 s (3 min 47 s) Moderate dose: 62 min 12 s (3 min 34 s) High dose: 63 min 12 s (3 min 25 s)	55 min 44 s (5 min 41 s) Low dose: 50 min 24 s (3 min 55 s) Moderate dose: 54 min 48 s (1 min 39 s) High dose: 62 min (3 min)
Session density	Active: 45 min	Active: 45 min	Active: 45 min	Active: 45 min
Proportion of total session length spent active	Inactive: Low dose: 9 min 36 s (2 min 36 s) Moderate dose: 12 min 36 s (3 min 30 s) High dose: 18 min (2 min 21 s)	Inactive: Low dose: 6 min 48 s (3 min 34 s) Moderate dose: 10 min (1 min 25 s) High dose: 14 min 36 s (5 min 49 s)	Inactive: Low dose: 5 min 36 s (3 min 47 s) Moderate dose: 17 min 12 s (3 min 34 s) High dose: 18 min 12 s (3 min 25 s)	Inactive: Low dose: 5 min 24 s (3 min 55 s) Moderate dose: 9 min 48 s (1 min 39 s) High dose: 17 min (3 min)
	Density Low dose: 0.82 Moderate dose: 0.78 High dose: 0.71	Density Low dose: 0.87 Moderate dose: 0.82 High dose: 0.76	Density Low dose: 0.89 Moderate dose: 0.72 High dose: 0.71	Density Low dose: 0.89 Moderate dose: 0.82 High dose: 0.73
Episodes per session (naming opportunities per session)	Low dose: 15 (120) Moderate dose: 30 (240) High dose: 45 (360)	Low dose: 17 (136) Moderate dose: 34 (272) High dose: 51 (408)	Low dose: 15 (120) Moderate dose: 30 (240) High dose: 45 (360)	Low dose: 16 (128) Moderate dose: 32 (256) High dose: 48 (384)
Total number of naming opportunities received/prescribed (%)	Low dose: 596/600 (99.3) Moderate dose: 1192/1200 (99.3) High dose: 1800/1800 (100) Total: 3588/3600 (99.7)	Low dose: 676/680 (99.4) Moderate dose: 1360/1360 (100) High dose: 2040/2040 (100) Total: 4076/4080 (99.9)	Low dose: 587/600 (97.8) Moderate dose: 1196/1200 (99.7) High dose: 1800/1800 (100) Total: 3583/3600 (99.5)	Low dose: 640/640 (100) Moderate dose: 1280/1280 (100) High dose: 1920/1920 (100) Total: 3840/3840 (100)
Episode length	Low dose: 60 s Moderate dose: 90 s High dose: 180 s	Low dose: 159 s Moderate dose: 79 s High dose: 53 s	Low dose: 180 s Moderate dose: 90 s High dose: 60 s	Low dose: 169 s Moderate dose: 84 s High dose: 56 s
Episode difficulty	<i>Not measured</i>	<i>Not measured</i>	<i>Not measured</i>	<i>Not measured</i>
Average self-reported rating of session difficulty (0-100, mean [SD])	Low dose: 5.8 (4.3) Moderate dose: 9.0 (3.7) High dose: 10.6 (6.5)	Low dose: 23.2 (3.9) Moderate dose: 28.2 (9.4) High dose: 33.2 (10.2)	Low dose: 29.5 (6.6) Moderate dose: 27.6 (13.4) High dose: 40.6 (11.2)	Low dose: 10.0 (7.1) Moderate dose: 43.8 (33.3) High dose: 30.4 (12.0)
Episode intensity	1 picture per episode	1 picture per episode	1 picture per episode	1 picture per episode
	Time per naming opportunity Low dose: 22.5 s Moderate dose: 11.25 s High dose: 7.5 s	Time per naming opportunity Low dose: 19.9 s Moderate dose: 9.9 s High dose: 6.6 s	Time per naming opportunity Low dose: 22.5 s Moderate dose: 11.25 s High dose: 7.5 s	Time per naming opportunity Low dose: 21.1 s Moderate dose: 10.5 s High dose: 7.0 s

NB: min = minute(s), s = second(s)

Supplemental materials

1. Participant recruitment procedure

Modified communicatively accessible invitations to participate in the study were sent in written format to four established stroke survivor and aphasia support networks as well as to speech pathologists working in neurological rehabilitation, and via the Aphasia CRE Community of Practice. Interested parties were screened for eligibility via phone/video call using a standardised screening call procedure. Consent was gathered in person or via video call using an aphasia-friendly consent form and communication support materials. Following consent, participants attended a screening visit in which their final eligibility to participate was determined.

