Playing with BEARS: Balancing Effort, Accuracy, and Response Speed in a Semantic Feature Verification Anomia Treatment Game.

William S. Evans,[1]* Robert Cavanaugh,[1,2] Yina Quique,[3] Emily Boss,[4] Jeffrey J. Starns,[5] and William D. Hula[2,1]

[1]Department of Communication Sciences and Disorders, University of Pittsburgh, Pittsburgh, PA, USA.

[2]Geriatric Research Education and Clinical Center, VA Healthcare System, Pittsburgh, PA, USA.

[3]Center for Education in Health Sciences, Northwestern University, Chicago, IL, USA.

[4]Integrative Reconnective Aphasia Therapy, Pittsburgh, PA, USA.

[5]Department of Psychology, University of Massachusetts, Amherst, MA, USA.

*Corresponding author: will.evans@pitt.edu

Abstract:

Purpose: The purpose of this study was to develop and pilot a novel treatment framework called *BEARS* (Balancing Effort, Accuracy, and Response Speed). People with aphasia (PWA) have been shown to maladaptively balance speed and accuracy during language tasks. BEARS is designed to train PWA to balance speed-accuracy tradeoffs and improve system calibration (i.e., to adaptively match system use with its current capability), which was hypothesized to improve treatment outcomes by maximizing retrieval practice and minimizing error learning. In this study, BEARS was applied in the context of a semantically-oriented anomia treatment based on semantic feature verification (SFV).

Methods: Nine PWA received 25 hours of treatment in a multiple baseline single-case series design. BEARS + SFV combined computer-based SFV with clinician-provided BEARS meta-cognitive training. Naming probe accuracy, efficiency, and proportion of “pass” responses on inaccurate trials were analyzed using Bayesian generalized linear mixed-effect models. Generalization to discourse and correlations between practice efficiency and treatment outcomes were also assessed.

Results: Participants improved on naming probe accuracy and efficiency of treated and untreated items, although untreated item gains could not be distinguished from the effects of repeated exposure. There were no improvements on discourse performance, but participants demonstrated improved system calibration based on their performance on inaccurate treatment trials, with an increasing proportion of “pass” responses compared to paraphasia or timeout nonresponses. In addition, levels of practice efficiency during treatment were positively correlated with treatment outcomes, suggesting that improved practice efficiency promoted greater treatment generalization and improved naming efficiency.

Conclusions: BEARS is a promising, theoretically-motivated treatment framework for addressing the interplay between effort, accuracy, and processing speed in aphasia. This study establishes the feasibility of BEARS + SFV and provides preliminary evidence for its efficacy. This study highlights the importance of considering processing efficiency in anomia treatment, in addition to performance accuracy.

Introduction

Aphasia is a language disorder caused by stroke and other acquired brain injuries that affects roughly one-third of stroke survivors and more than 2 million people in the United States (Simmons-Mackie, 2018). Anomia, the inability to successfully retrieve and produce words, is a cardinal feature of aphasia (Goodglass, 1980) and experienced to some degree by all people with aphasia (PWA). Therefore, it is important to continue to improve anomia treatment outcomes, and the current work attempts to contribute to this endeavor. The current study piloted a novel game-based intervention which combined an established semantically-oriented anomia treatment (Semantic Feature Verification; Kiran & Roberts, 2010) with feedback and clinician-provided "system calibration training" (described below), designed to help PWA to balance speed, accuracy, and effort during word retrieval.

In the sections to follow, we will explain our conceptualization of system calibration and adaptation deficits in aphasia and explain how they apply to speed-accuracy tradeoffs and retrieval effort in anomia rehabilitation and functional communication. This will in turn motivate our novel system calibration training framework, BEARS ("Balancing Effort, Accuracy, and Response Speed"). The introduction will conclude with goals and study predictions for the current pilot.

System calibration and adaptation deficits in aphasia

In their classic work on Adaptation Theory, Heeschen and colleagues (Heeschen & Schegloff, 1999; Kolk & Heeschen, 1990; Kolk & Heeschen, 1992) argued that a distinction should be made between aphasia symptoms caused by underlying impairments ('impairment symptoms') and those caused by an individual's response to those impairments ('adaptation symptoms'). Their key evidence came from observations regarding the nature of 'telegraphic speech' in Broca's aphasia, where individuals who typically produced single key content words at a slow rate in spontaneous speech instead produced lengthier paragrammatic output when directed to do so in more constrained contexts such as a sentence elicitation task (Kolk & Heeschen, 1992). In contrast, individuals with Wernicke's aphasia display task insensitivity, consistently produced paragrammatic output regardless of task. As a result, they argued that two key symptoms of the Broca's aphasia, slowed speaking rate and telegraphic speech, were adaptation symptoms, reflecting a strategic adaptation to an underlying grammatical output impairment. Their key claim, that a distinction should be made between an individual's core linguistic impairments and their strategic response to these impairments, applies beyond the context of understanding classic aphasia syndromes and has wide-ranging implications.

Building upon Adaptation Theory, we have argued that language performance and communication success in aphasia are determined by a combination of impairment and adaptation factors, with ultimate performance based on how well an individual makes use of their language system in its current state (Evans et al., 2019). PWA who do not respond well to their language system changes may demonstrate *adaptation deficits*, poorer-than-necessary language performance exacerbated by the use of maladaptive strategies and habitual responses (e.g., an over-reliance on ineffectual self-cuing approaches, consistently struggling to retrieve difficult words with considerable effort and frustration in contexts where this rarely results in success).

PWA who make effective use of their current core language capability demonstrate good system calibration, calibrating the demands they make of their system to its current capabilities in ways that are most likely to result in success. One analogy we have used with PWA to describe this concept is driving a car with a manual transmission. A good driver makes best use of the transmission and engine in its current condition (e.g., knows how to work with a worn-out clutch). However, someone unfamiliar with the car or a new driver may not know how to apply the necessary finesse, and thus may experience unnecessary issues such as stalling the engine or grinding the clutch. Adaptation deficits in aphasia consist of behaviors such as making repeated inaccurate retrieval attempts with increasing frustration instead of moving on or switching to an alternative communication strategy. A schematic for understanding the relationship between system capability, use, and calibration can be seen in Figure 1.

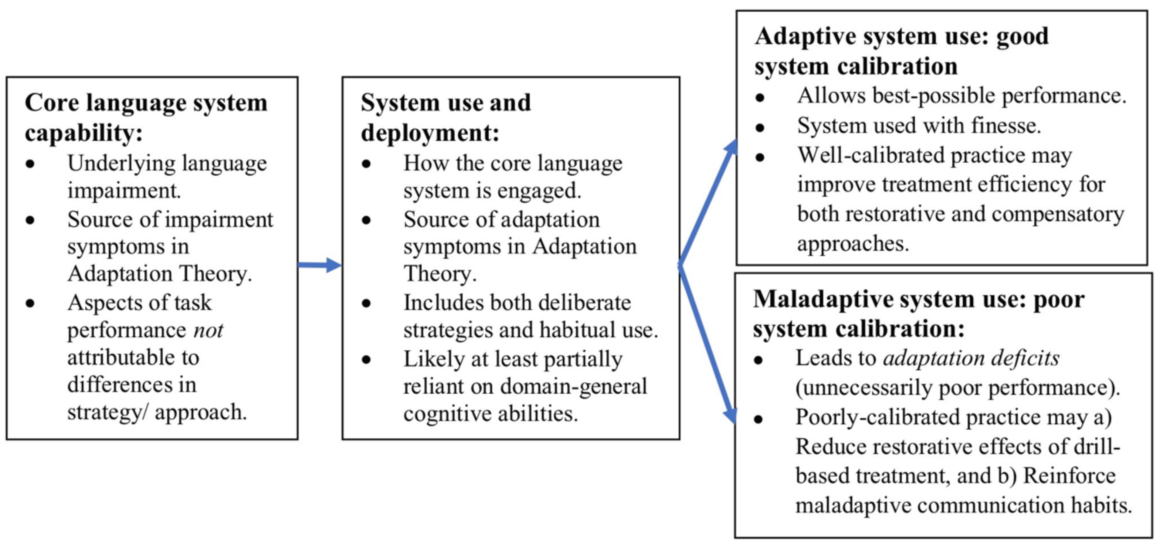

Figure 1. Schematic representing how core system capability (impairment) and its deployment (adaptation) together determine overall language task and communication performance, leading to *adaptive* or *maladaptive system calibration*. The BEARS training framework is intended to improve system calibration specifically as they relate to the level of effort and speed-accuracy tradeoffs during language performance.

The concept of flexible adaptation to current capability has often been associated with compensatory treatments and augmentative alternative communication (Hunt et al., 2002). In the current work, we wish to expand on this idea to propose that adaptive system calibration also makes best use of the original intended modality (e.g., successfully producing a difficult word after taking a breath to relax instead of needing to shift to an alternative communication strategy). We also propose that in drill-based treatment tasks, adaptive system calibration can be defined as engaging the language system in ways that improve treatment outcomes. Broadly construed, adaptive system calibration is about PWA making best-possible use of their current system to maximize language performance during treatments or functional communication activities.

A benefit of this conceptualization of language performance in aphasia is that treatment can target PWA's underlying language impairments (e.g., strengthen specific retrieval mechanisms), strategic and habitual response to these impairments, or both. For instance, recent work has demonstrated that a brief 5-session mindfulness meditation intervention for PWA temporarily improves verbal fluency (Marshall et al., 2018), even such training is unlikely to modify core linguistic capabilities in such a short time period. One explanation is that this intervention could help PWA make more adaptive use of their current language system by reducing maladaptive responses to language impairments related to extralinguistic factors such as "linguistic anxiety" (Cahana-Amitay et al., 2011).

A key consideration when seeking to address system calibration is that some aspects of language use may be more malleable and open to adaptive deployment than others. Speed-accuracy tradeoffs have shown potential for malleability and evidence for adaptation deficits in aphasia and are therefore worth pursuing from a rehabilitation perspective.

