Applying adaptive distributed practice to self-managed computer-based anomia treatment: a single-case experimental design

Yina M. Quique[1], Robert Cavanaugh[2], Erica Lescht[2], William S. Evans[2]

[1]Center for Education in Health Sciences, Northwestern University, Chicago, IL, USA.
[2]Department of Communication Sciences and Disorders, University of Pittsburgh, Pittsburgh, PA, USA.

Corresponding author: Yina M. Quique
yinaquique@northwestern.edu
Center for Education in Health Sciences
633 N. Saint Clair St.
20th Floor
Chicago, IL 60611

Keywords: aphasia, anomia, adaptive distributed practice, self-managed treatment

Funding and acknowledgments: Research reported in this publication was supported by the UPMC Competitive Medical Research Fund (CMRF) awarded to William Evans. Our sincere gratitude to Brandon Nguy and Gillian Reed for their assistance in finding the picture exemplars and Joshua Peckman for his assistance in scoring.

Declaration of Interest statement: The authors report no declarations of interest.

Abstract

Introduction: There is a pressing need to improve computer-based treatments for aphasia to increase access to long-term effective evidence-based interventions. The current single case design incorporated two learning principles, adaptive distributed practice and stimuli variability, to promote acquisition, retention, and generalization of words in a self-managed computer-based anomia treatment.
Methods: Two participants with post-stroke aphasia completed a 12-week adaptive distributed practice naming intervention in a single-case experimental design. Stimuli variability was manipulated in three experimental conditions: high exemplar variability, low exemplar variability, and verbal description prompt balanced across 120 trained words. Outcomes were assessed at 1-week, 1-month, and 3-months post-treatment. Statistical comparisons and effect sizes measured in the number of words acquired, generalized, and retained were estimated using Bayesian generalized mixed-effect models.
Results: Participants showed large and robust acquisition, generalization, and retention effects. Out of 120 trained words, participant 1 acquired ~77 words (trained picture exemplars) and ~63 generalization words (untrained picture exemplars of treated words). Similarly, participant 2 acquired ~57 trained words and ~48 generalization words. There was no reliable change in untrained control words for either participant. Stimuli variability did not show practically meaningful effects.
Conclusions: These case studies suggest that adaptive distributed practice is an effective method for re-training more words than typically targeted in anomia treatment research (~47 words on average per Snell et al., 2010). Generalization across experimental conditions provided evidence for improved lexical access beyond what could be attributed to simple stimulus-response mapping. These effects were obtained using free, open-source flashcard software in a clinically feasible, asynchronous format, thereby minimizing clinical implementation barriers. Larger-scale clinical trials are required to replicate and extend these effects.

1. Introduction

Approximately one-third of stroke survivors are diagnosed with aphasia, and more than two million people with aphasia are currently living in the United States (Simmons-Mackie & Cherney, 2018). Aphasia negatively affects communication ability, life participation, and long-term quality of life (Cruice et al., 2006; Dalemans et al., 2010). Although people with aphasia continue to improve their communication abilities in response to long-term speech and language pathology services (Allen et al., 2012; Brady et al., 2012; Brady et al., 2016), access to aphasia rehabilitation services is often limited by constraints such as insurance coverage or transportation (Ostwald et al., 2009). For example, Cavanaugh et al. (2021) reported that people with aphasia received a median of 10 outpatient sessions of speech-language pathology treatment in their first year post-stroke. As a result, there is a critical need to develop alternative methods for accessing evidence-based aphasia treatment alongside or in lieu of existing access to clinical rehabilitation services.

Self-managed, computer-based treatments are one promising option for offering low-cost, accessible, and prolonged interventions for people with aphasia. Computer-based treatments have been developed for aphasia for independent practice or supplemental treatment purposes showing promising results (e.g., Davis & Copeland, 2006; Meltzer et al., 2018; Stark & Warburton, 2018). When self-managed or administered with a small degree of support from caregivers, computer-based interventions can help people with aphasia regain independence and agency during their recovery process (Palmer et al., 2020; Palmer et al., 2012). Moreover, cost analyses have shown that these interventions can be less expensive than standard care (Wenke et al., 2014) because they increase the total amount of direct treatment while reducing the costs of face-to-face interventions.

Despite the established benefits of current computer-based treatments (e.g., Cherney, 2010; Palmer et al., 2020), there is a need to improve their effectiveness and *treatment efficiency* (i.e., the largest-possible treatment outcomes in the shortest amount of treatment time). Treatment efficiency is particularly important because self-managed treatment comes with an opportunity cost: every hour spent on self-managed treatment is an hour not spent in social communication environments or meaningful life activities. Computer-based treatments may be particularly well-suited to improving treatment efficiency, as they have the capability of implementing complex algorithms focused on optimizing the “return on investment” for time spent in therapy.

To improve treatment efficiency in aphasia treatment, computerized treatments should be informed by a theoretical understanding of aphasia recovery and the processes that promote learning in aphasia. The current study incorporated two principles from the learning literature thought to improve acquisition, retention, and generalization: distributed practice and stimuli variability. We chose to focus on anomia (i.e., word-finding deficits) treatment since anomia is a hallmark of aphasia (Goodglass, 1980) and endorsed as a primary frustration (e.g., Johansson et al., 2012). We examined distributed practice and stimuli variability by implementing a self-managed computer-based anomia treatment using Anki open-source flashcard software[1]. This software incorporates *adaptive distributed practice* and enables the customization of multimedia flashcards, which allowed us to manipulate the *stimuli variability* of trained items.

1.1 Learning effects of distributed practice in neurotypicals and people with aphasia.

The effects of spaced practice on learning have a long history. In general, distributed practice, where trials are more spaced out over time, has been found to enhance learning and retention in a wide range of tasks compared to massed practice, where trials are spaced more closely together in time (Karpicke, 2017). The distributed practice effect is one of the most replicated and reliable findings in the learning literature (Cepeda et al., 2006; Delaney et al., 2010; Toppino & Gerbier, 2014). One influential account of this effect is that distributed practice allows for memory decay during learning, which increases the effort required for subsequent retrieval. The increased effort has been hypothesized to provide a *desirable difficulty*, where successful effortful retrieval attempts improve memory consolidation and encoding (Bjork & Bjork, 1992).

Middleton et al. (2020) recently published a comprehensive review summarizing the evidence supporting distributed practice for anomia treatment in aphasia. They reported that spacing practice trials over time (Middleton et al., 2019; Middleton et al., 2016), increasing trial spacing within a session (Middleton et al., 2016), increasing spacing across sessions (Schuchard et al., 2020), and increasing the time between learning sessions (Ramsberger & Marie, 2007; Sage et al., 2011) can all enhance learning and retention during anomia treatment.

Recent research in education and language learning has begun to explore various *adaptive* distributed practice algorithms with promising results (Eglington & Pavlik Jr, 2020; Settles & Meeder, 2016; Tabibian et al., 2017; Tabibian et al., 2019). In *adaptive distributed practice*, item scheduling depends on ongoing performance accuracy, where difficult-to-learn items receive more frequent practice than more easily learned items over time. One common implementation of adaptive distributed practice is based on the forgetting curve model of memory decay (Ebbinghaus, 1885; Murre & Dros, 2015). The forgetting curve describes a phenomenon in which memory retention decreases over time unless learned information is re-reviewed and re-retrieved.

Anki is an open-source flashcard software program that schedules reviews using adaptive distributed practice. Each flashcard goes through an initial “learning phase” where flashcards are practiced frequently until they are answered accurately a certain number of times (the user self-rates accuracy, which requires error awareness). Then, flashcards are scheduled at increasingly expanding intervals determined by an algorithm based on the Ebbinghaus forgetting curve (Ankiweb, 2006; Ebbinghaus, 1885). If at a future point, a given flashcard is answered incorrectly, it goes back through a “re-learning phase” where it has to be answered correctly again a certain number of times before returning to expanding interval practice (Figure 1). Flashcard A in Figure 1 represents a word that is easily re-acquired, which quickly moves through the learning phase into expanding interval practice without subsequent memory lapses. In contrast, Flashcard B represents a harder-to-acquire word, requiring more initial practice and

[1] https://apps.ankiweb.net/

multiple times passing through a re-learning phase. For both flashcards, final accuracy during the practice period is consistently high, but different practice trajectories are required to arrive at the same final criteria (i.e., 28 reviews for Flashcard A vs. 42 reviews for Flashcard B). Anki has been shown to be effective in other learning contexts, such as academic vocabulary learning, second language acquisition (Altiner, 2019; Rana et al., 2020; Seibert Hanson & Brown, 2020), and anomia treatment in primary progressive aphasia (Evans et al., 2016).

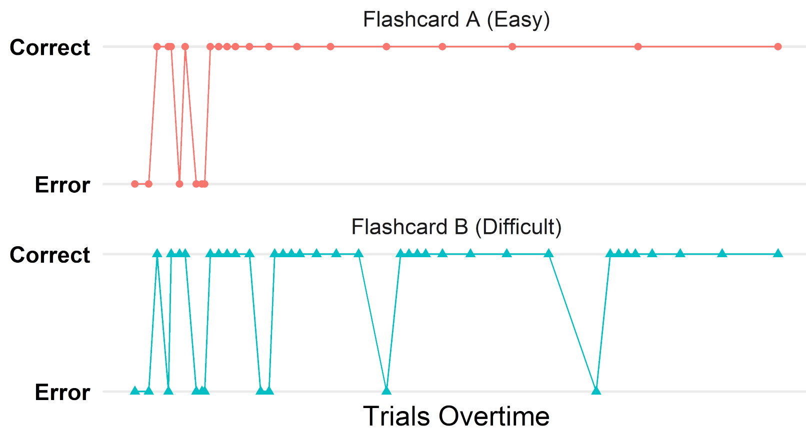

Figure 1. Schematic representation of adaptive distributed practice for an easy word (flashcard A) and a more difficult word (flashcard B). Correct responses lengthen retrieval intervals while errors shorten them.