Screening call	Screening visit
Conduct: Screening assessment Consent Determine: Age Stroke chronicity Vision acuity Hearing acuity English speaking Concurrent SLT Medical history	Conduct: SADQ-10 PNT WABR ASRS Determine: Presence of depressed mood Anomia severity (PNT) Aphasia severity and type (WABR AQ), auditory-verbal comprehension Verbal apraxia severity (ASRS)
Outcome: Consent and enrolment	Outcome: If assessment results demonstrate participant meets inclusion criteria, proceed to Assessment 1. If participant does not meet inclusion criteria (i.e., PNT is 0 or >140, WAB auditory comprehension <2 SD below the normed reference mean, if SADQ-10 indicates depressed mood (score >14), or the ASRS indicates severe AoS (>8)), participant becomes ineligible.

2. Modifications to assessment procedures to allow remote administration

All assessments were conducted remotely using Zoom videoconferencing software.

Test, subtest, item	Modification
WAB-R Spontaneous Speech B. Picture description	Stimulus image scanned into PDF format
WAB-R Auditory Verbal Comprehension B. Auditory word recognition, items 1-6	Photo of items presented on screen Participant given control of mouse in order to point to items
WAB-R Auditory Verbal Comprehension B. Auditory word recognition, items 7-36	Stimulus images scanned into PDF format Participant given control of mouse in order to point to items
WAB-R Auditory Verbal Comprehension B. Auditory word recognition, items 37-42	Items visible in participant's room selected where possible. Some assumptions were made about items out of view (i.e., light/ceiling) and clarification was sort at end of assessment when accuracy of response was uncertain.
WAB-R Auditory Verbal Comprehension B. Auditory word recognition, items 56-57	Left knee changed to left foot Left ankle changed to left hip
WAB-R Auditory Verbal Comprehension C. Sequential commands, items 5-11	Participants asked to gather pen, comb, and book (or other items e.g., hairbrush). Participant asked to place items in front of them and position camera so items in field of view.
WAB-R Naming and Word Finding A. Object naming	Objects held up to assessor's camera. A gestural cue demonstrating the object in use was given in place of a tactile cue, as required.
Philadelphia Naming Test	Test delivered (i.e., stimulus items shown) via PowerPoint slideshow. One stimulus item per slide. Participants given 12 seconds to name each item.
Corsi Block Tapping Test	Test delivered via PowerPoint slideshow. Each slide shows 9 orange squares positioned at random on a white background. A large green oblong with the word 'Done' is positioned in the bottom right corner of each slide. A 'pulse' effect is applied to selected orange squares. The pulse effect causes the squares to briefly become larger in size and paler in colour before reverting to original size and colour. When the presenter advances the slide, these squares pulse in sequence at 1 second intervals. The number of pulsing squares begins at 2 and progresses to 9 as the slideshow continues. There are two trials at