System calibration, processing speed, and speed-accuracy tradeoffs in anomia

Anomia is a word retrieval deficit measurable both in terms of accuracy and processing speed (Moineau et al., 2005). Improving word retrieval accuracy has been the focus of most anomia work to date (e.g., Best et al., 2013; Fridriksson et al., 2005), although speed has also been considered in a number of instances (e.g., Neto & Santos, 2012; Prather et al., 1997). However, the interactive relationship between speed and accuracy has not been adequately considered in anomia treatment. In speed-accuracy tradeoffs, spending more time on a task tends to increase accuracy, while spending less time lowers accuracy. Speed-accuracy tradeoffs are a robust and widespread phenomenon in both human psychology (e.g., Wickelgren, 1977) and beyond (Ceccarini et al., 2020). In humans, speed-accuracy tradeoffs appear to be partially under volitional control: individuals are able to flexibly adjust speed vs. accuracy in the context of shifting task instructions, feedback, or rewards that prioritize speed or accuracy (Campanella et al., 2016; Starns & Ratcliff, 2010; Touron et al., 2007; Wagenmakers et al., 2008). Critically, speed-accuracy tradeoffs are often nonlinear (Starns & Ratcliff, 2010), such that overly cautious responses may be much slower but provide only marginal gains in accuracy, while overly impulsive responses may be faster but result in much lower accuracy performance (Figure 2). In previous response time modeling work, we have shown that speed-accuracy tradeoffs are present in PWA during lexical decision and picture naming tasks. In Evans et al. (2019), we applied the Diffusion Model (Ratcliff, 1978) to lexical decision data from 20 PWA, and found that 40% demonstrated adaptation deficits in speed-accuracy tradeoffs, with impaired speed or accuracy performance attributable to overly impulsive or overly cautious responses. In subsequent work (Evans et al., 2020), we developed a novel multinomial ex-gaussian response time model of picture naming in aphasia to estimate an “optimal response time cutoff,” the point at which additional processing time was unlikely to produce additional gains in accuracy. We fit this model to picture naming data from PWA, and found that for 8/10 participants, their average response time (RT) for incorrect responses exceeded their own optimal RT cutoff. Together, these results suggest that PWA do not always set speed-accuracy tradeoffs to optimize task performance in language-dependent tasks.

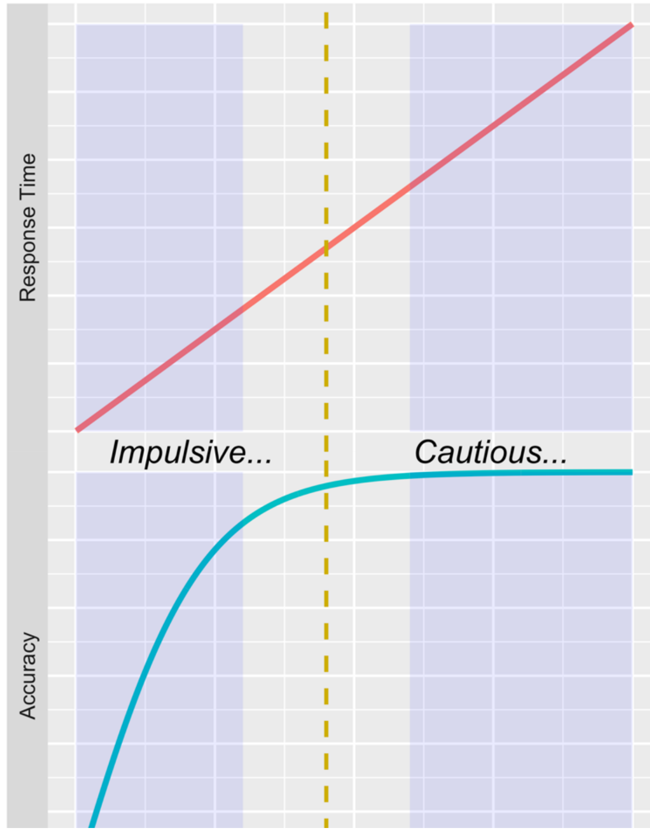

Figure 2. Schematic for speed-accuracy tradeoffs. Overly cautious responses slow response time (RT, red line) without further increasing accuracy (blue line). Overly impulsive responses improve RT, but with considerable lower levels of accuracy. Good system calibration (yellow vertical bar) balances these extremes and improves overall performance efficiency. Figure modified from Evans et al. (2019).

If present, maladaptive speed-accuracy tradeoffs likely have negative consequences for everyday communication and for treatment outcomes. In everyday communication, impulsive responses increase the chances of making fast retrieval errors, self-corrections, and conversation repairs. On the other hand, responses that are too cautious maximize accuracy in everyday communication at the cost of timely transfer of information, slowed processing, and may make online communication processes more susceptible to competition from internal or external distractions (e.g., PWA forgetting their idea before they can finish sharing it).

In treatment, maladaptive speed-accuracy tradeoffs may have specific negative consequences for dosage. Previous literature, particularly in the anomia context, has found that dose-form typically consists of the number of successful/accurate retrieval events (Harvey et al., 2020). Therefore, overly cautious responses would decrease dosage within a given treatment time by decreasing the number of trials while only providing negligible additional gains in trial accuracy. On the other hand, overly impulsive responses would increase error rates and interference effects from error learning (Fillingham et al., 2006) without appreciably increasing treatment dosage. As a result, PWA who display good system calibration may respond more optimally during drill-based treatment – not too quickly, but willing to move on after their chances of providing a correct response diminish – and as a result, maximize their treatment dosage. Recent evidence within usual care from the VERSE trial suggests that treatment is most successful when it is effortful while minimizing errors (but is not errorless), which provides empirical evidence for this claim (Brogan et al., 2020). Thus, maladaptive speed-accuracy tradeoffs in word retrieval likely have significant negative consequences for both everyday communication effectiveness for PWA and negatively affect drill-based treatment dosage. In the following section, we propose a general treatment framework intended to improve system calibration as it relates to effort and speed-accuracy tradeoffs in the context of word retrieval.

BEARS: A treatment framework to address strategic responses to core processing abilities

Following the findings that PWA set maladaptive speed-accuracy tradeoffs and that language performance in general may be affected by extralinguistic factors such as pressure, frustration, or anxiety, we conceptualized an aphasia treatment framework called *BEARS* (Balancing Effort, Accuracy, and Response Speed) intended to address these factors holistically. The framework is intended to address PWA's underlying linguistic deficits through more effective practice as well as increase strategic adaptation to underlying linguistic deficits through increased awareness and training. Our claim underlying the BEARS framework is straightforward: for an intervention to maximize its treatment dosage and its impact on everyday communication, it should strive to strike a balance between processing effort, performance accuracy, and response speed. Following the effortful retrieval practice literature (e.g., Middleton et al., 2016) and findings that patient-generated responses are likely to compose key active ingredients in treatment protocols (Evans et al., 2020; Gravier et al., 2018), responses within treatment protocols should be effortful, but only up until the point where effort becomes counterproductive.

Additionally, treatment should not only focus on constraining responses to this optimal calibration point where effort, speed, and accuracy are balanced, but should provide explicit metacognitive training and feedback to the person with aphasia, so that they are able to identify this calibration point on their own without the need for input from a clinician. Instruction and feedback can be implemented in terms of education on the relationship between these components, metacognitive training including self-monitoring of frustration or tension which reduce performance, and the successful use of strategies reduce these feelings which are detrimental to successful performance. PWA should be encouraged to balance speed and accuracy. Individuals who tend to make impulsive errors should be taught to slow down, while individuals who tend to persevere and who are stuck for longer than is therapeutically ideal should be taught to let unsuccessful attempts go and move on. Individuals who make both fast and slow types of errors should be taught to notice and respond to both (see Appendix 1 for a detailed description of the BEARS metacognitive training provided in the current study). Given its nature, the BEARS treatment framework could be used to augment most evidence-based restorative aphasia treatments targeting linguistic impairment, so that treatment not only addresses the underlying impairment, but also a person with aphasia's strategic response to these impairments. In the current study, we have applied the BEARS framework to one such treatment, the semantic-feature verification (SFV; Kiran & Roberts, 2010).

Semantic Feature Verification anomia treatment + BEARS

Semantically-oriented anomia treatments such as Semantic Feature Analysis (SFA; Boyle, 2004; Boyle, 2010; Coelho et al., 2000) are among the most well-studied treatments for naming impairment in aphasia. In SFA, the clinician shows the person with aphasia a pictured object and elicits a naming attempt. In the traditional version of SFA (Boyle, 2010; Massaro & Tompkins, 1994), the clinician then guides the PWA in verbally generating semantic features for the target, using a chart specifying feature categories (Boyle & Coelho, 1995). Correct naming of the target is elicited at the end of each trial. SFA has been modified in many ways since the original papers, for example, to focus on verbs and actions (Wambaugh & Ferguson, 2007) or implemented within the context of discourse (Peach & Reuter, 2010). Kiran and Roberts (2010) developed a variant of SFA in which repeated, guided practice centers around the verification of semantic features for target words rather than their verbal production. As in generation-based SFA, the verification-based variant (SFV) is hypothesized to improve retrieval of both treated and untreated semantically-related words by strengthening the activation of related concepts in the lexicon (e.g., Collins & Loftus, 1975). SFA has been found to improve treated words for almost all participants and semantically related, untreated words for a large proportion of PWA (Efstratiadou et al., 2018, Oh et al., 2016, Quique et al., 2018), including for the SFV variant (e.g., Gilmore et al., 2020).

The current study employed a computer-based version of SFV as we were interested evaluating the BEARS system calibration training in the context of an established anomia treatment and the two-choice nature of SFV is well-suited for computer-based implementation and feedback. The resulting BEARS + SFV treatment protocol included structured naming practice and feature verification as well as education on speed-accuracy tradeoffs, metacognitive training focused on the self-monitoring of effort, frustration, and timeliness of responses, and computer-based performance feedback on the efficiency of both naming and feature verification responses using a game-based points system (see Methods section). Thus, BEARS + SFA is not only intended to strengthen the production of target words and underlying semantic networks through more efficient drill-based practice, but also intended to improve participants' adaptive system calibration, learning to make more adaptive use of their core language system during both picture naming and feature-verification.