The potential positive impact of adaptive distributed practice on anomia treatment outcomes is three-fold. *First*, adjusting practice schedules at the item level, based on ongoing performance, may maximize the benefits of effortful retrieval (Middleton et al., 2015) while minimizing the potential costs of failed retrieval attempts and subsequent error learning (Evans et al., 2019). In other words, adaptive distributed practice may maximize the strength of memory encoding per trial by maintaining item-specific desirable difficulty. *Second*, adaptive distributed practice increases the total number of items that may be practiced within a limited period of time since items are only scheduled when necessary. *Third*, by scheduling each trial based on item-specific performance, this approach dynamically adjusts treatment difficulty to meet individual needs, which may provide a practical approach to precision medicine for anomia treatment. For example, people with more severe aphasia may receive shorter spacing and fewer additional items based on their ongoing performance.

In sum, adaptive distributed practice may lead to better treatment efficiency and long-term retention. While anomia interventions in aphasia research typically train small sets of words (e.g., 47 words on average, with a larger cluster of anomia studies training between 30-40 words; Snell et al., 2010), adaptive distributed practice could allow more words to be trained. In addition, the effects of aphasia treatment are generally not well maintained (e.g., Menahemi-Falkov et al., 2021), but a treatment based on the forgetting curve model of memory decay has the potential to engender lasting retention effects. However, the benefits of *adaptive* distributed practice for anomia treatment have not been examined in stroke-related aphasia. Also, algorithms for adaptive distributed practice are calibrated based on neurotypical learning patterns, and it's unknown if they are well suited for people with aphasia. Therefore, the first aim of the current work is to evaluate the initial efficacy of an adaptive distributed practice intervention on acquisition, retention, and generalization of more words than are typically targeted in anomia treatment research.

1.2 Stimuli variability as a facilitator of generalization in aphasia

Although adaptive distributed practice may improve treatment efficiency and retention, such gains would be even more meaningful if they generalize outside the trained context. Aphasia researchers have distinguished between "within-level generalization," the change to untreated stimuli within the same linguistic level as the treatment (e.g., from treated words to untreated words, or from trained exemplars to untrained exemplars of the same word), and "across-level generalization," the change at a linguistic level different to the treatment (e.g., from words to sentences; Webster et al., 2015). Although direct treatment effects on treated words have generally been robust in published anomia interventions (e.g., Wisenburn & Mahoney, 2009), effects of within and across-level generalization have generally been modest (e.g., Carragher et al., 2013; Kendall et al., 2015; Quique et al., 2019; Thompson, 1989; Webster et al., 2015).

One potential reason for these modest generalization outcomes is that many anomia treatments rely on naming a single picture exemplar of the target word during treatment and assessment (e.g., Gravier et al., 2018; Kendall et al., 2019; Middleton et al., 2015). Regardless of other active ingredients of the therapy, overtraining a single picture exemplar through stimulus-response mapping could create a potent memory retrieval cue (e.g., Holland, 2008; Spence, 1950; Thorndike, 1898) that may limit within and across-level generalization.

Early knowledge transfer literature suggested that increasing the number and variability of exemplars can benefit learning and generalization (Gick & Holyoak, 1987 for a summary). Similarly, the developmental literature (e.g., Aguilar et al., 2018; Perry et al., 2010; Twomey et al., 2014) and applied behavioral analysis (e.g., Stokes & Baer, 1977; Stokes & Osnes, 2016; Swan et al., 2016) recommend the use of multiple exemplars to promote learning and generalization. Based on some of these findings, Thompson (1989) suggested that training multiple examples might promote generalization in aphasia. However, it is possible that increasing exemplar variability could reduce treatment efficiency for acquisition if additional time is required to train multiple exemplars to the same level of performance. Therefore, the second aim of the current work is to evaluate the effects of stimulus variability (defined here in terms of exemplar number and prompt type) on acquisition, retention, and within-level generalization.

1.3 The current study

In summary, while previous anomia treatment research has focused on training small sets of words (47 on average; Snell et al., 2010) in the hopes of inducing generalization effects (e.g., Boyle, 2010; Boyle, 2011; Webster et al., 2015), the current work proposes a novel alternative: using adaptive distributed practice to directly train a larger set of words in a treatment-efficient manner. In addition, mechanisms thought to improve generalization should be explored to improve functional outcomes. Therefore, we evaluated the initial efficacy of adaptive distributed practice and stimuli variability by implementing a self-managed anomia treatment using Anki open-source flashcard software.

Prediction 1: Anki-based anomia treatment will lead to efficient acquisition, retention, and within-level generalization to untrained picture exemplars of trained words for more words than are typically targeted in anomia treatment research. In addition, we do not predict within-level generalization to untreated words, which serve as a control condition.

Prediction 2: Increasing stimuli variability (i.e., more picture exemplars or a verbal description compared to training only a single picture exemplar) will lead to improved within-level generalization to untrained picture exemplars of trained words for both acquisition and retention to a practically meaningful degree.

2. Methods

2.1 Study design.

We used an A-B design with control and follow-up (Tate & Perdices, 2019, p. 93). Each participant was probed during 3 baseline sessions, 12 weekly treatment sessions, and 3 follow-up sessions. Follow-up sessions were planned for 1 week, 1 month, and 3 months post-treatment.

2.2 Participants

Participants were recruited from the University of Pittsburgh program, from the Aphasia Recovery Connection online community, and from local clinician referrals. The Institutional Review Board at the University of Pittsburgh approved the study (IRB# 19060039). Informed consent was obtained before any research procedures were completed. Enrolled participants did not receive concurrent speech-language treatment for the duration of the study.

For initial assessment, participants completed the Comprehensive Aphasia Test (CAT; Swinburn et al., 2004), the Cactus and Camel Test (CCT; Adlam et al., 2010; Bozeat et al., 2000), and the Duffy protocol (Duffy, 2005). The first and third authors administered these assessments. While assessments were initially administered in-person at a lab at the University of Pittsburgh, the study protocol shifted to online administration part-way through the study due to COVID-19.

Participants were enrolled in this study using a two-stage enrollment criteria. In the first stage, participants were required to have an existing diagnosis of aphasia secondary to a stroke at least six months post-onset. The presence of aphasia was confirmed via initial assessment using the CAT, where participants had to demonstrate impairment on at least 2/8 test domains to qualify. Participants needed to be native English speakers. Participants were excluded if they reported a concomitant history of a neurodegenerative disorder or showed a severe motor speech disorder, as measured by the Duffy protocol (Duffy, 2005). To meet the second stage of enrollment criteria, participants had to demonstrate the ability to complete all treatment steps independently and with fidelity following four synchronous training sessions (see Treatment Procedures below). This was required because the intervention relied on self-managed asynchronous independent home practice during the last ten weeks of the treatment phase, where participants had to self-rate their performance accurately in the Anki software for the adaptive algorithms to function appropriately.

Seven people with aphasia were recruited, three met the first stage criteria, and two of which met the second stage criteria. Participant 1 was a 50-year-old male with aphasia from a left hemisphere stroke of approximately 24 months post-onset. His scores on the CAT suggested moderate language impairments (mean T-score = 46.8). He demonstrated more severe impairments in repetition (T-score = 32) relative to his overall severity. He demonstrated relative strengths in naming (T-score = 54). At the time of enrollment, he reported that he was independent in activities such as scheduling sessions, driving, and attending in-person or online sessions. He also demonstrated strong motivation to participate in treatment.

Participant 2 was a 53-year-old male with aphasia from a left hemisphere stroke of approximately 18 months post-onset. His scores on the CAT suggested moderate language impairments (mean T-score = 45.7). His scores on the CAT indicated more severe impairments in spoken comprehension (mean T-score = 38) relative to his overall severity. He demonstrated relative strengths in reading (mean T-score = 49). He required the support of his spouse to schedule online sessions. Baseline testing on the CAT and CCT is shown in Table 2. Naming probe accuracy for each participant by session and treatment condition is presented in Figure 3.

2.3 Treatment conditions and stimuli development

This study consisted of one untreated control condition and three experimental conditions: high exemplar variability, low exemplar variability, and verbal description prompt (see Figure 2). Each condition included 40 items, for a total of 120 treated and 40 control items (total n =160 items per participant). In the *low exemplar variability condition*, each item was trained using one picture exemplar. This low-variability condition was intended to match typical anomia treatment protocols, which might lead to reduced within-level generalization by over-training specific stimulus-response mappings. In contrast, in the *high exemplar variability condition*, each item was trained using three picture exemplars. This was intended to increase stimuli variability and promote within-level generalization (e.g., Aguilar et al., 2018; Perry et al., 2010; Twomey et al., 2014). In the *verbal description condition*, each item was trained by presenting a short verbal description prompt in both written and auditory formats. This was intended to increase stimuli variability by providing a prompt without relying on trained pictures (Edmonds et al., 2014; Edmonds et al., 2009), which also increased the overall complexity of the stimulus-response mapping.