	each level. When the squares finish pulsing, the 'Done' oblong pulses to indicate the end of the sequence. The participant is given control of the mouse to point to squares in the order that they pulsed.
Test of Everyday Attention Elevator Counting subtest	A digitised version of the original audio stimulus was edited using Logic Pro X (Apple, Inc) to produce separate audio files for each trial of the subtest (n=9). A PowerPoint slideshow was created. The first slide provides task instructions in an aphasia-friendly format (reduced sentence length and complexity, use of images). Each trial slide is animated to display written text synchronised to the spoken prompts. Then when the tones are playing, the screen is white (blank). At the end of the trial, the numbers 1 through 15 appear at the bottom of the slide. The participant is asked to report how many tones were heard.
Test of Everyday Attention Visual Elevator subtest	Stimulus images/pages scanned into PDF format
Picture Span Test	Delivered via PowerPoint slideshow. The first two slides provide aphasia-friendly instructions and demonstration of the task. The five stimulus sheets appear one per slide with a blank slide between. During administration, the assessor starts on a blank slide, provides the auditory stimulus, then moves to the slide containing the corresponding visual stimuli. The participant is given control of the mouse to point to the visual stimuli.

3. Treatment software

A purpose-built web application was designed to deliver this experiment. Two people with aphasia tested the software during the development phase. This resulted in changes to allow alternative input modalities (i.e., mouse, keyboard, touchscreen) and changes to improve clarity of written and pre-recorded spoken instructions provided throughout treatment sessions. The software consisted of three major components which are described briefly here.

A. Database

The software code was stored in Firebase (firebase.google.com). Firebase was also used to store assets (e.g., pictures used in treatment, pre-recorded sound files used as auditory cues), anonymised participant details (e.g., participant ID number), session details (e.g., session number, dose) and user inputted data (e.g., fatigue, motivation, and difficulty ratings).

B. Admin interface

The admin interface was hosted on netlify.app. The admin interface was used to do the following:

- Add assets to the database (pictures, sound files) and link these assets together (e.g., link a picture of a bee to the sound files with the corresponding cues for *bee*).
- Create users
- Create and edit training sets, allocate pictures to set, allocate set to user, assign naming opportunity duration for items in a given set, and allocate set to a dose condition.

Create Training Set

Training Set Name:

User: x | v

Naming Duration:

Dose Condition: x | v

<input type="checkbox"/> bee	<input type="checkbox"/> chicken	<input type="checkbox"/> grapes	<input type="checkbox"/> hand	<input type="checkbox"/> mattress	<input type="checkbox"/> staples
<input type="checkbox"/> strainer	<input type="checkbox"/> turtle	<input type="checkbox"/> accordion	<input type="checkbox"/> anchor	<input type="checkbox"/> ant	<input type="checkbox"/> apple
<input type="checkbox"/> apron	<input type="checkbox"/> armadillo	<input type="checkbox"/> arrow	<input type="checkbox"/> ashtray	<input type="checkbox"/> avocado	<input type="checkbox"/> axe
<input type="checkbox"/> bacon	<input type="checkbox"/> bagel	<input type="checkbox"/> ball	<input type="checkbox"/> banana	<input type="checkbox"/> banjo	<input type="checkbox"/> basketball
<input type="checkbox"/> bat	<input type="checkbox"/> bath	<input type="checkbox"/> battery	<input type="checkbox"/> bear	<input type="checkbox"/> bed	<input type="checkbox"/> belt

- Create and edit naming probe lists for specific user. Only items allocated to a user could be selected for that user's probe lists.
- Create sessions, allocate session to user, include/exclude data collection, allocate naming probe list. Each session generated a unique session key which was used by the participant to access the session via the user interface.

Session Setup

Session Details

User ID

Session

Flanker

Training Set

Session Duration

Other Options

☐ Edit Total Episodes

☒ Include Pre/Post Scores

☒ Include Naming Probe

Choose Naming Probe



- Retrieve session reports which included all session related data in exportable xlsx document format

C. User interface

- Log in via unique session key. A new link with the session key embedded was sent to the participant each treatment day.
- Pre-session fatigue and motivation rating (images used with permission from Gill Pearl at Speakeasy (speakeasy-aphasia.org.uk)).

How tired do you feel?

0 100

- Picture naming probes

Picture naming test

Now you will see some pictures.

Say the name of each picture.

Use one word to name each picture.

You will have 12 seconds to name each picture.

You will not receive any feedback.

The test will take about 10 minutes.

- Treatment was delivered for 45-minutes (time on task) in which pictures were presented on the screen with a series of auditory and visual cues and prompts. The episode structure was identical and is depicted below.