Study purpose:

The purpose of this study is to develop and pilot a BEARS-augmented anomia treatment (BEARS + SFV) using a multiple baseline single-case series experimental design. Its goals are to a) establish the feasibility of this approach, b) replicate previous SFV findings on performance accuracy and determine whether BEARS + SFV improves naming and discourse production efficiency, c) assess whether BEARS training improves how PWA respond in instances where they cannot produce a target word, and d) explore relationships between overall practice efficiency and treatment outcomes. Positive findings will support further research developing this intervention, which could establish comparative effectiveness of BEARS-augmented compared to standard interventions.

Study predictions:

1. *BEARS + SFV will replicate previous SFA/SFV findings and improve naming accuracy for both treated and semantically-related untreated words.*
2. *BEARS + SFV will increase naming efficiency.* By improving lexical access and system calibration, BEARS + SFV will improve the efficient retrieval of trained and untrained words, as measured in the number of correct words per minute.
3. *BEARS+SFV will improve discourse informativeness and efficiency.* This would indicate that BEARS system calibration training generalizes beyond the single-word level where it was trained.
4. *BEARS + SFV will improve system calibration for self-monitoring and error awareness.* We predicted BEARS training would lead to a shift in the nature of how participants responded on incorrect trials over time, producing a higher proportion of "pass" responses and a corresponding reduction in overt errors (paraphasias) and timeout nonresponses.
5. *Efficient practice performance during BEARS + SFV treatment will be positively associated with good treatment outcomes.* While the current study cannot distinguish correlation from causation, it is important to explore relationships between system calibration, practice efficiency, and treatment outcomes to determine whether further development of this work is warranted. We predicted that more efficient practice, would be associated with larger treatment effect sizes.

Methods

Participants

Participants were recruited from the Western Pennsylvania Research Registry, the Audiology and Speech Pathology Research Registry maintained by the VA Pittsburgh Healthcare System (VAPHS), and local clinician referral. No participants enrolled in this study received any concurrent speech-language treatment outside of the study-related sessions for the duration of the study.

To be included in the study, participants were required to be at least 6 months post-onset of stroke, have a diagnosis of aphasia (as defined by impairments in 2/8 subtest of the Comprehensive Aphasia Test), be community-dwelling and at least wheel-chair ambulatory, have spoken English as their primary language since childhood, and be age 18 or older. Participants were also required to demonstrate less than or equal to 50% correct performance on at least 80 treatment item probes during the pre-treatment study phase. Potential participants were excluded if they had a history of neurodegenerative disease, active, unmanaged psychopathology or alcohol/substance abuse, severe motor speech disorder (i.e., apraxia of speech or dysarthria) or were participating in any other speech/language therapy during the time of the study. Based on the complex multi-step nature of the treatment and our previous clinical trial experience evaluating semantically-oriented anomia treatment (e.g., Evans et al., 2020), we excluded participants who presented with very severe anomia, as measured by a CAT Naming modality T-score of less than 40. Data collection took place at the VA Pittsburgh Healthcare System with IRB approval (Study ID: Pro00002040).

Assessment

Participants were tested with standardized measures at study onset and again post-treatment. They were assessed with the Comprehensive Aphasia Test (CAT; Swinburn et al., 2004), the Philadelphia Naming Test (PNT; Roach et al., 1996), Cactus and Camel Test (CCT; Bozeat et al., 2000), and selected subtests of the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA; Kay et al., 1996). Motor speech was assessed via the Duffy protocol (Duffy, 2020) with diagnosis determined via consensus expert opinion between the first and fourth authors who are certified SLPs.

Changes in monologue-based discourse informativeness and efficiency were evaluated through the Nicholas and Brookshire protocol (Nicholas & Brookshire, 1993), which includes two sets of discourse stimuli, each with two picture descriptions, one narrative, one procedural, and one personal story. Discourse informativeness was measured by calculating the proportion of correct information units (CIUs; effectively, words that are both accurate and relevant) to total words. Efficiency was measured by the number of CIUs produced during the cumulative time taken for each narrative task within (CIUs/minute). Calculation of CIUs, words, and time followed the protocol described by Nicholas and Brookshire (1993). Informativeness and efficiency is known to be relatively equivalent between sets and reasonably stable between administrations (Nicholas & Brookshire, 1993). Sets were ordered pseudo-randomly. Samples were scored by the treating, certified speech-language pathologist (fourth author) who is well-trained in scoring CIUs but was not blinded to timepoint.

Stimuli matching and selection

Treatment and probe stimuli for study participants consisted of picturable nouns from two freely available photographic databases (Brodeur et al., 2010; Brodeur et al., 2014; Moreno-Martínez & Montoro, 2012). For each stimulus, we collected linguistic characteristics from available corpora (Balota et al., 2007; Brysbaert et al., 2012) consisting of lexical frequency, number of phonemes, and age of acquisition (Kuperman et al., 2012). Potential trained and untrained items were then matched for production complexity based on an item complexity algorithm from Fergadiotis et al. (2015). In this approach, item complexity was estimated using the following equation: B = -1.22 -.36(log word frequency) + .21(Age of Acquisition) + .15(number of phonemes), which they reported to account for 63% of variance in naming difficulty. This complexity score was used to create difficulty-matched triplets of items: a trained item, an untrained related item from the same category, and an untrained unrelated item from a different semantic category. Item triplets that had item complexity difference scores above 2 standard deviations were removed. The final set of stimuli consisted of 224 item triplets with potential trained items across 15 semantic categories: body parts, building, clothing, decoration, electronics, food, fruits and vegetables, furniture, kitchen utensils, mammals, nature, outdoor activities, stationary, tools, and vehicles.

For each participant, treatment lists were generated on the basis of performance on a confrontation picture-naming task. Pictures of the 224 potential treatment targets were presented one at a time and participants were given 15 seconds to name each picture. Accuracy was judged based on the ‘first complete response’ as per the scoring rules on the PNT. Per these rules, self-corrections were not accepted if they already made a complete response as indicated by pausing and/or prosody. If a participant indicated they did not know what the picture was, the item was marked as incorrect but not selected as a treatment target. Each participant completed the naming task on two separate occasions. Items that were named with less than or equal to 50% accuracy across both administrations were included as potential treatment items. To qualify for treatment, a category had to have at least 8 qualifying items. A total of 5 categories with 8 qualifying items in each category were selected for treatment for a total of 40 treatment targets per participant. Selection took into account participants personal interests and the quality of stimuli pictures. Since each of the treatment targets had difficulty-matched related and unrelated generalization items, administering all items would have created probe lists of 120 items. To reduce the considerable testing burden, only one generalization item was randomly selected for each treatment target, leading to a probe list of 80 items (40 treated words, 20 related untreated words, and 20 unrelated untreated words).

For the semantic feature verification portion of the treatment task, eight semantic "yes/no" feature questions for each of the 224 potential treatment targets were created by undergraduate lab volunteers and were rated by 3 independent raters on a 1-10 scale (with "10" being a good question and "1" being a poor question). Questions with an average score below eight were re-written and rescored by three additional independent raters. For each treatment target, four questions had a "yes" response and four had a "no" response. All questions were audio-recorded and edited using Audacity software.

Probe Administration

Probes were administered for each participant in a multiple baseline across participants design. Baseline probe performance was established via multiple probe assessments where each participant was randomly assigned a number of baseline probe sessions (3, 4, or 5 sessions), which helped control for the direct effects of probe exposure in the absence of treatment. Treated and untreated items were assessed at the beginning of each treatment session prior to initiating treatment, and within 1 week of finishing treatment. Items were also probed in a single follow-up session approximately 3 weeks after the completion of treatment.

Probe administration during baseline, treatment, and follow-up included both confrontation picture naming and a written lexical decision task, with item presentation randomized within each task. For naming probes, participants were given 15 seconds to name each picture and accuracy was judged based on the ‘first complete response’ as per stimuli selection. The written lexical decision task was 320 trials in length, presenting all 80 probe words and 80 matched pseudowords twice, and was always presented after naming probes each session. Written lexical decision probes were collected for secondary response time modeling analyses and are therefore not reported here.

Both probes and treatment software were programmed in PsychoPy software (Peirce et al., 2019) and administered on a Dell XPS13 laptop using a USB microphone headset. Assessment audio recordings were collected on a Surface Pro laptop using an external USB microphone.

Measuring naming response times

Given the focus placed on speed-accuracy tradeoffs and efficiency in the current work, trial response times during naming probes and treatment were collected via software voice key and clinician button press (marked immediately after the first complete response had been provided). Voice key responses were used to provide online computer-based points feedback during treatment (see below), and therefore participants were trained to produce a single verbal response and were reminded of these instructions each session. Voice key sensitivity was scored online by the treating clinician, with apparent false or failed triggers noted by a key press. All naming responses were also audio recorded for later review to establish rater reliability.

To assess voice key accuracy and reliability on naming probes, a trained independent third rater hand-coded 10% of trials which the clinician had marked were triggered appropriately. To do this, they viewed the recording waveforms in Audacity software and measured the distance between stimuli onset and the first complete response by hand. Trials with large disagreement between voice key and hand-coding were reviewed by the study team, and more than 90% of these trials were due to ambiguity in PNT scoring rules for determining the first correct response. Reliability between the voice key and hand-coded response times was good for 6 of the 9 participants (r's ranging between .92 and .99), with poorer reliability for the remaining 3 participants (r = .83 for participant 1, r = .24 for participant 4, and r = .69 for participant 9).

However, after assessing the reliability of trials with 'good' voice key triggers above, we determined that naming trials with failed voice key triggers appeared to exclude data not at random for some participants, as incorrect naming attempts were much more likely to be marked as inaccurate voice key response times due to early partial production attempts. As a result, we chose to use the clinician button-press measure of total trial time (from stimulus onset to immediately after participants gave their first complete response) to inform measure of reward rate used as a dependent variable in our efficiency analyses (see below).