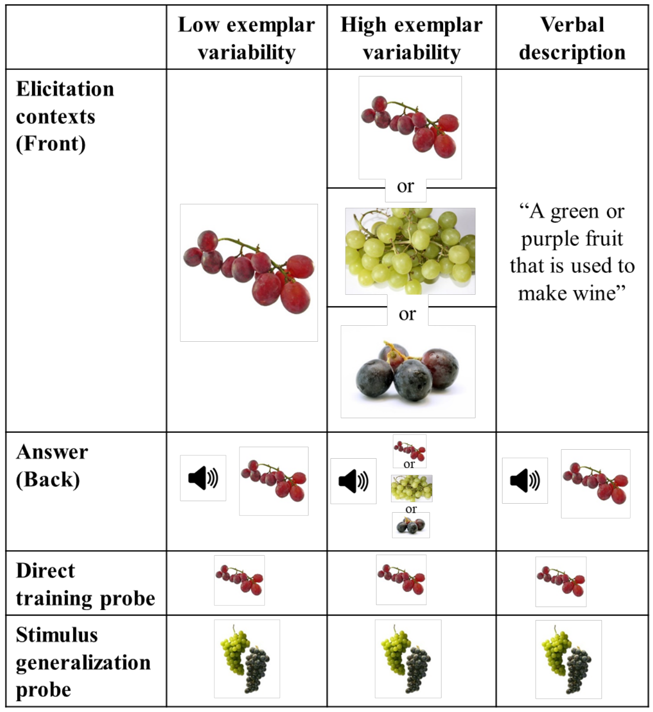

Figure 2. Examples of the experimental conditions, including the front of the flashcard, the back, and the picture exemplars used for probes.

In order to select 160 items for each condition, participants who met the first stage of enrollment criteria were asked to name a battery of 362 noun pictures for stimuli selection. This naming battery was assembled from publicly available normative picture databases (Brodeur et al., 2010; Brodeur et al., 2014; Moreno-Martínez & Montoro, 2012), excluding images that were included in the CAT. Each item in the naming battery was characterized in terms of lexical frequency, number of phonemes, and age of acquisition (Balota et al., 2007; Brysbaert et al., 2012; Kuperman et al., 2012) for the purposes of matching condition lists for difficulty.

In order to manipulate stimuli variability in terms of picture exemplars, we selected three additional images for each item, including variations in either the item's perspective, color, size, or number to ensure each picture was visually distinct. Two independent raters reviewed each additional image to ensure they were acceptable exemplars of the intended target. Exemplar images identified as problematic by either rater (e.g., not representative of the target image) were replaced by new images and re-evaluated until considered acceptable by both raters. All exemplar images were drawn from internet searches of images labeled public domain or creative commons reuse with modification. We also created a short verbal description prompt highlighting each item's main characteristics. Two independent raters evaluated these verbal descriptions to ensure they were acceptable descriptions of the intended target. Descriptions identified as problematic by either rater (e.g., not an adequate description of the target word) were revised and re-evaluated until both raters considered them acceptable. As a result, all 362 naming battery items had three additional picture exemplars and a verbal description prompt. This allowed any item to be selected for any condition to facilitate difficulty matching between conditions for a given participant and provided an untrained picture exemplar for each item to assess within-level generalization during probes.

2.4 Probe selection and administration

Participants were asked to complete the 362-item naming battery of the original target images on two different occasions at least one day apart. Items were presented in random order on each occasion. We selected 160 items from those with an accuracy of ≤ 50% across testing points for each participant. To match production difficulty across conditions, we used a published algorithm that provides weights based on word frequency, number of phonemes, and age of acquisition (Fergadiotis et al., 2015). We calculated the resulting difficulty score for each item. Then, items were divided into four groups (three experimental conditions and a control group) by arranging them in order of difficulty and sequentially assigning them into each group to ensure they contained the same range and density of difficulty scores for each participant.

Treatment probes consisted of a confrontation naming task assessing trained and untrained picture exemplars of treated words and untrained control words. Trained picture exemplars of treated words assessed direct treatment effects; untrained picture exemplars of treated words assessed within-level generalization; untreated words served as a control condition to distinguish treatment effects from simple probe exposure. Treatment effects were assessed in terms of initial acquisition (naming performance up to the one-week follow-up) and retention (up to the three months follow-up). Naming accuracy was scored online during probe administration and was confirmed and hand-corrected offline by a second independent rater using an audio recording of the session. Raters were not blinded to time points or study hypotheses, although the second rater was blinded to experimental conditions.

There were 120 treated words with two pictures per word (one trained, one untrained) and 40 untrained control words with one picture per word, for a total of 280 probe pictures. Trained picture probes were provided as part of the flashcard "answer" for each target (Figure 2). Only one of the three trained pictures was probed in the high exemplar variability condition. Untrained picture probes were not provided during treatment.

Pictures were pseudo-randomized into two lists, each containing 140 pictures: 60 trained picture exemplars, 60 untrained picture exemplars, and 20 untreated control pictures. We administered both lists in a single session during baseline and follow-up and in alternative sessions during the treatment phase to reduce the overall testing burden and exposure effects. Probes were administered using PsychoPy 2 version 2.0.44 (Peirce et al., 2019) on a Dell laptop computer.

2.5 Treatment administration

Treatment started with two weeks of initial training. Each of these weeks had two sessions focused on ensuring that participants knew how to use Anki and how to complete each step of the flashcard practice with fidelity and independence (which was required for the second stage of enrollment). Training relied on systematic instruction techniques (Mateer & Sohlberg, 2003; Sohlberg & Turkstra, 2011) provided by the senior author.

For participants who met the criteria for continued enrollment, the initial two weeks of training were followed by ten weeks of combined synchronous practice (in-person or online) and asynchronous independent home practice. This treatment approach was designed to reflect a clinically feasible treatment design and was based on a case report using Anki in the clinic for an individual with primary progressive aphasia (Evans et al., 2016). Synchronous practice sessions lasted up to 60 minutes per session or until they ran out of scheduled flashcard reviews, which allowed for reinforcement of the treatment steps for home practice and provided an opportunity to administer weekly probes. The independent home practice was asynchronous (Cherney et al., 2011), and participants were instructed to practice at least four days a week for at least 20 minutes a day (or until they ran out of flashcards scheduled for that day).

The anomia treatment consisted of flashcard-based distributed practice in which participants attempted to effortfully retrieve (Middleton et al., 2016) target words in response to cues (pictures or verbal descriptions). Specifically, in each trial, participants saw a picture or written prompt of the target in Anki and attempted to name it with a single complete response, then pressed "show answer" to see and hear the target's correct form along with its trained picture. They then self-rated their performance as accurate or inaccurate by pressing the "good" (correct) or "again" (incorrect) buttons, which adaptively adjusted the future scheduling of the flashcard (Figure 2). Whenever they could not name the word accurately, they were asked to listen to the target (by clicking a "replay audio" button) and repeat it three times correctly before self-rating their performance as inaccurate and moving on to the next practice trial.

As noted above, participants had to demonstrate the ability to complete all treatment steps with fidelity and independence by the end of the four initial synchronous practice sessions to continue study enrollment. These steps consisted of being able to a) attempt to name the target in response to the initial prompt, b) click "show answer" after making an attempt, and c) being able to self-rate performance with ≥85% accuracy (which was crucial to ensure appropriate flashcard scheduling based on the adaptive algorithm).

Anki has a number of settings that can be specified. Flashcards in the "active learning" state need to be answered correctly a certain number of times before being categorized as "learned." We set the adaptive distributed practice parameters such that words needed to be answered correctly three times in a row (immediately, then at one-minute, and five-minute intervals) to be categorized as learned. Once a flashcard was categorized as "learned," it was automatically scheduled at ever-increasing intervals (up to a maximum set at 15 days). Intervals continue to increase to the maximum until answered incorrectly, at which point the flashcard returned to the "active learning" state, requiring three correct responses in a row before returning again to the "learned" expanding interval state. Up to 24 new flashcards were provided per day.

2.6 Analyses

Bayesian generalized linear mixed-effects models were used to evaluate change in naming accuracy, modeled separately by participant. Bayesian analyses are advantageous because they match researchers' general interpretation of statistical results as "the probability of a hypothesis given the data," which is a misinterpretation of p-values in the frequentist framework. This approach also characterizes the uncertainty of the effects of interest through posterior distributions and their credible intervals (e.g., a 90% credible interval can be interpreted as a 90% probability that an effect falls within the interval; see Kruschke & Liddell, 2018). Analyses were completed using Stan (Carpenter et al., 2017; Stan Development Team, 2020) accessed via the BRMS package (Bürkner, 2018) using the R statistical software version 4.0.2 (R Core Team, 2020). Data and analysis scripts for the following analyses have been made publicly available at https://osf.io/sfutm.

For all models, a Bernoulli probability distribution with a logit link function was used to model trial-level responses (i.e., correct and incorrect naming attempts). All models included weakly informative priors for all beta coefficients using a student-T distribution (3 degrees of freedom, mu = 0, and sigma = 2.5). This prior is characterized by a symmetrical distribution roughly between -5 and 5 logits centered around zero (no effect), which comprises a reasonable range of values for all beta coefficients. Default BRMS priors were used for all other aspects of these analyses. A visual prior predictive check was used to ensure that these priors characterized reasonable possible estimates.

For prediction 1 analyses, we evaluated the initial efficacy of adaptive distributed practice as an efficient method for acquiring, generalizing, and retaining more words than is typically targeted in anomia treatment research (collapsing across experimental conditions). For this purpose, models were fit to describe performance across study timepoints and then used to estimate effect sizes in terms of the number of words *acquired* between treatment baseline and treatment exit (measured at 1-week post-treatment) and the number of words *retained* between treatment exit and follow-up. We chose the number of words acquired and retained as an unstandardized measure of effect size because it maps directly to our prediction, is interpretable for clinicians, and it allows the comparison of treatment efficacy between anomia studies in functional terms (i.e., how many words were gain in relation to how much time was spent in treatment). Moreover, this choice of effect size is based on a well-established model structure (Baek et al., 2014; Huitema, 2011; Huitema & McKean, 2000) and builds upon previous work in our field (e.g., Swiderski et al., 2021; Wiley & Rapp, 2019).