START EPISODE	Show picture				Picture obscured				END EPISODE	BUFFER
	Confrontation cue	Orthographic cue	Spoken word repetition	Delayed recall	Semantic cue	Phonological & phonemic cues	Spoken word repetition	Delayed recall		
	What is this?	E.g., Bat What is it?	It's a bat. Say 'bat'.	Keep it in mind for a moment. What's it called?	It can fly. It's nocturnal. It sleeps upside down. What is it?	It starts with the letter B. /b/ What is it?	It's a bat. Say 'bat'.	Keep it in mind for a moment. What's it called?		
	Naming opportunity	Naming opportunity	Naming opportunity	Naming opportunity	Naming opportunity	Naming opportunity	Naming opportunity	Naming opportunity		

4. Stimulus sets

Participant MR

Total pictures: 60 (15 per set)

Number of probe items: 44 (11 per set). Probe items were items that were named inaccurately three times pre-treatment.

Sets were balanced: Word frequency [$f(3, 56) = 0.287, p = .835$]; Word length [$f(3, 56) = 0.611, p = .61$]

Item #	Word	Probe	Word frequency	Word length	Condition
35	bread	No	2.9557	4	Untreated
38	broccoli	Yes	1.8388	7	Untreated
74	corkscrew	Yes	1.6532	7	Untreated
97	eye	No	3.5075	1	Untreated
99	fingerprint	Yes	2.0453	10	Untreated
111	garlic	Yes	2.2856	5	Untreated
118	gorilla	Yes	2.1959	6	Untreated
124	handcuffs	Yes	2.3404	8	Untreated
164	match	No	3.1926	3	Untreated
165	mattress	Yes	2.3636	6	Untreated
177	nail	Yes	2.8519	3	Untreated
207	rabbit	Yes	2.6513	5	Untreated
222	saw	Yes	3.8075	2	Untreated
250	squid	Yes	1.8808	5	Untreated
291	watch	No	3.7696	4	Untreated

Item #	Word	Probe	Word frequency	Word length	Condition
30	bolt	Yes	2.3560	4	Low
39	broom	Yes	2.2553	4	Low
52	candle	Yes	2.4654	5	Low
55	car	No	3.7122	2	Low
68	clock	Yes	3.2148	4	Low
90	ear	No	3.0441	1	Low
91	eggplant	Yes	1.6232	7	Low
116	globe	Yes	2.3201	4	Low
145	knee	Yes	2.7275	4	Low
161	lock	No	3.2711	3	Low
205	potato	No	2.5866	6	Low
210	rake	Yes	1.9445	3	Low
211	raspberry	Yes	1.8573	7	Low
243	slipper	Yes	1.7160	5	Low
259	sunglasses	Yes	2.1367	4	Low

Item #	Word	Probe	Word frequency	Word length	Condition
16	basketball	Yes	2.6243	9	Moderate
29	blueberry	Yes	1.9494	7	Moderate
102	foot	Yes	3.3128	3	Moderate
123	hand	No	3.7453	4	Moderate
125	hanger	Yes	1.7404	4	Moderate
135	insulation	Yes	1.6435	10	Moderate
218	ruler	Yes	2.1430	4	Moderate
229	screw	No	3.1274	4	Moderate
236	shoe	No	2.9513	2	Moderate
239	sink	Yes	2.8109	4	Moderate
244	slug	Yes	2.2788	4	Moderate
248	sponge	Yes	2.3579	5	Moderate
265	table	No	3.4609	4	Moderate
274	tissues	Yes	1.8633	5	Moderate
294	wheelchair	Yes	2.3365	5	Moderate