Treatment procedures

Participants each received 25 hours of treatment administered by a licensed speech-language pathologist (mostly by the 4th author[1]). Sessions were typically scheduled 3-4 days per week with 2-3 hours of treatment time per day interspersed by breaks. Participants received *BEARS+ SFV*, a hybrid clinician-computer treatment with 2 core components: computer-based semantic feature verification treatment with points feedback, and BEARS meta-cognitive system calibration training from a clinician. While being introduced to the treatment task, each participant was educated on the speed-accuracy tradeoff and how to appropriately balance effort, accuracy, and response speed. Participants were encouraged to find the balance and the “right speed” for their processing ability in that moment. They were taught to become more aware of instances when they are very unlikely to produce a correct response, and instructed to say “pass” instead of producing an overt error or waiting until the response deadline had run out. This was intended to reduce error learning and to increase the number of successful completed trials during treatment. Additional details regarding BEAR meta-cognitive strategy training and how it was individualized for each participant are described in Appendix 1.

The treatment game was an implementation of SFV anomia treatment. Each treatment trial consisted of 3 steps (Figure 3), with a naming attempt (step 1) followed by four feature verification questions (step 2), followed by a second naming attempt (step 3).

In step 1, the target picture was presented, and the participant was asked to name it with a single verbal response. Production accuracy, voice key success, and response time was judged online by the clinician and input by key press. If the voice key trigger was successful, participants then received immediate feedback consisting of their response accuracy and RT.

[1] The first author co-treated with the 4th for the first 5-6 sessions for the first 2 participants to ensure consistent application of the BEARS components. The first author was actively consulted and answered questions for the remainder of participants. The second author covered 1 session for participant 6 while the treating clinician was on vacation. All three are certified speech-language pathologists.

In step 2, the target picture was shown again along with an audio recording of the correct response, followed by the auditory and written presentation of a semantic feature question. Participants gave a yes/ no response to each question via key press. Immediate accuracy and response time feedback was provided after they answered each semantic feature question. Four of the eight semantic feature questions were randomly selected for presentation on each treatment trial. After completing four feature verification questions, Step 3 was initiated.

In step 3, the participant was asked to name the picture again "as quickly and accurately as possible," and accuracy and RT feedback were provided. Steps 1-3 were repeated for each target until all eight category items had been practiced, which completed a "round." At this point cumulative point-based feedback for the round was provided based on speed-accuracy performance, and then the process was repeated in a new round for a new category. Every time all 40 items across the five categories were practiced, participants were told they had

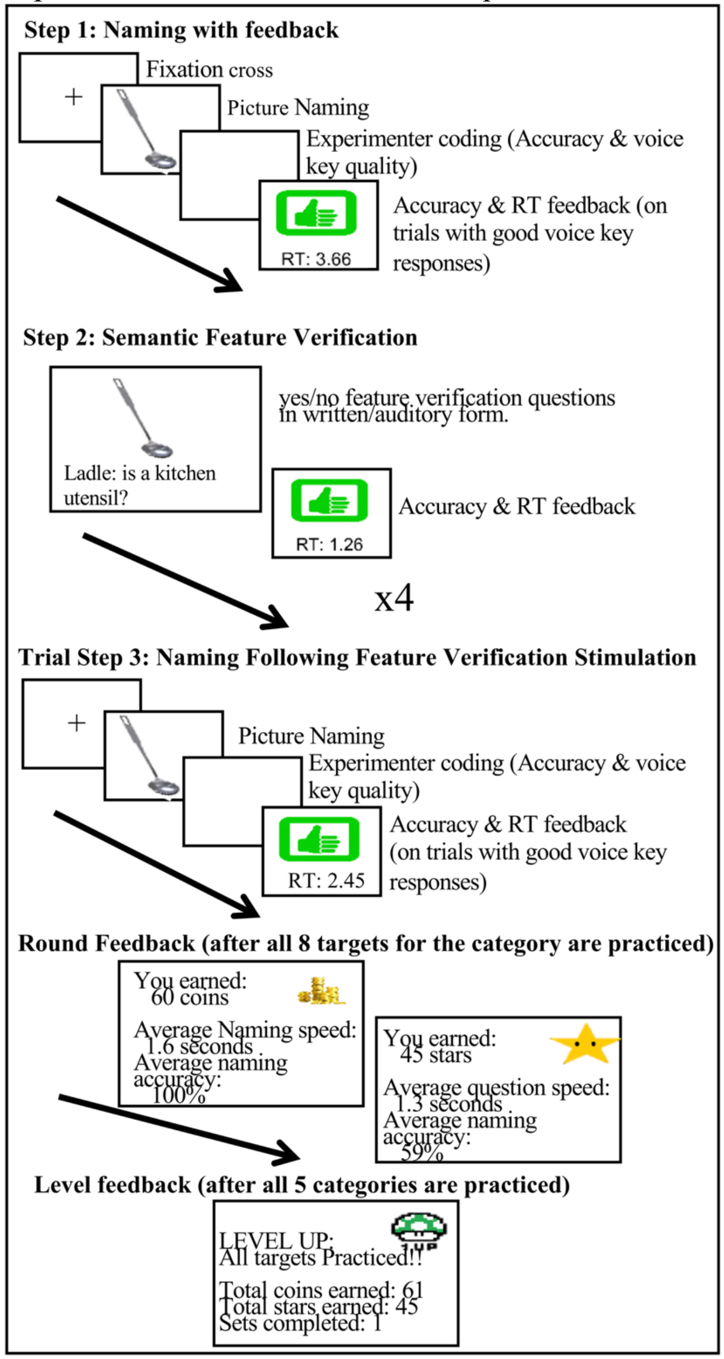

Figure 3. BEARS + SFV treatment schematic of computer-administered components.

completed a “level up” and point-based feedback was provided cumulative across all five rounds. With- and between-category presentation order were determined randomly.

To support BEARS system calibration training, participants received both points feedback and metacognitive strategy training throughout the intervention. Point feedback was based on a modified reward rate algorithm (Bogacz et al., 2006) which rewarded efficiency (both fast and accurate) performance, with correct responses earning 5 points and error responses losing 1 point, divided by the total time spent on each treatment block. Points were awarded separately for naming efficiency (“Coins”) and feature verification efficiency (“Stars”). Points were presented on the round and level up screens and the clinician reviewed the points with the participant, including comparisons to previous blocks and treatment sessions to ensure comprehension of the feedback system. Metacognitive strategy training throughout treatment included ongoing education on the speed-accuracy tradeoff and participants’ level of effort and frustration. Discussion included how speed can occasionally result in more errors, but how slowing down does not always result in retrieving the target word. The clinician reviewed how speech can be modulated and adjusted based on perceived difficulty/ accessibility of each target and introduced the option to “pass” or “move on” once they recognize the feeling that the target word is not accessible at the given time. Adaptive speed-accuracy tradeoffs were reinforced for participants throughout treatment. A detailed description of this training and which participants received which types of training are in Appendix 1 and Supplementary Materials S1.

Analysis

Data were analyzed in R statistical software, version 4.0.2 (R Core Team, 2020). Naming probe data at all baseline and treatment timepoints were used for analysis of treatment response for predictions 1, 2, and 4. For prediction 1 (*BEARS + SFV will replicate previous SFV findings and improve naming accuracy for both treated and semantically-related untreated words*), treatment outcomes were evaluated using the dependent variable of naming probe accuracy. For prediction 2 (*BEARS + SFV will increase naming efficiency*), treatment outcomes were evaluated using the dependent variable of naming probe reward rate (i.e., the number of correct responses per minute). For prediction 3 (*BEARS + SFV will improve discourse informativeness and efficiency*), treatment outcomes were evaluated using the dependent variables of CIUs/minute and proportion of CIUs. For prediction 4 (*BEARS + SFV will improve system calibration for self-monitoring and error awareness*), self-monitoring ability was evaluated using the dependent variable of naming probe "pass rate" (i.e., the proportion of inaccurate trials where participants indicated they could not produce the target by saying "pass" instead of producing an overt error or giving no response by the end of the 15-second response window). For prediction 5 (*Efficient practice performance during BEARS + SFV treatment will be positively associated with good treatment outcomes)*, treatment practice efficiency was evaluated using the dependent variable of the total number of feedback points earned across treatment sessions ("coins" for naming performance and "stars" for feature verification question performance), while dependent variables for treatment outcomes were measured in terms of individual treatment effect sizes.

For predictions 1, 2, and 4, group-level performance was evaluated using Bayesian generalized linear mixed-effect models using the R package BRMS (Bürkner, 2017) following the interrupted time series approach described by Huitema and McKean (2000) and Moeyaert et al. (2017). Bayesian implementations of generalized linear mixed effect models are largely similar to their frequentist variant, but also permit estimation of the probability of a given effect and individual effect sizes, as discussed below. The interrupted time series approach includes fixed effects for baseline slope, level change, and slope change. Together, these fixed effects can characterize the presence of a stable, rising, or declining baseline (i.e., baseline slope), any immediate changes in performance at the onset of treatment (i.e., level change), and whether or not the slope of treatment-related change exceeds that of the slope established during the baseline phase (i.e., slope change). Therefore, positive level change and slope change fixed effects may be found even in the presence of rising baseline slope, which provides evidence that changes in probe performance over time are attributable to the treatment. Models were implemented separately for each item condition (i.e., treated, related untreated, unrelated untreated). Details regarding modeling fitting are reported in Appendix 2.

Individual effect sizes for naming accuracy were estimated for each participant by taking the difference between the model's posterior prediction for each subject at the last treatment probe and final baseline session, resulting in an estimate of the median number of words improved and associated 90% credible interval. An equivalent approach was used to estimate the individual improvements in naming reward rate. Group-level effect sizes were estimated by calculating the difference between posterior samples at session 13 from session 4 at the mean of the random effects for an average PWA, which accounts for performance during baseline.