For acquisition and generalization, we implemented interrupted time series models described by Huitema and McKean (2000). To examine *acquisition* effects, we modeled item-level responses for all *trained* pictures during the baseline and treatment phases. To examine *within-level generalization* effects, we modeled item-level responses for *untrained* pictures of trained items during the baseline and treatment phases. The interrupted time series model structure has population-level[2] effects (often described as "fixed effects" in the frequentist framework) for baseline slope, level change, and slope change. Baseline slope characterizes the presence of a stable, rising, or declining trendline during the baseline phase. Level change characterizes the immediate change in naming performance at the onset of treatment (i.e., the

[2] Since models were run separately by participant, the "population" here refers to the item level, in this case, picturable nouns in English. In other words, analyses allow us to draw inferences about how these particular people with aphasia would respond to training for any similar words in English, which is appropriate for the case study approach employed here.

difference between the baseline and treatment trendlines at the first treatment probe). Slope change assesses whether the trendline during the treatment phase differs from the trendline during the baseline phase. Positive level and slope changes provide evidence of treatment effects (i.e., changes in performance over time attributable to treatment).

Each model was also evaluated with a quadratic slope change term to account for a substantial non-linear change in naming accuracy during treatment (e.g., a diminishing rate of improvement over time). We compared models with and without this quadratic term using Bayes factors (Rouder et al., 2009). The more parsimonious model without the quadratic term was used unless the Bayes factor indicated at least moderate (>3) evidence of better fit (Jeffreys, 1998). The interrupted time series model was also applied to *untreated* words to establish experimental control. There are a number of benefits of using this modeling approach, including specific model-based checks for distinguishing between rising baseline performance and treatment-related change (e.g., Evans et al., 2021). However, for the current case study, we applied this approach to determine whether effect sizes fit our data (full model results are available at https://osf.io/sfutm).

We estimated effect sizes from each model by taking the difference in the posterior distributions for the estimated number of words correct between treatment exit and the last baseline probe. Hence, effect sizes represent an estimate of the median number of words gained between these time points, accompanied by a 90% credible interval for each participant.

To examine *retention*, the model for each participant included population-level effects for timepoint (probe sessions at treatment exit, one-month, and three-month follow-up, with treatment exit as the reference level), stimulus item type (trained versus untrained exemplars, with trained as the reference level), and their interaction. The random effects for these models included a random intercept for items and a random slope for timepoint. Effect sizes were estimated using the same method, taking the difference in performance between one-month follow-up and treatment exit, and three-month follow-up and treatment exit. As a result, retention effect sizes with a negative value reflect the number of words forgotten from treatment exit to a given follow-up time point.

For prediction 2 analyses, we examined the effects of stimuli variability in acquisition, within-level generalization, and retention from sessions 3 to 16 (from the last baseline to 1-week follow-up) to determine if increasing stimuli variability improves treatment outcomes to a practically meaningful extent.

For acquisition and generalization, models included population-level effects for session (last baseline session to the first follow-up session) and condition (high exemplar variability, low exemplar variability, verbal description, with low exemplar variability as the reference level), and their interaction. With this approach, interaction effects tested whether, compared to the low exemplar variability condition, the higher stimulus variability conditions (high exemplar variability or verbal description) resulted in greater differences in treatment response during treatment. The random effects for these models included a random intercept for items and a random slope for session. To examine *acquisition* effects for this question, we modeled item-level responses for *trained* pictures during the treatment phase. To examine *within-level* *generalization* effects, we modeled item-level responses for *untrained* pictures of trained items during the treatment phase.

For *retention*, models included population-level effects for session (treatment exit, one-month, and three-month follow-up, with treatment exit as the reference level), condition (same as above), and their interaction. The interaction effects tested whether, compared to the low exemplar variability condition, the higher stimulus variability conditions (high exemplar variability or verbal description) resulted in greater differences in retention. The random effects for these models included a random intercept for items and a random slope for session.

Models were run separately by participants and stimuli item type, resulting in four models per participant. As a reminder, we addressed prediction 2 by evaluating the condition by session interaction terms reported in each model, resulting in a total of 12 interaction terms per participant (Table 3). Looking across these interaction terms provides statistical tests of whether stimuli variability affected treatment outcomes (measured in terms of acquisition, one-month retention, and three-month retention) for direct training and generalization (see complete model structures and output at https://osf.io/sfutm).

To determine whether interaction effects reflected a meaningful difference, we used the Region of Practical Equivalence (ROPE; Kruschke & Liddell, 2018). ROPE allowed us to set stringent criteria to assess whether increasing stimuli variability would have a large enough effect on treatment outcomes to be practically relevant, not merely statistically reliable. The ROPE approach allows redefining the null hypothesis from a point-null hypothesis to a range of values considered small to have practical significance. In other words, rather than testing whether a number is different from zero, ROPE defines a range of values for which the effect would be considered, practically speaking, no different from zero (combining significance testing with effect sizes; see Harms & Lakens, 2018; Makowski et al., 2019). Thus, we used ROPE to evaluate null effects of stimuli variability (i.e., whether or not the lack of a reliable effect could be interpreted as an effect practically equivalent to zero).

To define the range of ROPE values determined to be equivalent to the null region (i.e., practically equivalent to zero in terms of outcomes), we used a standardized probe value of ± 0.18, as recommended by Kruschke (2018), which roughly corresponds to a negligible effect according to Cohen (1988). The ROPE range of parameters was then compared against the posterior distribution. If > 97.5% of the posterior distribution fell within the ROPE, the effect was considered practically equivalent to zero (i.e., “null effect”), indicating that effects were not large enough to warrant further consideration for clinical implementation. If < 2.5% of the posterior distribution fell outside the ROPE, the effect was considered practically meaningful and worthy of further consideration. ROPE percentages between these values were considered to provide uncertain evidence (see Kruschke & Liddell, 2018). We calculated ROPE using the R package BayestestR (Makowski et al., 2019).

2.7 Model Convergence and Fit

For all models, we ran four independent Hamiltonian Markov Chain Monte Carlo (MCMC) chains with 6000 iterations. The initial 1000 chains were used as a warmup and were not included in the parameter's estimation. We examined model convergence by checking the Gelman-Rubin Potential Scale Reduction statistic (ensuring that $\hat{R}$ values were less than or equal to 1.01), the number of effective samples (ensuring >400 effective samples), and trace plots. Model fit was examined using posterior predictive checks. Model details, including priors, model convergency, the number of effective samples, and posterior predictive checks, are available in the OSF link provided above.

3. Results

Seven people with aphasia were recruited, three of which met the first stage criteria for initial enrollment and two of which met the second stage criteria for enrollment in treatment (participants 1 and 2, see Table 1 for demographic information). The four people with aphasia who did not meet initial enrollment criteria were excluded based on very mild CAT performance.[3] The person who did not meet the second enrollment stage was excluded due to difficulties demonstrating fidelity to the treatment steps necessary for independent practice (including accurate self-rating of responses). For fully enrolled participants 1 and 2, apraxia of speech was absent as measured by the Duffy protocol (Duffy, 2005). Assessments and study procedures were administered in-person for participant 1, except for his last two follow-up sessions, which had to shift to an online format due to pandemic-related restrictions. Also, follow-up sessions for this participant coincided with the onset of COVID-19 in the US, which led to a delay in the planned assessment intervals (the first follow-up session was completed 11 days post-treatment, the second 2 months post-treatment, and the third 3 months post-treatment). All assessments and study procedures were administered via Zoom for participant 2. Both participants completed the planned number of baselines, treatment, and follow-up sessions without any adverse events.

Table 1. Participant Demographics

Participant	Age (years)	Race[1]	Gender[2]	Education (years)	MPO[3]	Handedness[4]
1	50	AA; NA	M	14	24	R
2	53	C	M	18	18	L

Notes. [1]AA=African-American, NA=Native American, C=Caucasian; [2]M=Male, F=Female; [3]MPO=months post-onset; [4]Handedness R=right, L=left.

[3] One of our primary referral sources, Pitt+Me, recruited individuals based on an aphasia diagnosis generally entered into their medical record during their initial episode of care. Therefore, many of the individuals we recruited had mostly or completely resolved aphasia by the time of initial testing.

Table 2. Language assessments

	Comprehensive Aphasia Test T-Scores							CCT[1]
Participant	Comp. Spoken[2] *(cutoff 56)*	Comp. Written[3] *(cutoff 59)*	Rep[4] *(cutoff 59)*	Naming *(cutoff 62)*	Reading *(cutoff 57)*	Writing *(cutoff 57)*	Mean	Total Correct *(cutoff 56)*
1	50	50	32	54	49	46	46.8	52
2	38	43	48	48	49	48	45.7	50

Notes. [1]Camel and Cactus Test (CCT; Bozeat et al., 2000); [2]Comprehension of Spoken Language; [3]Comprehension of Written Language; [4] Repetition. The tests' cutoffs are shown in parenthesis (all participant scores are below the cutoffs).