Item #	Word	Probe	Word frequency	Word length	Condition
18	bath	Yes	2.3522	6	High
21	bed	No	3.6143	3	High
27	binoculars	Yes	1.8451	10	High
60	chain	Yes	2.8692	4	High
88	drill	Yes	2.6693	4	High
96	escalator	Yes	1.6902	5	High
98	feather	Yes	2.3655	4	High
103	fork	No	2.5011	3	High
113	ginger	Yes	2.3096	5	High
153	leg	No	3.2011	3	High
155	lettuce	Yes	2.0934	5	High
171	mitten	Yes	1.5911	4	High
176	mushroom	Yes	1.9243	6	High
257	strawberry	Yes	2.2227	8	High
275	toast	No	3.0434	4	High

Participant IP

Total pictures: 68 (17 per set)

Number of probe items: 60 (15 per set)

Sets were balanced: Word frequency [$f(3, 64) = 0.182, p = .908$]; Word length [$f(3, 64) = 0.593, p = .622$]

Item #	Word	Probe	Word frequency	Word length	Condition
26	bin	Yes	2.2253	3	Untreated
29	blueberry	Yes	1.9494	7	Untreated
76	couch	Yes	2.8993	4	Untreated
87	doughnut	Yes	2.2122	5	Untreated
119	grapes	Yes	2.1139	4	Untreated
150	lamp	Yes	2.5855	4	Untreated
202	pizza	Yes	2.9253	5	Untreated
226	scissors	Yes	2.3181	5	Untreated
240	skateboard	Yes	1.7853	7	Untreated
243	slipper	Yes	1.7160	5	Untreated
258	sugar	Yes	3.0596	4	Untreated
262	switch	Yes	3.0004	5	Untreated
267	teabag	Yes	0.3010	5	Untreated
271	thermometer	Yes	1.8976	8	Untreated
281	towel	Yes	2.7210	4	Untreated
276	toaster	No	2.0755	5	Untreated
249	spoon	No	2.4330	4	Untreated

Item #	Word	Probe	Word frequency	Word length	Condition
9	avocado	Yes	1.6435	7	Low
18	bath	Yes	2.3522	6	Low
19	battery	Yes	2.6021	6	Low
38	broccoli	Yes	1.8388	7	Low
40	bulb	Yes	2.1732	4	Low
47	cabbage	Yes	2.0414	5	Low
100	flower	Yes	2.7993	4	Low
108	frisbee	Yes	1.7709	6	Low
114	giraffe	Yes	1.7634	5	Low
115	glasses	Yes	3.0039	6	Low
166	medal	Yes	2.4728	4	Low
176	mushroom	Yes	1.9243	6	Low
225	scarf	Yes	2.2201	4	Low
261	swing	Yes	2.9335	4	Low
280	toothbrush	Yes	2.2856	7	Low
255	oven	No	2.4518	4	Low
35	bread	No	2.9557	4	Low

Item #	Word	Probe	Word frequency	Word length	Condition
28	blender	Yes	1.7782	6	Moderate
86	dolphin	Yes	1.8633	6	Moderate
113	ginger	Yes	2.3096	5	Moderate
155	lettuce	Yes	2.0934	5	Moderate
160	lipstick	Yes	2.5211	7	Moderate
162	mailbox	Yes	2.2014	7	Moderate
165	mattress	Yes	2.3636	6	Moderate
199	pillow	Yes	2.6571	4	Moderate
203	plate	Yes	2.9666	4	Moderate
210	rake	Yes	1.9445	3	Moderate
239	sink	Yes	2.8109	4	Moderate
257	strawberry	Yes	2.2227	8	Moderate
274	tissues	Yes	1.8633	5	Moderate
275	toast	Yes	3.0434	4	Moderate
288	tweezers	Yes	1.6532	6	Moderate
68	clock	No	3.2148	4	Moderate
277	toilet	No	2.9800	5	Moderate

Item #	Word	Probe	Word frequency	Word length	Condition
25	bike	Yes	2.7597	3	High
43	bullet	Yes	3.0030	5	High
61	chair	Yes	3.1915	3	High
82	cupcake	Yes	2.0043	6	High
92	elbow	Yes	2.3979	4	High
107	fridge	Yes	2.5944	4	High
111	garlic	Yes	2.2856	5	High
121	hairdryer	Yes	1.0000	7	High
125	hanger	Yes	1.7404	4	High
163	mango	Yes	1.7924	5	High
169	microwave	Yes	2.2201	8	High
175	mug	Yes	2.4166	3	High
227	scooter	Yes	1.8976	5	High
259	sunglasses	Yes	2.1367	4	High
289	umbrella	Yes	2.3222	7	High
181	onion	No	2.1847	5	High
236	shoe	No	2.9513	2	High