A major benefit of this Bayesian mixed-effect approach is that a single model can estimate effect sizes and group-level fixed effects, calculating 90% credible intervals and posterior probabilities (i.e., the probability that the effect size or model parameter is greater than zero) in each instance. Together, this a) provides an interpretable point estimate and range for expected treatment effects, b) characterizes the degree of statistical robustness for effect sizes based on posterior probabilities, and c) provides an appropriately conservative model check of whether effect sizes are attributable to the treatment, which is done by comparing posterior probabilities for the fixed effects of baseline slope, slope change, and level change.

For prediction 3, changes in discourse efficiency were calculated for each participant as a measure of far generalization on the Nicholas and Brookshire protocol by calculating CIUs/min and the proportion of CIUs . Changes in both proportion of CIUs and CIUs/minute were analyzed via non-parametric bootstrap test for paired differences (Dwivedi et al., 2017) using the R package *infer* (Bray et al., 2020). This approach is advantageous for small sample sizes as it does not rely on underlying assumptions typical of parametric statistical tests and also demonstrates better power than non-parametric tests in smaller sample sizes.

Prediction 5 was evaluated via Pearson correlations exploring relationships between practice efficiency and treatment outcomes by condition (as calculated for predictions 1 through 4 above). Correlations between pass rates and treatment outcomes by condition were also evaluated in the same correlation matrix as secondary analyses for prediction 4.

Results

Thirteen people with chronic aphasia were enrolled, and nine of the 13 met enrollment criteria during initial assessment. At the time of their enrollment, participants ranged from 9 months to 44 years post-onset of a left-hemisphere stroke. Participants were between ages 55-73 years and were all native speakers of English. Apraxia of speech was either absent or very mild and indistinguishable from phonological output deficits in the majority of participants, with the exceptions of participant 1, who presented with moderate apraxia of speech and participant 2 who presented with probable mild apraxia of speech. Baseline testing on the Comprehensive Aphasia Test and demographic information is presented in Tables 1 and 2, and additional assessment results are reported in supplementary materials (S2).

Table 1. Participant Descriptive Characteristics

Participant	Sex	Age	Etiology	MPO	Years Edu	Premorbid handedness	Race/ ethnicity	Hemiparesis
p1	M	63	CVA	259	14	R	African American	R UE
p2	M	73	CVA	194	14	R	Caucasian	None
p3	M	68	CVA	21	12	L	Caucasian	None
p4	M	70	CVA	522	13	R	Caucasian	R UE
p5	M	70	CVA	39	16	R	Caucasian	None
p6	F	71	CVA	8	14+ years	R	African American	None
p7	M	70	CVA	9	18+ years	R	Caucasian	R UE and LE
p8	M	54	CVA	18	16	R	African American	None
p9	M	72	CVA	58	14	R	Caucasian	None

Note: P= participant, M = Male, F = Female, CVA = Cerebrovascular accident, UE = Upper Extremity, LE = Lower Extremity. P1 had concomitant moderate apraxia of speech and P2 had probable mild apraxia of speech. All participants had a diagnosis of aphasia following left-hemisphere CVA.

Table 2. Comprehensive Aphasia Test Performance (modality T score)

Participant	Semantic Memory	Recognition Memory	Comp of Spoken Language	Comp of Written Language	Repetition	Naming	Reading (aloud)	Writing
p1	60	48	52	51	47	48	45	51
p2	60	59	48	53	47	53	52	50
p3	51	48	55	58	55	59	57	60
p4	60	59	53	53	52	54	48	54
p5	60	48	55	50	53	49	53	52
p6	60	59	57	55	52	54	52	57
P7	40	59	44	35	60	46	44	47
P8	60	48	49	51	47	53	50	58
P9	60	59	49	61	46	55	49	52

For predictions 1 and 2, aggregate naming accuracy and reward rate performance for each participant by session are presented in Figure 4. For predictions 1, 2, and 4, model fixed effect estimates and posterior probabilities are presented in Table 3 by dependent variable and item condition, with full model results in Appendix 2. For predictions 1 and 2, group and individual effect sizes for naming accuracy and reward rate are presented in Figure 5.

Results for prediction 1: BEARS + SFV will replicate previous SFV findings and improve naming accuracy for both treated and semantically-related untreated words.

Fixed effect beta-coefficients are presented as odds ratios. For treated items, credible intervals excluded zero for baseline slope (β = 0.20; 90% credible interval (CI) = 0.10, 0.29), level change (β = 0.56, 90% CI = 0.15, 0.98), and the quadratic term for slope change (β = -0.02; 90% CI = -0.03, -0.01) but not slope change (β = 0.03; 90% CI = -0.10, 0.15). For the untreated conditions, only a positive trend of baseline slope (β = 0.19, 90% CI = 0.08, 0.30) and downward trend for slope change (β = -0.17; 90% CI=-0.28, -0.05) for untreated items was evident. For both untreated conditions, the quadratic term for slope change did not improve model fit and was therefore not included.

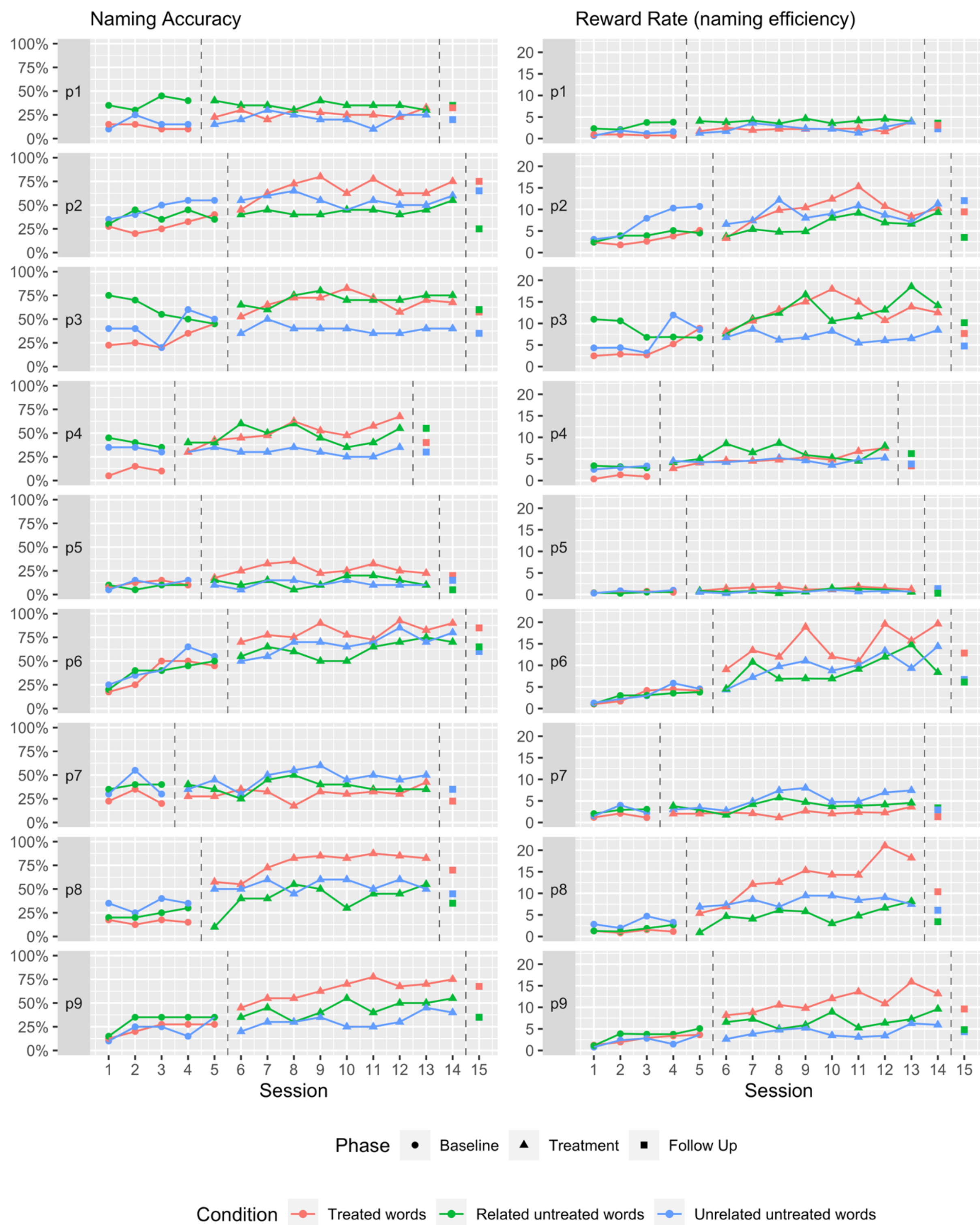

Figure 4. Individual performance on naming probe accuracy and reward rate over time.

Calculation of overall group effect sizes revealed that an average PWA would be expected to name an additional 13.7 treated words (90% CI: 7.3, 19.17) and demonstrate small but meaningful generalization of 2.31 additional related, untreated words (90% CI: 0.31,4.36). No improvements were seen for unrelated words (Effect size: 0.67, 90% CI: -1.37, 2.87). Group effect sizes were consistent with individual effect size estimates (Figure 5). For treated items, individual effect sizes ranged between two and 25 additional words named accurately, and the 90% credible intervals for excluded zero for all participants except for participant 7. The 90% credible intervals excluded zero for four participants for untreated, related items, and excluded zero for two participants for unrelated untreated items.

Table 3 provides a summary of treatment effect sizes by item condition along with fixed effects and posterior probabilities to guide in their interpretation. Taken together, results showed that participants improved on treated words (+13.7 additional words named accurately, posterior probability > 0.99) and these gains were attributable to the treatment (posterior probability for level change = 0.98) despite the presence of rising baselines (posterior probability > 0.99). BEARS + SFV also produced improvements on related untreated words (+2.31 additional words named accurately, posterior probability = 0.97), but these gains could not be clearly distinguished from repeated probe exposure (level change and slope change posterior probabilities ≤ .65). Finally, there were no improvements on unrelated untreated words (+0.67 additional words named accurately, posterior probability = 0.71).