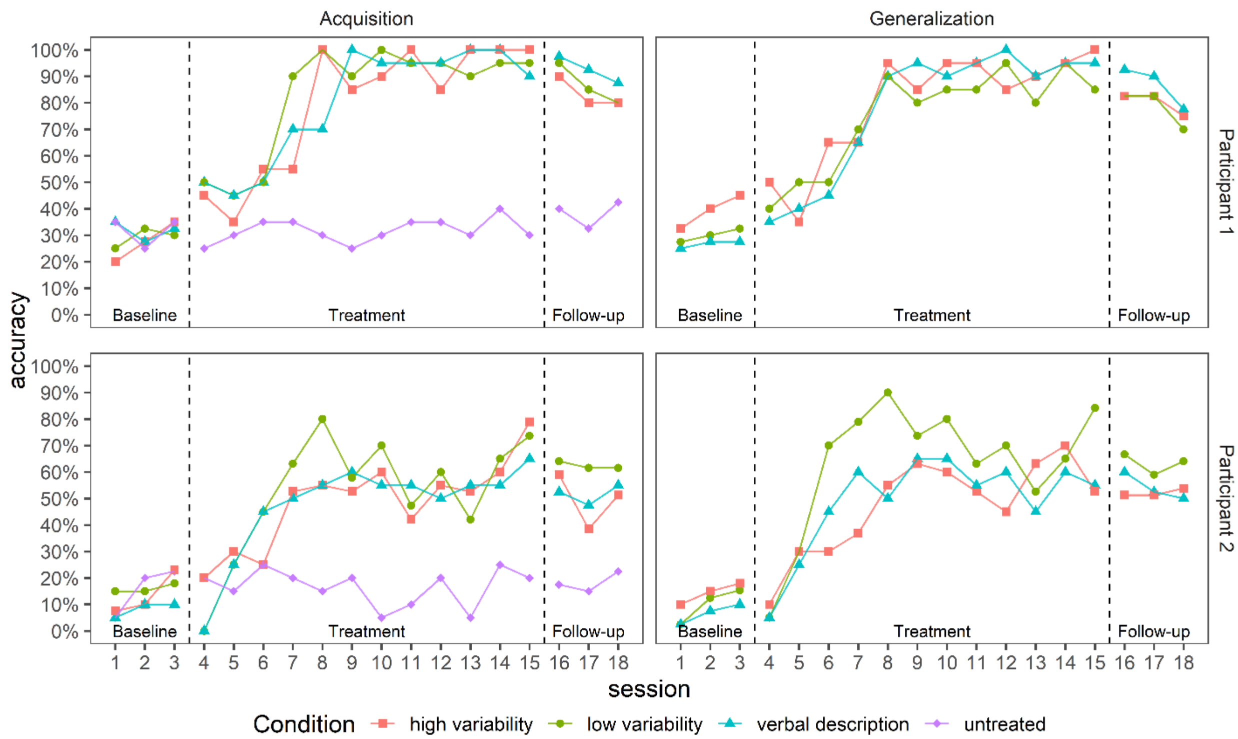

Figure 3. Raw probe data for participant 1 (top row) and participant 2 (bottom row). This figure depicts probe naming accuracy during the 3 baselines, 12 treatment sessions, and 3 follow-up sessions. The high exemplar variability condition is depicted in red and squares. The low exemplar variability condition is depicted in green and circles. The verbal description condition is depicted in purple and diamonds. The untreated control condition is depicted in light blue and triangles. Each condition included 40 items, for a total of 120 treated and 40 control items (total n =160 items per participant).

During the treatment phase, participant 1 spent a total of 5.4 hours practicing Anki during one-on-practice synchronous treatments sessions, and 8.6 hours practicing Anki independently, for a total of 14 hours of practice (as retrieved from Anki logs). Participant 2 spent a total of 7.5 hours practicing Anki during one-on-practice synchronous treatments sessions, and 16.4 hours practicing Anki independently, for a total of 23.9 hours of practice (as retrieved from Anki logs). These differences in total practice time were due to differences in average trial time and the total number of reviews (participant 2 took longer on average per trial and had lower average trial accuracy, leading to more scheduled reviews). Participant 1 completed 3844 practice trials, with an average of 230.4 trials per week. Participant 2 completed 4918 practice trials, with an average of 309.2 trials per week.

3.1 Prediction 1 results

In terms of treatment *acquisition*, effect sizes and credible intervals revealed that, out of 120 trained words, participant 1 acquired 77.24 direct training probes (i.e., trained pictures of trained words) across experimental conditions (90% CI = [72.48, 82.68]) and acquired 63.25 within-level generalization probes (i.e., untrained exemplars of trained words, 90% CI = [57.27, 69.06]). He showed no meaningful or statistically reliable improvements on untreated control probes, with credible intervals including zero words gained for this condition (effect size: 2.40 words, 90% CI = [-0.47, 5.17]). Effect sizes for participant 2 showed that he acquired 57.62 direct training probes across experimental conditions (90% CI = [50.94, 64.25]) and 48.06 within-level generalization probes (90% CI = [40.83, 55.37]). He showed no improvement on untreated control probes (effect size: -2.65 words, CI = [-5.96, 0.63]).

In terms of treatment *retention*, effect sizes revealed that participant 1 forgot 8.88 direct training probes (90% CI = [-14.23, -3.16]) at one-month post-treatment, and 12.94 direct training probes at three months post-treatment (90% CI: [-18.54, -7.10]) when compared to initial acquisition effect sizes. He forgot 1.60 within-level generalization probes at one month post-treatment (90% CI: [-8.27, 5.25]) and 14.49 at three months post-treatment (90% CI = [-21.64, -7.14]). Participant 2 forgot 10.44 direct treatment probes (90% CI: [-19.39, -1.62]) at one month post-treatment, and 2.67 direct treatment probes at three months post-treatment (90% CI: [-10.97 5.42]). He forgot 6.14 within-level generalization probes at one month post-treatment (90% CI: [-15.06, 3.11]) and 4.08 at three months post-treatment (90% CI: [-12.31, 4.19]). Model fixed effects support this interpretation of effect sizes, with robust effects for slope change and level change, indicating that naming gains can be attributed directly to the treatment (see full interrupted time series model results at).

There was no statistically reliable change in untrained control probes performance for participant 1 (2.40, 90% CI = [-0.47, 5.17]) or participant 2 (-2.65, 90% CI = [-5.96, 0.63]), suggesting no effect of repeated probe exposure.

3.2 Prediction 2 results

In most cases, the two-way interaction models testing for differences in stimuli variability on treatment response for acquisition, retention, direct training, and within-level generalization showed uncertain or robustly null results with small effect sizes. Table 3 reports estimates, credible intervals, and the ROPE for all two-way interaction effects (complete model results available at https://osf.io/sfutm). The only practical effect (i.e., exceeded ROPE) was found in the retention of direct training for participant 2, where words trained in the low variability condition were better retained at 1-month follow-up. However, this difference was smaller at the 3-month follow-up, where it did not reliably exceed the ROPE (See Figure 2). Taken together, these results indicate that stimuli variability did not lead to practically meaningful differences in treatment outcomes for participants 1 and 2.

Table 3. Interaction effects for prediction 2

ID	Models	Interaction term	Estimate	90% CI	ROPE	Interp.
1	Acquisition of direct training	timepoint (sessions 3 to 16) by stimuli variability (low vs. high)	-0.09	-0.33, 0.16	70.5%	uncertain
		timepoint (sessions 3 to 16) by stimuli variability (low vs. verbal)	0.02	-0.22, 0.27	77.2%	uncertain
	Acquisition of within-level generalization	timepoint (sessions 3 to 16) by stimuli variability (low vs. high)	-0.07	-0.33, 0.18	71.4%	uncertain
		timepoint (sessions 3 to 16) by stimuli variability (low vs. verbal)	0.29	0.02, 0.58	25.7%	uncertain
	Retention of direct training	timepoint (session 17) by stimuli variability (low vs. high)	-0.12	-2.28, 2.02	11.6%	uncertain
		timepoint (session 18) by stimuli variability (low vs. high)	0.56	-1.62, 2.89	10.7%	uncertain
		timepoint (session 17) by stimuli variability (low vs. verbal)	0.58	-1.94, 3.47	9.5%	uncertain
		timepoint (session 18) by stimuli variability (low vs. verbal)	0.22	-2.26, 2.95	9.9%	uncertain
	Retention of within-level generalization	timepoint (session 17) by stimuli variability (low vs. high)	-0.01	-1.92, 1.95	12.8%	uncertain
		timepoint (session 18) by stimuli variability (low vs. high)	0.74	-1.49, 3.25	10.2%	uncertain
		timepoint (session 17) by stimuli variability (low vs. verbal)	0.08	-2.11, 2.39	11.1%	uncertain
		timepoint (session 18) by stimuli variability (low vs. verbal)	-0.12	-2.50, 2.61	10.2%	uncertain
2	Acquisition of direct training	timepoint (sessions 3 to 16) by stimuli variability (low vs. high)	-0.03	-0.14, 0.07	98.7%	null effect
		timepoint (sessions 3 to 16) by stimuli variability (low vs. verbal)	-0.03	-0.13, 0.08	98.9%	null effect
	Acquisition of within-level generalization	timepoint (sessions 3 to 16) by stimuli variability (low vs. high)	-0.07	-0.18, 0.04	94.6%	uncertain
		timepoint (sessions 3 to 16) by stimuli variability (low vs. verbal)	-0.04	-0.15, 0.07	97.9%	null effect

Retention of direct training	timepoint (session 17) by stimuli variability (low vs. high)	-1.88	-4.60, 0.24	4.4%	practical effect
	timepoint (session 18) by stimuli variability (low vs. high)	-0.60	-3.28, 1.95	9.4%	uncertain
	timepoint (session 17) by stimuli variability (low vs. verbal)	-0.42	-2.66, 1.64	11.9%	uncertain
	timepoint (session 18) by stimuli variability (low vs. verbal)	0.23	-2.33, 2.81	10.3%	uncertain
Retention of within-level generalization	timepoint (session 17) by stimuli variability (low vs. high)	0.20	-2.22, 2.44	11.4%	uncertain
	timepoint (session 18) by stimuli variability (low vs. high)	0.02	-2.53, 2.52	10.7%	uncertain
	timepoint (session 17) by stimuli variability (low vs. verbal)	-0.26	-2.65, 1.99	11.9%	uncertain
	timepoint (session 18) by stimuli variability (low vs. verbal)	-1.01	-3.87, 1.57	8.2%	uncertain

Notes: ID = participant. Interaction term "timepoint" = includes the treatment and follow-up sessions (model sessions are shown in parenthesis, with sessions 3-16 reflecting the treatment phase, session 17 reflecting ~1-month retention, and session 18 reflecting ~3-month retention); Interaction term "stimuli variability" = refers to the conditions included in each interaction (low exemplar variability, high exemplar variability, verbal description). ROPE = Region of Practical Equivalence (this column represents percentages *inside* ROPE). Interpretation = If > 97.5% of the posterior distribution falls within the ROPE, the effect is considered null. If < 2.5% of the posterior distribution falls outside the ROPE, the effect is considered practically meaningful. ROPE percentages between these values are considered to provide uncertain evidence. Interp. = interpretation.