Participant LR

Total pictures: 60 (15 per set)

Number of probe items: 60 (15 per set)

Sets were balanced: Word frequency [$f(3, 56) = 0.344, p = .794$]; Word length [$f(3, 64) = 0.593, p = .622$]

Item #	Word	Probe	Word frequency	Word length	Condition
47	cabbage	Yes	2.0414	5	Untreated
64	cherries	Yes	1.9031	5	Untreated
92	elbow	Yes	2.3979	4	Untreated
101	flyswatter	Yes	0.4771	8	Untreated
118	gorilla	Yes	2.1959	6	Untreated
126	harmonica	Yes	1.6532	8	Untreated
130	hinge	Yes	1.4472	4	Untreated
139	jellyfish	Yes	1.6128	7	Untreated
157	lighthouse	Yes	1.8195	6	Untreated
190	peacock	Yes	1.3222	5	Untreated
191	peanut	Yes	2.5599	5	Untreated
208	raccoon	Yes	1.7076	5	Untreated
211	raspberry	Yes	1.8573	7	Untreated
262	switch	Yes	3.0004	5	Untreated
266	tambourine	Yes	1.5185	8	Untreated

Item #	Word	Probe	Word frequency	Word length	Condition
23	belt	Yes	2.9238	4	Low
39	broom	Yes	2.2553	4	Low
58	cauliflower	Yes	1.3802	8	Low
74	corkscrew	Yes	1.6532	7	Low
79	crocodile	Yes	1.8513	8	Low
110	funnel	Yes	1.6721	4	Low
128	headphones	Yes	1.8062	7	Low
131	hippopotamus	Yes	1.2788	11	Low
161	lock	Yes	3.2711	3	Low
168	microscope	Yes	2.0170	9	Low
180	nutcracker	Yes	1.5051	8	Low
186	panda	Yes	1.5798	5	Low
238	shovel	Yes	2.4031	4	Low
259	sunglasses	Yes	2.1367	4	Low
271	thermometer	Yes	1.8976	8	Low

Item #	Word	Probe	Word frequency	Word length	Condition
8	ashtray	Yes	2.0719	5	Moderate
38	broccoli	Yes	1.8388	7	Moderate
42	bulldozer	Yes	1.6232	7	Moderate
48	cactus	Yes	1.9345	6	Moderate
70	coconut	Yes	2.1492	7	Moderate
81	cucumber	Yes	1.8633	8	Moderate
104	forklift	Yes	1.5315	7	Moderate
113	ginger	Yes	2.3096	5	Moderate
122	hammer	Yes	2.5977	4	Moderate
144	kiwi	Yes	1.3802	4	Moderate
148	ladle	Yes	1.4472	4	Moderate
162	mailbox	Yes	2.2014	7	Moderate
204	platypus	Yes	0.6021	8	Moderate
223	saxophone	Yes	1.6532	8	Moderate
243	slipper	Yes	1.7160	5	Moderate

Item #	Word	Probe	Word frequency	Word length	Condition
11	bacon	Yes	2.6021	5	High
27	binoculars	Yes	1.8451	10	High
28	blender	Yes	1.7782	6	High
29	blueberry	Yes	1.9494	7	High
135	insulation	Yes	1.6435	10	High
174	mousetrap	Yes	1.4472	7	High
185	paintbrush	Yes	1.2304	8	High
244	slug	Yes	2.2788	4	High
250	squid	Yes	1.8808	5	High
257	strawberry	Yes	2.2227	8	High
268	teapot	Yes	1.5563	5	High
275	toast	Yes	3.0434	4	High
276	toaster	Yes	2.0755	5	High
279	toolbox	Yes	1.6532	7	High
293	wheelbarrow	Yes	1.3802	7	High