These results provide clear evidence of a treatment effect for trained words and some inconclusive evidence of response generalization to semantically-related untreated words, which is broadly consistent with prediction 1.

Table 3. Summary of group treatment effect sizes, fixed effects, and posterior probabilities for naming probe accuracy, reward rate, and “pass” rate.

Model		Treated words		Related, Untreated words		Unrelated, untreated words	
		Value	*PP*	*Value*	*PP*	*Value*	*PP*
Accuracy							
	Effect Size	13.70	1.00	2.31	0.97	0.67	0.71
	Baseline Slope	0.20	1.00	0.06	0.81	0.19	1.00
	Level Change	0.56	0.98	0.08	0.65	-0.28	0.11
	Slope Change	0.03	0.63	-0.01	0.44	-0.17	0.01
Reward Rate							
	Effect Size	5.23	1.00	2.83	1.00	1.52	0.96
	Baseline Slope	0.22	1.00	0.16	1.00	0.27	1.00
	level change	0.49	0.99	-0.03	0.43	-0.32	0.03
	slope change	-0.15	0.00	-0.09	0.02	-0.22	0.00
Pass Rate							
	Effect Size	0.24	0.98	0.24	0.99	0.18	0.99
	Baseline Slope	-0.14	0.13	0.01	0.53	-0.06	0.37
	level change	1.73	0.99	1.13	0.99	1.70	0.99
	slope change	0.55	1.00	0.14	0.80	0.51	0.98

Note: results are summarized from the Bayesian interrupted times series models completed for predictions 1 through 3. PP = The posterior probability (i.e., the probability that the effect size or model parameter is greater than zero). Effect sizes are reported in additional words named correctly for accuracy, in additional words named per minute for reward rate, and in increased proportion of “pass” vs. other inaccurate trial types for pass rate. Model fixed effects of baseline slope, level change, and slope change are reported as odds ratios. Full model results are reported in Appendix 2.

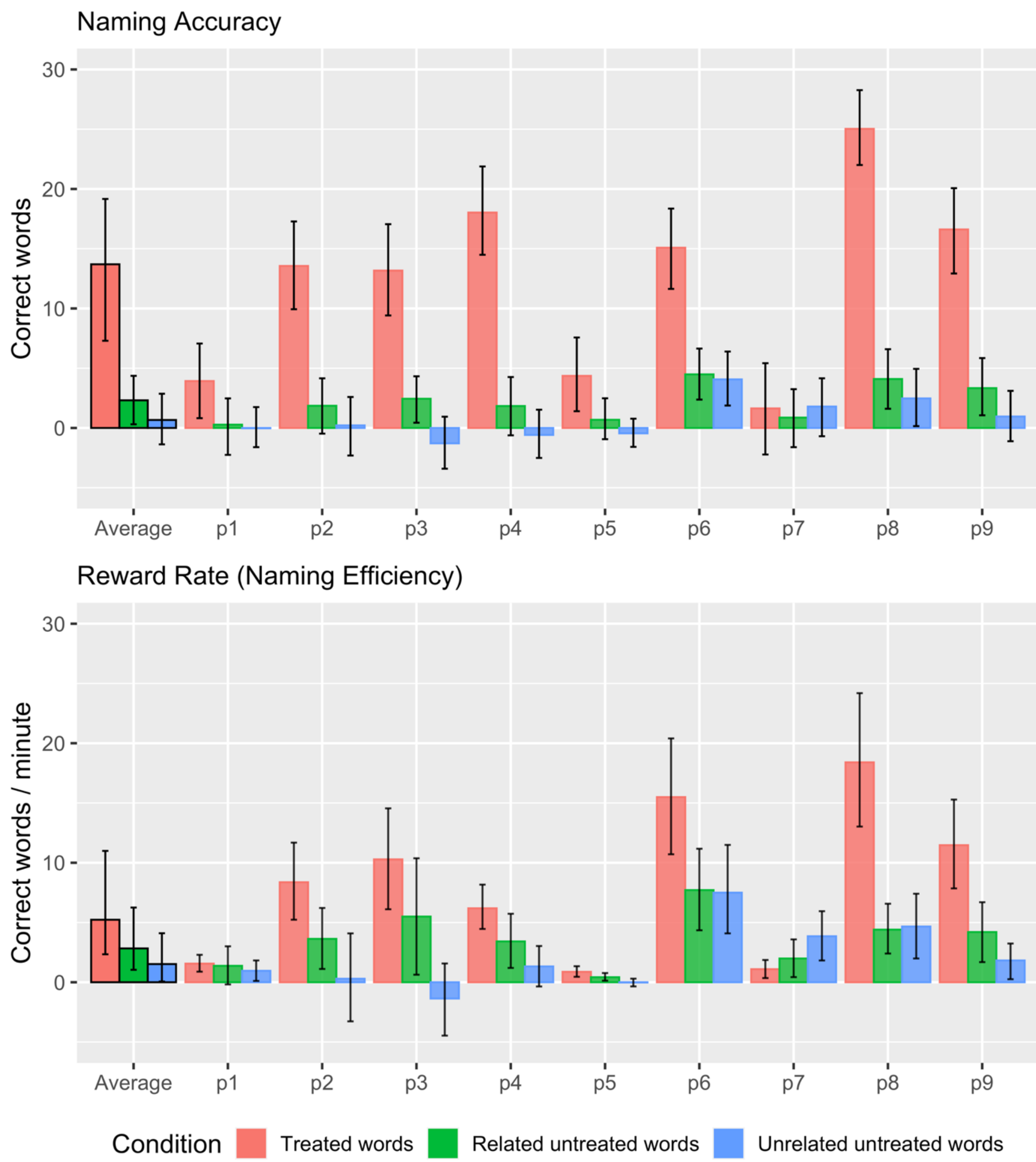

Figure 5. Group and Individual effect sizes for naming probe accuracy (top) and reward rate (bottom). Error bars reflect 90% Bayesian credible intervals.

Results for prediction 2: BEARS + SFV will increase naming efficiency (reward rate).

For treated items, there was a positive baseline slope (β = 0.22; 90% CI = 0.15, 0.29) along with a positive level change (β = 0.49; 90% CI = 0.17, 0.83) and a negative slope change (β = -0.15; 90% CI = -0.22, -0.09), suggesting that reward rate improved during baseline, responded initially to treatment, and then slowed in its rate of improvement. For related and unrelated untreated items, a rising baseline slope was evident (related: β = 0.16; 90% CI = 0.08, 0.23; unrelated: β = 0.27; 90% CI = 0.20, 0.33) but level change and slope change were not meaningfully different from zero, indicating that the rate of improvement during treatment did not exceed the slope established at baseline.

Calculation of overall group effect sizes revealed that the average participant's reward rate improved by 5.23 additional words/minute for treated words (90% CI: 2.34, 10.99), and demonstrated small gains of 2.83 additional words/minute (90% CI: 1.04, 6.25) for related untreated words and 1.52 additional words/minute (90% CI: -0.06, 4.10) for unrelated untreated words. Group effect sizes were consistent with individual effect size estimates (Figure 5). For treated items, individual effect sizes ranged between 1.04 and 17.72 additional words named per minute, and 90% credible intervals excluded zero for all nine participants. For semantically-related untreated items, the 90% credible intervals excluded zero for all participants except for participant 1. For unrelated untreated items, the 90% credible intervals excluded zero for five participants.

As summarized in Table 3, results show that participants improved in naming efficiency for treated words (+5.23 additional words/minute; posterior probability > 0.99) with gains attributable to the treatment (posterior probability for level change = 0.99) despite the presence of rising baselines (posterior probability > 0.99). Participants also demonstrated small, though statistically robust, improvements in naming efficiency for related untreated words (2.83 additional words/minute; posterior probability > 0.99), and unrelated untreated words (1.52 additional words/minute; posterior probability = 0.96) but given rising baseline performance (baseline slope posterior probabilities > 0.99) and lack of an increasing level or rate of improvement during the treatment phase (level change and slope change posterior probabilities ≤ .43) these effects could not be clearly distinguished from repeated probe exposure.

The treatment-related gains in naming efficiency for treated items provides clear evidence in support of prediction 2, while the inconclusive evidence for efficiency gains on untreated items does not.

Results for prediction 3: BEARS + SFV will improve discourse informativeness and efficiency.

Performance on the Nicholas and Brookshire discourse elicitation task is reported in Table 4. On average, participants improved their CIUs/minute by 4.07 and their proportion of CIUs by 3.93% between entry and exit. However, these differences were not significant for either CIUs/minute (p = .18) or proportion of CIUs (p = .13). Changes in discourse efficiency (Table 4) were highly variable, with pre-post improvement by up to 21.29 CIUs/minute (participant 7) and decreases as large as 12 CIUs/minute (participant 5[2]). These results do not support prediction 3.

[2] Participant 5's discourse performance was highly tangential, which appeared to increase the overall session-to-session variability in his performance.

Table 4. Pre- and Post-treatment Discourse Results (Nicholas and Brookshire protocol)

	CIUs/minute Pre-treatment	CIUs/minute Post-treatment	Total CIUs Pre-treatment	Total CIUs Post-treatment	% CIUs Pre-treatment	% CIUs Post-treatment	Total words Pre-treatment	Total words Post-treatment
p1	24.4	26.72	194	167	37.09	39.02	523	428
p2	30.82	29.26	187	118	29.12	33.81	642	349
p3	66.92	74.31	377	379	40.1	59.4	940	638
p4	50.11	58.81	228	248	48.1	56.11	474	442
p5	30	18	213	182	25.09	13	849	1399
p6	20.06	25.73	237	193	53.26	53.17	445	363
p7	38.2	59.49	198	235	52.94	63.17	374	372
p8	21.73	30.45	151	101	28.76	34.12	525	296
p9	42.22	38.29	254	224	47.65	45.71	533	490

Results for prediction 4: BEARS + SFV will improve system calibration for self-monitoring and error awareness (pass rate).