4. Discussion

While previous anomia treatment research has focused on training small sets of words (e.g., 47 words on average; Snell et al., 2010) in the hopes of inducing generalization (see examples of such studies in Boyle, 2010; Boyle, 2011), the current work tested a novel alternative: training a larger set of words, in a time-efficient manner, to engender long-term retention and within-level generalization. We did so by employing a retrieval-based anomia treatment using Anki open-source flashcard software, which incorporates adaptive distributed practice. We implemented this treatment using a clinically feasible self-managed treatment paradigm combining synchronous and asynchronous practice. We predicted that the intervention would lead to efficient acquisition, retention, and within-level generalization for more words than are typically targeted in anomia treatment research. Since generalization of trained exemplars to untrained exemplars (i.e., within-level generalization) is a crucial goal of anomia treatment, we also varied stimuli variability within participants, training sets of words using either a single picture exemplar, multiple picture exemplars or verbal description prompts. We predicted that increasing stimuli variability (i.e., more picture exemplars or verbal description) would improve within-level generalization for acquisition and retention. Research questions are discussed as follows.

4.1 Does adaptive distributed practice using Anki flashcard software lead to efficient acquisition, retention, and within-level generalization for more words than typically targeted during anomia treatment research?

Overall, direct training effects (measured on probe naming of trained picture exemplars for trained words) were large in both participants. Within-level generalization effects (measured on probe naming of untrained picture exemplars for trained words) were equally large. Participants 1 and 2 acquired an average of 67.43 trained picture exemplars when probed immediately after treatment and retained an average of 59.61 trained picture exemplars three months post-treatment (i.e., ~50% of the 120 words trained). In terms of within-level generalization, participants acquired an average of 55.65 untrained picture exemplars and retained 46.37 untrained exemplars three months post-treatment.

Compared to previous anomia treatment approaches, these effect sizes are large and suggest treatment efficiency. For example, Kendall et al. (2019) provide rough outcomes benchmarks for existing semantically- and phonologically-oriented anomia treatments. Specifically, Kendall et al. (2019) compared the effectiveness of Phonomotor Treatment (PMT, Kendall et al., 2008) and Semantic Feature Analysis (SFA, Boyle & Coelho, 1995). They administered 56-60 hours of PMT or SFA to 58 participants with aphasia in a between-group randomized control trial. At three months post-treatment, they found that SFA led to an average retention effect of 12.4 trained words above baseline (+15.6% of the 80 trained probes) and a generalization effect of two related words above baseline (+5%). PMT led to an average effect of 7 trained words above baseline (+18% of 39 trained probes) and a generalization retention effect of 1 word above baseline (+4.4%).

In contrast to the roughly 60 hours of treatment in Kendall et al. (2019), and as retrieved from Anki logs, participants in the current study completed an average of 6.5 hours of synchronous practice (and an average of 19 hours of total treatment, including independent home practice). This resulted in an average of 59.61 words retained at three months post-treatment, in contrast to the average of < 15 trained or related untrained words retained three months post-treatment by participants in Kendall et al. (2019). Notably, SFA and PMT seek to improve treatment efficiency via generalization to *all* closely related untrained words (not just those probed). Therefore, Kendall et al.'s effect sizes may under-estimate their participants' overall naming gains. However, our effect sizes and lower dosage suggest that direct training using adaptive distributed practice may be a feasible and treatment-efficient alternative to existing anomia treatment research, especially since it allows for the training of more words than typically targeted (e.g., 47 words on average; Snell et al., 2010).

Given the limited services available, considering treatment efficiency is important for clinical implementation. For example, stroke survivors with aphasia in Western Pennsylvania receive only a median of 7.5 hours of outpatient aphasia treatment in their first year post-stroke (Cavanaugh et al., 2021), which allows for the implementation of our current study dosage. Additionally, while this approach does require more stimuli to be created than most word-finding treatments, our recommendation to clinicians is to guide people with aphasia and their families on how to continuously add personally relevant targets to Anki as their treatment progresses. Thus, the burden is not on the clinician to create hundreds of stimuli in their limited time. Instead, the person with aphasia and their family are empowered to choose treatment targets that are more meaningful to them. This patient-centered approach has been successfully implemented in the clinic for an individual with primary progressive aphasia (Evans et al., 2016). This lays out the potential impact of direct training approaches using self-managed flashcard software that may be implemented within the current constraints of outpatient clinical practice. However, additional candidacy factors come into play for this type of clinical implementation (e.g., patient language ability, level of family support).

4.2 Does increasing stimuli variability lead to improved within-level generalization?

We did not find any practically meaningful or statistically robust differences in treatment outcomes when comparing stimuli variability conditions. This result has implications for clinical feasibility since it indicates that the retention and within-level generalization effects, observed across experimental conditions, may be achieved while training a single picture exemplar.

Within-level generalization to untrained picture exemplars also suggests that this approach trains more than stimulus-response mapping between specific words and pictures. Instead, it improved lexical access for trained words (e.g., by strengthening lexical-semantic representations that can be elicited by untrained visual input). While assessing anomia treatment outcomes using multiple picture exemplars of trained words has not been previously evaluated in post-to aphasia research (to our knowledge), these findings are consistent with current theories regarding the nature of post-stroke anomia, in contrast to semantic dementia/ semantic variant primary progressive aphasia (Jefferies & Lambon Ralph, 2006).

4.3 Clinical implications

Our findings suggest that training people with aphasia to practice using self-managed flashcard software independently is an efficient use of limited clinic time. While we treated participants synchronously for a total of 12 treatment sessions in our study, participants independently completed all treatment steps with a high level of fidelity after only four initial training sessions. Therefore, it may take relatively little effort to establish a successful Anki home practice program in the clinic, allowing practicing clinicians to shift to other treatment targets.

An additional advantage of implementing Anki open-source flashcard software in the clinic is that it is flexible, free, and widely accessible. It is fully customizable through settings and plugins, and has free versions available for both computers and Android devices. In addition, people with aphasia can be trained to make their own flashcards to ensure personal relevance, as demonstrated in a previous clinical case report (Evans et al., 2016). In addition, the software's multimedia flashcards may be customized to include pictures, text, audio, and video. Therefore, different treatment approaches could be implemented in future studies, such as verb-focused treatments, script training, writing treatments, etc.

This implementation of adaptive distributed practice provides a bottom-up approach to precision medicine for anomia treatment. People with aphasia adaptively receive the amount of practice needed to ensure good acquisition and retention of specific words, which varies depending on specific patient and word-level characteristics. We suggest that this may be a way to provide a performance-based item-level optimal dose, which would be a substantial treatment innovation. This approach, where people with aphasia practice independently, and well-learned words are scheduled less frequently over time, may also provide a long-term practice option that addresses the limited retention often observed following anomia treatment (Menahemi-Falkov et al., 2021).

One promising future direction would be to harness adaptive distributed practice to train more words than were targeted in the current study. We trained 120 words, primarily due to experimental design limitations (i.e., item development, probe selection, and probe administration time) and not due to limitations in how many words our participants could have practiced concurrently in a reasonable amount of time. Both participants showed good acquisition and retention over time, which led to progressively fewer scheduled reviews and minutes spent practicing per week. For example, in the final week of treatment, participant 1 completed all scheduled reviews in 19.23 minutes (6.4 minutes average per day practiced). Since Anki allows automatically rolling in the practice of new items over time (e.g., up to 5 new flashcards per day), he likely could have continued to add new words while retaining acquired words via continued occasional reviews. This approach of gradually adding new words to increase the total number of words trained while maintaining old ones via adaptive distributed practice reflects an innovation in anomia treatment research. Standard anomia treatments typically target an average of 47 words (Snell et al., 2010). However, adaptive distributed practice could efficiently target hundreds (or perhaps even thousands) of words. With appropriate within-level generalization, this would impact everyday communication, as 2.000-word families account for approximately 90% of day-to-day spoken English (Nation, 2006).