Participant TS

Total pictures: 64 (16 per set)

Number of probe items: 44 (11 per set)

Sets were balanced: Word frequency [$f(3, 60) = 0.376, p = .771$]; Word length [$f(3, 60) = 1.48, p = .299$]

Item #	Word	Probe	Word frequency	Word length	Condition
28	blender	Yes	1.7782	6	Untreated
42	bulldozer	Yes	1.6232	7	Untreated
62	champagne	Yes	2.9279	6	Untreated
41	bull	No	2.8235	3	Untreated
225	scarf	No	2.2201	4	Untreated
13	ball	No	3.3162	3	Untreated
106	frame	No	2.7226	4	Untreated
224	scales	No	2.5933	4	Untreated
151	lawnmower	Yes	1.3802	7	Untreated
173	mouse	Yes	2.6031	3	Untreated
198	pigeon	Yes	2.2601	5	Untreated
202	pizza	Yes	2.9253	5	Untreated
210	rake	Yes	1.9445	3	Untreated
219	saltshaker	Yes	0.7782	8	Untreated
229	screw	Yes	3.1274	4	Untreated
259	sunglasses	Yes	2.1367	4	Untreated

Item #	Word	Probe	Word frequency	Word length	Condition
49	calculator	Yes	1.7243	10	Low
58	cauliflower	Yes	1.3802	8	Low
91	eggplant	Yes	1.6232	7	Low
155	lettuce	Yes	2.0934	5	Low
167	microphone	Yes	2.2330	8	Low
206	pumpkin	Yes	2.4409	7	Low
221	sandcastle	Yes	0.6990	8	Low
239	sink	Yes	2.8109	4	Low
264	syringe	Yes	1.8921	6	Low
275	toast	Yes	3.0434	4	Low
294	wheelchair	Yes	2.3365	5	Low
31	book	No	3.5112	3	Low
14	banana	No	2.5132	6	Low
277	toilet	No	2.9800	5	Low
113	ginger	No	2.3096	5	Low
181	onion	No	2.1847	5	Low

Item #	Word	Probe	Word frequency	Word length	Condition
38	broccoli	Yes	1.8388	7	Moderate
87	doughnut	Yes	2.2122	5	Moderate
92	elbow	Yes	2.3979	4	Moderate
101	flyswatter	Yes	0.4771	8	Moderate
116	globe	Yes	2.3201	4	Moderate
145	knee	Yes	2.7275	4	Moderate
168	microscope	Yes	2.0170	9	Moderate
174	mousetrap	Yes	1.4472	7	Moderate
61	chair	No	3.1915	3	Moderate
170	mirror	No	2.9405	4	Moderate
231	seal	No	2.6749	3	Moderate
111	garlic	No	2.2856	5	Moderate
184	owl	No	2.1399	2	Moderate
243	slipper	Yes	1.7160	5	Moderate
256	strainer	Yes	1.0792	6	Moderate
261	swing	Yes	2.9335	4	Moderate

Item #	Word	Probe	Word frequency	Word length	Condition
18	bath	Yes	2.3522	6	High
22	bee	Yes	2.3962	2	High
29	blueberry	Yes	1.9494	7	High
88	drill	Yes	2.6693	4	High
118	gorilla	Yes	2.1959	6	High
128	headphones	Yes	1.8062	7	High
144	kiwi	Yes	1.3802	4	High
262	switch	No	3.0004	5	High
258	sugar	No	3.0596	4	High
76	couch	No	2.8993	4	High
169	microwave	No	2.2201	8	High
274	tissues	No	1.8633	5	High
187	paperclip	Yes	0.4771	8	High
220	sandal	Yes	0.9542	5	High
223	saxophone	Yes	1.6532	8	High
244	slug	Yes	2.2788	4	High

5. Analysis workflow

A detailed summary of the analysis workflow is available here: [Outcomes from a pilot dose comparison study of naming therapy in aphasia - Analysis workflow](#)