There was no effect of baseline slope for treated or untreated items (Appendix 2 and Table 3). A positive level change was present in all three conditions while a slope change was evident in the treated and unrelated conditions, suggesting that participants increased their proportion of "pass" responses compared to other error types after the onset of treatment and as treatment progressed. However, pass rates were highly variable and did not change equally for all participants (supplementary materials, S3).

As summarized in Table 3, results show that BEARS + SFV produced improvements in pass rate for treated words (mean rate increase of 0.24; posterior probability = 0.98) attributable to the treatment (posterior probability for level change = 0.99 and for slope change > 0.99). BEARS + SFV also produced similar changes in pass rate for related, untreated words (mean rate increase of 0.24; posterior probability = 0.99), and unrelated untreated words (mean rate increase 0.18; posterior probability = 0.99). Because of the stable baselines and high posterior probabilities for slope and level change across conditions, we attribute changes in pass rate directly to BEARS + SFV, indicating that the participants were able to increase their pass rate in response to training. These results support prediction 4.

As a secondary analysis, we looked at correlations between pass rates, treatment effect sizes, and aphasia severity (Figure 6). There was a particularly strong relationship between pass rate and treated item effect sizes for naming accuracy (r = -0.93), such that pass rates were higher for participants with lower effect sizes. Correlations between pass rate and other naming probe effect sizes were also negative (rs between -0.39 and -0.67), and the correlation between pass rates and aphasia severity was weak (r = -0.31). Visual inspection of pass rates by participant over time (supplementary materials S3) shows that the three participants with the lowest effect sizes for treated naming probe accuracy (participants 1, 5, and 7, Figure 5) demonstrated the largest increases in pass rates as a result of treatment. These results are also broadly consistent with prediction 4.

Results for prediction 5: Efficient practice performance during BEARS + SFV treatment will be positively associated with good treatment outcomes.

Pearson correlations between practice efficiency (i.e., total feedback points earned with “coins” for naming performance and “stars” for feature verification), naming probe treatment effect sizes for accuracy and reward rate, and pre-post changes in discourse performance are presented in Figure 6. Individual practice efficiency performance (coins and stars earned by session) is presented in supplemental materials (S3). There was a strong positive correlation between total coins and stars earned (r = 0.94) and these measures therefore demonstrated similar relationships to other variables. For correlations between practice efficiency and naming probe effect sizes, there were a) strong positive correlations for reward rate on untrained related items (coins, $r = 0.94$; stars, $r = 0.89$), b) strong positive correlations for reward rate on treated items and on accuracy for related untreated items (*r*s between 0.67 and 0.73), and c) weak positive correlations on accuracy for treated items (coins, $r = 0.31$; stars, $r = 0.25$). There were weak positive correlations between practice efficiency and measures of discourse improvement (*r*s between 0.22 and 0.33).

Aphasia severity has been found to predict treatment outcomes for semantically-oriented anomia treatment (Quique et al., 2019) and could potentially play a role in how well individuals were able to efficiently practice BEARS + SFV. Therefore, we examined correlations between aphasia severity, practice efficiency, and treatment outcomes to explore whether practice efficiency served as a proxy measure of aphasia severity in the current data (Figure 6). There were moderate correlations between aphasia severity and practice efficiency (coins, $r = 0.37$; stars $r = 0.42$), weak-to-moderate correlations between aphasia severity and naming probe effect sizes (*r*s between 0.29 and 0.47), and essentially null correlations between aphasia severity and discourse-related changes (proportion of CIUs, $r = 0.028$; CIUs per minute, $r = -0.14$).

Overall, these exploratory analyses demonstrated strong positive correlations between practice efficiency and most measures of naming probe effect size, and only moderate correlations between overall aphasia severity and practice efficiency, which is broadly consistent with prediction 5.

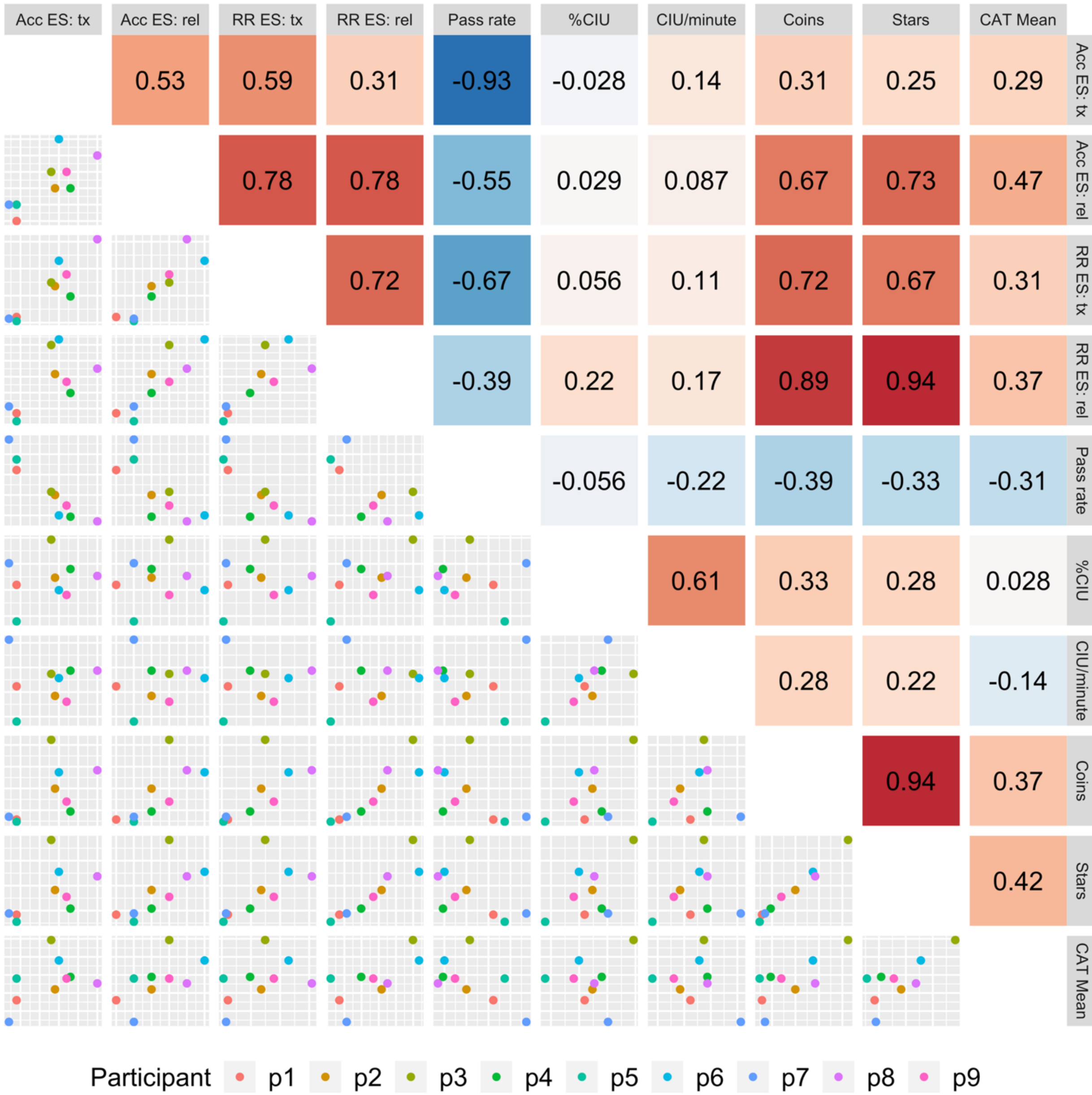

Figure 6. Correlations and scatterplots between treatment outcomes, measures of treatment practice efficiency, and aphasia severity. "Coins" = total feedback points earned for naming practice efficiency, "Stars" = total feedback points earned for feature verification practice efficiency. "Acc ES: tx" = effect size for naming probe accuracy on treated items. "Acc ES: rel" = effect size for naming probe accuracy on semantically-related untreated items."RR ES: tx" = effect size for naming probe reward rate on treated items."RR ES: rel" = effect size for naming probe reward rate on semantically-related untreated items items. %CIU = proportion of CIUs. CAT mean = aphasia severity (CAT mean modality T score). r values > .67 are significant at uncorrected alpha = .05. r values > .836 are significant at Bonferroni-corrected alpha = .005.

Discussion:

The purpose of this study was to develop and pilot a BEARS-augmented anomia treatment (BEARS + SFV) that combined computer-based feedback and clinician-provided metacognitive system calibration training. We developed BEARS + SFV based on the rationale that training PWA to better balance their effort, accuracy and response speed during drill-based anomia treatment would allow them to make more efficient use of their *current* language system for the purposes of both language adaptation and restorative treatment. We hypothesized that this in turn could directly improve performance efficiency during probe tasks via a) better adaptation (e.g., PWA learning to inhibit avoidable paraphasias and/or move on more quickly from unproductive word retrieval attempts) and b) better restorative treatment outcomes for trained and related untrained words by allowing for by higher dosage, more successful effortful trials, and fewer produced errors during the course of treatment. Isolating the unique effects of BEARS-augmented compared to un-augmented anomia treatment was outside the scope of the current study as this would require a larger-scale comparative effectiveness study. BEARS+SFV showed good feasibility as a pilot treatment. The nine participants who met eligibility criteria successfully completed all study procedures and the full 25 hours of treatment, with zero attrition or loss to follow-up. The remaining study goals are addressed in turn.

Does BEARS + SFV replicate previous SFV findings on performance accuracy and improve naming efficiency?

Analyses for predictions 1 and 2 showed similar results for treated and untreated items on naming probes. Direct training effects were found for treated words, measured both in terms of accuracy (prediction 1) and reward rate (prediction 2). Rising baselines were noted in both analyses, but additional treatment-related level increases in performance were noted in both instances. Group effect sizes indicated that after 25 hours of BEARS + SFV training, a typical individual with aphasia would be expected to accurately name an additional 13.7 out of 40 trained words and improve their efficiency by naming an additional 5.23 accurate trained words per minute. At the individual level, eight out of nine participants improved on accuracy and all nine improved in terms of reward rate. The accuracy effect size compares favorably to previously reported anomia studies, which tend to train fewer words, often for equal or longer periods of time (e.g., Snell et al., 2010). Together, results provide preliminary evidence that participants improved in both accuracy and efficiency of treatment-related words as a result of BEARS + SFV.