5 Study Limitations

There are four study limitations that should be considered. First, because Anki was not intended to be used with people with aphasia, it requires participants to self-rate their responses. This means that participants with poor error self-awareness are not ideal candidates for Anki as it currently stands. For example, participant 3 did not meet the second stage of enrollment criteria, as he was unable to learn to complete all treatment steps with fidelity and independence during initial training. This limitation is not unique to Anki, and is important for any self-managed computer-based treatment. Other challenges when using home-based practice via computers and tablets include lack of familiarity with technology (Kurland et al., 2014), motivation, cognition, expectations, family support (Chen & Bode, 2011) and the less dynamic and flexible nature of computer-based treatments relative to clinician-led practice. Second, while we examined within-level generalization to untrained picture exemplars, we did not assess across-level generalization to functional contexts (e.g., conversation). Future work should consider generalization more broadly to ensure that the treatment leads to gains in everyday communication. Third, this study does not compare adaptive vs. non-adaptive treatment; thus, the causal benefit of adaptivity is not being tested, which will require a future comparative effectiveness trial. Fourth, while these case studies are promising and demonstrate treatment feasibility, they require replication in well-powered group clinical trials to ensure that results are generalizable to the population of people with aphasia.

6 **Conclusions**

The current single case design suggests that self-managed computer-based flashcard software which incorporates adaptive distributed practice is an effective way to acquire and retain more words than are typically targeted during anomia treatment. Treatment resulted in within-level generalization across experimental conditions, indicating improved lexical access beyond what could be attributed to specific word-picture stimulus-response mapping. This approach has promise for precision medicine and training more words than were targeted here, but larger-scale clinical trials are required to replicate and extend these effects. Finally, this treatment relies on freely available open-source flashcard software and self-managed asynchronous telepractice (Cherney et al., 2011), making it feasible for real-world clinical implementation.

References

Adlam, A.-L. R., Patterson, K., Bozeat, S., & Hodges, J. R. (2010). The cambridge semantic memory test battery: Detection of semantic deficits in semantic dementia and alzheimer's disease. *Neurocase, 16*(3), 193-207. https://doi.org/10.1080/13554790903405693

Aguilar, J. M., Plante, E., & Sandoval, M. (2018). Exemplar variability facilitates retention of word learning by children with specific language impairment. *Lang Speech Hear Serv Sch, 49*(1), 72-84. https://doi.org/10.1044/2017_lshss-17-0031

Allen, L., Mehta, S., McClure, J. A., & Teasell, R. (2012). Therapeutic interventions for aphasia initiated more than six months post stroke: A review of the evidence. *Topics in Stroke Rehabilitation, 19*(6), 523-535. https://doi.org/10.1310/tsr1906-523

Altiner, C. (2019). Integrating a computer-based flashcard program into academic vocabulary learning. *Turkish Online Journal of Educational Technology-TOJET, 18*(1), 44-62.

Ankiweb. (2006). *What spaced repetition algorithm does anki use?* In https://faqs.ankiweb.net/what-spaced-repetition-algorithm.html?highlight=SuperMemo#what-spaced-repetition-algorithm-does-anki-use

Baek, E. K., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*(3-4), 590-606. https://doi.org/10.1080/09602011.2013.835740

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The english lexicon project. *Behavior Research Methods, 39*(3), 445-459. https://doi.org/10.3758/bf03193014

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes, 2*, 35-67.

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rehabilitation, 17*(6), 411-422. https://doi.org/10.1310/tsr1706-411

Boyle, M. (2011). Discourse treatment for word retrieval impairment in aphasia: The story so far. *Aphasiology, 25*(11), 1308-1326. https://doi.org/10.1080/02687038.2011.596185

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology, 4*(4), 94-98.

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia, 38*(9), 1207-1215. https://doi.org/10.1016/s0028-3932(00)00034-8

Brady, M. C., Kelly, H., Godwin, J., & Enderby, P. (2012). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*(5), CD000425. https://doi.org/10.1002/14651858.CD000425.pub3

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database Systematic Reviews, 1*(6), CD000425. 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

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (boss) phase ii: 930 new normative photos. *PloS One, 9*(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the subtlex-us word frequencies. *Behavior Research Methods, 44*(4), 991-997. https://doi.org/10.3758/s13428-012-0190-4

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *R Journal, 10*(1). https://doi.org/https://doi.org/10.32614/rj-2018-017

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Grantee Submission, 76*(1), 1-32.

Carragher, M., Sage, K., & Conroy, P. (2013). The effects of verb retrieval therapy for people with non-fluent aphasia: Evidence from assessment tasks and conversation. *Neuropsychological Rehabilitation, 23*(6), 846-887. https://doi.org/10.1080/09602011.2013.832335

Cavanaugh, R., Kravetz, C., Jarold, L., Quique, Y., Turner, R., & Evans, W. S. (2021). Is there a research-practice dosage gap in aphasia rehabilitation? *American Journal of Speech-Language Pathology, 30*(5), 2115-2129. https://doi.org/10.1044/2021_ajslp-20-00257

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380. https://doi.org/10.1037/0033-2909.132.3.354

Chen, C. C., & Bode, R. K. (2011). Factors influencing therapists' decision-making in the acceptance of new technology devices in stroke rehabilitation. *American Journal of Physical Medicine and Rehabilitation, 90*(5), 415-425. https://doi.org/10.1097/PHM.0b013e318214f5d8

Cherney, L. R. (2010). Oral reading for language in aphasia (orla): Evaluating the efficacy of computer-delivered therapy in chronic nonfluent aphasia. *Topics in Stroke Rehabilitation, 17*(6), 423-431. https://doi.org/10.1310/tsr1706-423

Cherney, L. R., Kaye, R. C., & Hitch, R. S. (2011). The best of both worlds: Combining synchronous and asynchronous telepractice in the treatment of aphasia. *Perspectives on Neurophysiology Neurogenic Speech Language Disorders, 21*(3), 83-93.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/https://doi.org/10.4324/9780203771587

Cruice, M., Worrall, L., & Hickson, L. (2006). Perspectives of quality of life by people with aphasia and their family: Suggestions for successful living. *Topics in Stroke Rehabilitation, 13*(1), 14-24. https://doi.org/10.1310/4jw5-7vg8-g6x3-1qvj

Dalemans, R., De Witte, L., Beurskens, A., Van Den Heuvel, W., & Wade, D. (2010). An investigation into the social participation of stroke survivors with aphasia. *Disability and Rehabilitation, 32*(20), 1678-1685. https://doi.org/10.3109/09638281003649938

Davis, L., & Copeland, K. (2006). Computer use in treatment of aphasia—a survey of practice patterns and opinions. *Contemporary Issues in Communication Sciences and Disorders, 33*, 138-146.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 53, pp. 63-147). Academic Press. https://doi.org/https://doi.org/10.1016/S0079-7421(10)53003-2

Duffy, J. R. (2005). *Motor speech disorders: Substrates, differential diagnosis, and management* (2nd ed.). Mosby.

Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Edmonds, L. A., Mammino, K., & Ojeda, J. (2014). Effect of verb network strengthening treatment (vnest) in persons with aphasia: Extension and replication of previous findings. *American Journal of Speech-Language Pathology, 23*(2), S312-S329. https://doi.org/10.1044/2014_AJSLP-13-0098

Edmonds, L. A., Nadeau, S. E., & Kiran, S. (2009). Effect of verb network strengthening treatment (vnest) on lexical retrieval of content words in sentences in persons with aphasia. *Aphasiology, 23*(3), 402-424. https://doi.org/10.1080/02687030802291339

Eglington, L. G., & Pavlik Jr, P. I. (2020). Optimizing practice scheduling requires quantitative tracking of individual item performance. *npj Science of Learning, 5*(1), 15. https://doi.org/10.1038/s41539-020-00074-4

Evans, W. S., Cavanaugh, R., Quique, Y., Boss, E., Starns, J. J., & Hula, W. D. (2021). Playing with bears: Balancing effort, accuracy, and response speed in a semantic feature verification anomia treatment game. *Journal of Speech, Language, and Hearing Research, 64*(8), 3100–3126. https://doi.org/doi:10.1044/2021_JSLHR-20-00543

Evans, W. S., Hula, W. D., & Starns, J. J. (2019). Speed–accuracy trade-offs and adaptation deficits in aphasia: Finding the “sweet spot” between overly cautious and incautious responding. *American Journal of Speech-Language Pathology, 28*(1S), 259-277. https://doi.org/10.1044/2018_AJSLP-17-0156

Evans, W. S., Quimby, M., Dickey, M. W., & Dickerson, B. C. (2016). Relearning and retaining personally-relevant words using computer-based flashcard software in primary progressive aphasia [Case Report]. *Frontiers in Human Neuroscience, 10*(561), 561. https://doi.org/10.3389/fnhum.2016.00561

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

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning* (pp. 9-46). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-188950-0.50008-4

Goodglass, H. (1980). Disorders of naming following brain injury. *American Scientist, 68*(6), 647-655. https://www.ncbi.nlm.nih.gov/pubmed/7224329

Gravier, M. L., Dickey, M. W., Hula, W. D., Evans, W. S., Owens, R. L., Winans-Mitrik, R. L., & Doyle, P. J. (2018). What matters in semantic feature analysis: Practice-related predictors of treatment response in aphasia. *American Journal of Speech-Language Pathology, 27*(1S), 438-453.

Harms, C., & Lakens, D. (2018). Making 'null effects' informative: Statistical techniques and inferential frameworks. *Journal of clinical and translational research, 3*(Suppl 2), 382-393. https://pubmed.ncbi.nlm.nih.gov/30873486

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6412612/

Holland, P. C. (2008). Cognitive versus stimulus-response theories of learning. *Learning & behavior, 36*(3), 227-241. https://doi.org/10.3758/lb.36.3.227

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Wiley.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38-58. https://doi.org/10.1177/00131640021970358

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain, 129*(8), 2132-2147. https://doi.org/10.1093/brain/awl153

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Johansson, M. B., Carlsson, M., & Sonnander, K. (2012). Communication difficulties and the use of communication strategies: From the perspective of individuals with aphasia. *International Journal of Language and Communication Disorders, 47*(2), 144-155. https://doi.org/https://doi.org/10.1111/j.1460-6984.2011.00089.x

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (Second ed., pp. 487-514). Elsevier Science.

Kendall, D. L., Moldestad, M. O., Allen, W., Torrence, J., & Nadeau, S. E. (2019). Phonomotor versus semantic feature analysis treatment for anomia in 58 persons with aphasia: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 62*(12), 4464-4482. https://doi.org/doi:10.1044/2019_JSLHR-L-18-0257

Kendall, D. L., Oelke, M., Brookshire, C. E., & Nadeau, S. E. (2015). The influence of phonomotor treatment on word retrieval abilities in 26 individuals with chronic aphasia: An open trial. *Journal of Speech, Language, and Hearing Research, 58*(3), 798-812. https://doi.org/10.1044/2015_JSLHR-L-14-0131

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Rothi, L. J. G., & Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language, 105*(1), 1-17. https://doi.org/10.1016/j.bandl.2007.11.007

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270-280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. https://doi.org/10.3758/s13423-016-1221-4

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods, 44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Kurland, J., Wilkins, A. R., & Stokes, P. (2014). Ipractice: Piloting the effectiveness of a tablet-based home practice program in aphasia treatment. *Seminars in Speech and Language, 35*(01), 051-064. https://doi.org/10.1055/s-0033-1362991

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). Bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software, 4*(40), 1541.

Mateer, C. A., & Sohlberg, M. M. (2003). Cognitive rehabilitation revisited. *Brain Impairment, 4*(1), 17-24.

Meltzer, J. A., Baird, A. J., Steele, R. D., & Harvey, S. J. (2018). Computer-based treatment of poststroke language disorders: A non-inferiority study of telerehabilitation compared to in-person service delivery. *Aphasiology, 32*(3), 290-311. https://doi.org/10.1080/02687038.2017.1355440

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/10.1080/09638288.2021.1955303

Middleton, E. L., Rawson, K. A., & Verkuilen, J. (2019). Retrieval practice and spacing effects in multi-session treatment of naming impairment in aphasia. *Cortex; a journal devoted to the study of the nervous system and behavior, 119*, 386-400. https://doi.org/10.1016/j.cortex.2019.07.003

Middleton, E. L., Schuchard, J., & Rawson, K. A. (2020). A review of the application of distributed practice principles to naming treatment in aphasia. *Topics in Language Disorders, 40*(1), 36-53. https://doi.org/10.1097/tld.0000000000000202

Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. *Journal of experimental psychology. Learning, memory, and cognition, 41*(4), 1253-1261. https://doi.org/10.1037/xlm0000091

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects.

Journal of Speech, Language, and Hearing Research, 59(5), 1111-1122. https://doi.org/10.1044/2016_jslhr-l-15-0303

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to snodgrass & vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS One, 7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Murre, J. M., & Dros, J. (2015). Replication and analysis of ebbinghaus' forgetting curve. *PloS One, 10*(7), e0120644. https://doi.org/10.1371/journal.pone.0120644

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian modern language review, 63*(1), 59-82.

Ostwald, S. K., Godwin, K. M., Cheong, H., & Cron, S. G. (2009). Predictors of resuming therapy within four weeks after discharge from inpatient rehabilitation. *Topics in Stroke Rehabilitation, 16*(1), 80-91. https://doi.org/10.1310/tsr1601-80

Palmer, R., Dimairo, M., Latimer, N., Cross, E., Brady, M., Enderby, P., Bowen, A., Julious, S., Harrison, M., Alshreef, A., Bradley, E., Bhadhuri, A., Chater, T., Hughes, H., Witts, H., Herbert, E., & Cooper, C. (2020). Computerised speech and language therapy or attention control added to usual care for people with long-term post-stroke aphasia: The big cactus three-arm rct. *Health technology assessment (Winchester, England), 24*(19), 1-176. https://doi.org/10.3310/hta24190

Palmer, R., Enderby, P., Cooper, C., Latimer, N., Julious, S., Paterson, G., Dimairo, M., Dixon, S., Mortley, J., & Hilton, R. (2012). Computer therapy compared with usual care for people with long-standing aphasia poststroke: A pilot randomized controlled trial. *Stroke, 43*(7), 1904-1911. https://doi.org/10.1161/STROKEAHA.112.650671

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally:Exemplar variability supports higher-order generalization and word learning. *Psychological Science, 21*(12), 1894-1902. https://doi.org/10.1177/0956797610389189

Quique, Y., 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*(1S), 230-246. https://doi.org/10.1044/2018_AJSLP-17-0155

R Core Team. (2020). *R: A language and environment for statistical computing.* In (Version 3.6.3) R Foundation for Statistical Computing. https://www.R-project.org

Ramsberger, G., & Marie, B. (2007). Self-administered cued naming therapy: A single-participant investigation of a computer-based therapy program replicated in four cases. *American Journal of Speech-Language Pathology, 16*(4), 343-358. https://doi.org/10.1044/1058-0360(2007/038)

Rana, T., Laoteppitaks, C., Zhang, G., Troutman, G., & Chandra, S. (2020). An investigation of anki flashcards as a study tool among first year medical students learning anatomy. *The FASEB Journal, 34*(S1), 1-1.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. https://doi.org/10.3758/PBR.16.2.225

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/10.1080/09602011.2010.528966

Schuchard, J., Rawson, K., & Middleton, E. (2020). Effects of distributed practice and criterion level on word retrieval in aphasia. *Cognition, 198*, 104216. https://doi.org/10.1016/j.cognition.2020.104216

Seibert Hanson, A. E., & Brown, C. M. (2020). Enhancing l2 learning through a mobile assisted spaced-repetition tool: An effective but bitter pill? *Computer Assisted Language Learning, 33*(1-2), 133-155.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers),

Simmons-Mackie, N., & Cherney, L. (2018). Aphasia in north america: Highlights of a white paper. *J Archives of Physical Medicine and Rehabilitation, 99*(10), e117.

Snell, C., Sage, K., & Lambon Ralph, M. A. (2010). How many words should we provide in anomia therapy? A meta-analysis and a case series study. *Aphasiology, 24*(9), 1064-1094. https://doi.org/10.1080/02687030903372632

Sohlberg, M. M., & Turkstra, L. S. (2011). *Optimizing cognitive rehabilitation: Effective instructional methods*. Guilford Press.

Spence, K. W. (1950). Cognitive versus stimulus-response theories of learning. *Psychological Review, 57*(3), 159-172. https://doi.org/10.1037/h0058250

Stan Development Team. (2020). *Rstan: The r interface to stan* In R package version 2.21.2. http://mc-stan.org/.

Stark, B. C., & Warburton, E. A. (2018). Improved language in chronic aphasia after self-delivered ipad speech therapy. *Neuropsychological Rehabilitation, 28*(5), 818-831. https://doi.org/10.1080/09602011.2016.1146150

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10*(2), 349-367. https://doi.org/10.1901/jaba.1977.10-349

Stokes, T. F., & Osnes, P. G. (2016). An operant pursuit of generalization - republished article. *Behavior Therapy, 47*(5), 720-732. https://doi.org/10.1016/j.beth.2016.08.012

Swan, A. J., Carper, M. M., & Kendall, P. C. (2016). In pursuit of generalization: An updated review. *Behavior Therapy, 47*(5), 733-746. https://doi.org/10.1016/j.beth.2015.11.006

Swiderski, A. M., Quique, Y. M., Dickey, M. W., & Hula, W. (2021). Treatment of underlying forms: A bayesian meta-analysis of the effects of treatment and person-related variables on treatment response. *Journal of Speech, Language, and Hearing Research, 64*(11), 4308-4328. https://doi.org/10.1044/2021_jslhr-21-00131

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive aphasia test*. Psychology Press.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schoelkopf, B., & Gomez-Rodriguez, M. (2017). Optimizing human learning. *arXiv preprint arXiv:1712.01856.*

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., & Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences, 116*(10), 3988. https://doi.org/10.1073/pnas.1815156116

Tate, R. L., & Perdices, M. (2019). Withdrawal/reversal designs. In *Single-case experimental designs for clinical research and neurorehabilitation settings: Planning, conduct, analysis and reporting* (pp. 90-107). Routledge. Taylor and Francis group.

Thompson, C. K. (1989). Generalization research in aphasia: A review of the literature. *Clinical aphasiology, 18*, 195-222.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i-109. https://doi.org/10.1037/h0092987

Toppino, T., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. *Psychology of Learning and Motivation, 60*, 113-189. https://doi.org/10.1016/B978-0-12-800090-8.00004-4

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. *Inf. Child. Dev, 23*(2), 105-122. https://doi.org/doi:10.1002/icd.1824

Webster, J., Whitworth, A., & Morris, J. (2015). Is it time to stop "fishing"? A review of generalisation following aphasia intervention. *Aphasiology, 29*(11), 1240-1264. https://doi.org/10.1080/02687038.2015.1027169

Wenke, R., Lawrie, M., Hobson, T., Comben, W., Romano, M., Ward, E., & Cardell, E. (2014). Feasibility and cost analysis of implementing high intensity aphasia clinics within a sub-acute setting. *International Journal of Speech-Language Pathology, 16*(3), 250-259. https://doi.org/10.3109/17549507.2014.887777

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://www.ncbi.nlm.nih.gov/pmc/articles/PMC7531584/pdf/nihms-1506946.pdf

Wisenburn, B., & Mahoney, K. (2009). A meta-analysis of word-finding treatments for aphasia. *Aphasiology, 23*(11), 1338-1352. https://doi.org/10.1080/02687030902732745