There was less robust evidence for treated-related gains for untrained words. Group effect sizes showed clear improvement for semantically-related untreated words on naming probe accuracy (consistent with prediction 1), and clear gains on naming probe reward rate for both related and unrelated untreated words (consistent with prediction 2). However, rising baselines were noted, and there was essentially no evidence of treatment-related slope or level change for either the accuracy or reward rate analyses (posterior probabilities ≤ 0.65). In other words, while participants improved their accuracy on related untreated words and their efficiency on both related and unrelated untreated words over the course of the study, these gains cannot be clearly distinguished from the positive effects of repeated exposure observed in the baseline phase[3].

[3] That being said, the calculation of fixed effects for treatment-related level and slope change rely on the modeling assumption that rising baselines would have continued linearly throughout the duration of the study even in the absence of treatment, which may be overly conservative. This highlights a general issue in single-subject designs when rising baselines are observed: it is not possible to determine whether a linear trend would have continued at the same rate (instead of leveling off) in the absence of treatment. This is problematic in regards to anomia treatment research in general, as repeated exposure to naming opportunities without feedback has been shown to improve naming abilities in at least some individuals (Creet et al., 2019). Given these considerations, we

These results broadly support prediction 1 ("*BEARS + SFV will replicate previous SFV findings and improve naming accuracy for both treated and semantically-related untreated words*"), and are generally consistent with previous work looking at SFA and SFV (e.g., Kiran & Roberts, 2010, Gilmore et al., 2020). We found robust direct training effects and weak evidence of generalization to related untrained words. While rising baselines and methodological issues (see Limitations) may have made generalization harder to capture, these results are also consistent with previous claims that SFV may produce more modest generalization effects than feature generation-based SFA (Boyle, 2010).

These results also partially supported prediction 2 ("*BEARS + SFV will increase naming efficiency (reward rate)*"). There was clear evidence of treatment-related improvements on naming efficiency observed across participants. Since reward rate is a measure of the total number of correct responses per unit of time, some portion of efficiency gains for trained words are attributable to improved word retrieval engendered by the SFV treatment component. However, the BEARS system calibration component also appeared to play some role in these gains, because even participant 7 who did not show clear gains in naming accuracy effect size improved in reward rate effect size.

We predicted that if the BEARS treatment component improved naming efficiency on its own via adaptive system calibration, participants would demonstrate efficiency gains on *unrelated* untreated items, since the SFV treatment component was not predicted to improve performance on these words. There were positive group-level gains in reward rate effect sizes for unrelated untrained words (displayed by five of the nine participants at the individual level,

---

will interpret positive effect sizes in the absence of robust treatment-related slope and level change as weak positive evidence (as opposed to null evidence) in the current study.

Figure 5), but as noted above, these gains could not be distinguished from rising baselines/ effects of repeated probe exposure based on the group-level fixed effects (Table 3). Overall, these results provide clear evidence that BEARS + SFV improves naming efficiency for trained words, but provides only weak evidence that BEARS training improves naming efficiency for untrained words in some participants.

Does BEARS+SFV improve discourse informativeness and efficiency?

We hypothesized that improved system calibration induced BEARS training at the single-word treatment during treatment could generalize to the level of connected speech and thereby improve discourse informativeness and efficiency (prediction 3). However, this study did not find any group-level changes in discourse performance (Table 4). While some individuals improved in discourse informativeness and efficiency, others did not, and changes in these patterns were not correlated with individual treatment response or aphasia severity (Figure 6). These findings are consistent with the generally mixed findings regarding discourse-related changes as a result of SFA (Rider et al., 2008; Silkes et al., 2020; Wallace & Kimelman, 2013). Previous work has also shown test-retest instability on the Nicholas and Brookshire discourse protocol (Cameron et al., 2010). This makes it difficult to detect treatment-related changes, and suggests our discourse analyses were likely underpowered. In addition, BEARS+SVF computer practice was completely focused on the single-word level, as was the great majority of the complementary BEARS training. Therefore, the current null results suggest that BEARS training at the single-word level is insufficient to induce system calibration improvements at the discourse level (see future directions).

Does BEARS + SFV improve system calibration for self-monitoring and error awareness?

As part of training, participants were educated about speed-accuracy tradeoffs and how to appropriately balance effort, accuracy, and response speed based on their own processing ability. They were trained to provide a single naming response and to become more aware of instances of anomia when they were very unlikely to produce a word correctly. In such instances, they were instructed to say "pass" instead of producing an overt error or waiting until the allotted time ran out, which was intended to increase overall efficiency and reduce the production of overt errors. Therefore, we hypothesized that participants would demonstrate improved adaptation and system calibration as by improving their proportion of "pass" responses relative to other error types (paraphasias and timeout nonresponses). Results fully supported this prediction. As a group, participants demonstrated improved pass rates attributable to the treatment. Correlation analyses showed that pass rates were inversely proportional to treatment effect sizes for naming probe accuracy, with higher pass rates for participants with lower effect sizes. This indicates participants who demonstrated only small restorative treatment gains in response to the restorative SFV component were still able to demonstrate strategic adaptation to their own poor performance. Increased pass rates do not appear to have caused lower treatment responses, because across participants, overt error and "pass" responses took longer on average than correct responses. This means that choosing to say "pass" was more likely to replace an overt error than a correct response, and therefore did not reflect giving up too early and losing opportunities for correct retrieval. Overall, we interpret these findings as evidence of improved system calibration as a result of the BEARS component of this intervention.

Participant 7 provides a helpful illustrative example of these effects: although he was the only participant who did not improve on probe accuracy (effect size 90% credible intervals all included zero, Figure 5 top panel), he demonstrated small but robust gains in naming efficiency across item categories (Figure 5, bottom panel). Given his steadily increasing pass rates (supplementary materials, Figure S3a), his gains in naming probe efficiency are likely attributable to improved system calibration as opposed to restored naming ability.

Is practice efficiency positively associated with treatment outcomes?

One of the primary rationales for developing BEARS + SFV was a consideration of how overly impulsive and overly cautious speed-accuracy tradeoffs could negatively affect treatment outcomes in drill-based restorative treatment. Overly impulsive responses may increase the number of avoidable errors and therefore may reduce treatment outcomes via error learning (Fillingham et al., 2006), while overly cautious responses are slow and may reduce treatment outcomes via reduced overall dosage. Based on this premise, the computer-based feedback in BEARS + SFV used an algorithm which awarded points based on maximizing the number of correct responses while minimizing the number of errors over time. This feedback was designed to help participants better balance speed and accuracy during practice in a way that maximized both effortful and errorless learning principles (Schuchard & Middleton, 2018). We hypothesized that more efficient practice, as measured by a great number of total "Coin" and "Star" feedback points earned, would be correlated with larger treatment effect sizes (prediction 5). This prediction was largely confirmed.

While our group-level analyses for prediction 1 only found weak support for generalization to related untrained items (i.e., small effect sizes without clear changes in treatment-related slope or level change), we actually found strong correlations between practice efficiency and generalization to related untrained naming probe accuracy using a case series correlational approach (Rapp, 2011). In other words, participants who produced more correct naming and feature verification responses and fewer errors over 25 hours of treatment also demonstrated better response generalization to semantically-related untrained items. In contrast, there were only weak correlations between practice efficiency and trained naming probe accuracy. We attribute this to the fact that best-possible accuracy performance for trained items was observed approximately half-way through the treatment for most participants (Figure 4). This suggests that dosage in the current study was sufficient to produce participant-specific ceiling effects for trained item accuracy (regardless of individual practice efficiency), but that the degree of response generalization specifically predicted by the SFV treatment component was affected by overall levels of practice efficiency.

In addition, we found strong positive correlations between practice efficiency and individual effects sizes for naming efficiency on both trained and related untrained words. It is not particularly surprising that more efficient practice during treatment corresponded to more efficient probe performance, but it does support the importance of considering not only accuracy but also efficiency in anomia treatment.

We would also note that these exploratory findings related to practice efficiency do not appear to be merely a result of aphasia severity. Correlations between aphasia severity, practice efficiency and effect sizes suggest that a relatively small amount of variance was shared between practice efficiency and aphasia severity ($r^2 = 0.15$), while higher proportions of variance were shared between practice efficiency and related untrained naming probe treatment response (r^2s ranging between 0.45 and 0.88). Individual predictors of this treatment response should be further explored. Overall, these results provide support for targeting practice efficiency during drill-based restorative anomia treatment. However, these results are preliminary and cannot distinguish correlation from causation. Future work should confirm these exploratory findings and evaluate the relative contributions of efficiency-focused practice in relation to learning theory.

Preliminary guidance for treatment candidacy based on individual case results.

Most of the participants who showed good treatment responses were those who improved in their practice efficiency during the course of treatment, and clear increases in practice efficiency were apparent by the fourth treatment session for these participants (Supplementary Figure S3b). This suggests that early treatment performance may be predictive of overall treatment response (Simic et al., 2020).

In contrast, the three participants with the lowest treatment responses in naming accuracy (participants 1, 5, and 7) also had the most severe anomia per PNT scores and participant 7 had the low semantic control performance of all participants as measured by the CCT (Supplementary language testing: S2), and Participant 1 was the only individual in the sample who presented with clear (more than suspected/very mild) apraxia of speech. Participants 5 and 7 were also the only two participants who consistently presented with overly conservative response patterns during treatment (see Appendix 1 and Supplementary materials S1). These characteristics may be negatively prognostic in terms of overall treatment response. However, some degree of nuance is required in considering this characterization of "non-responders".

Participants 1, 5 and 7 all demonstrated minimal naming gains naming probe accuracy, but all three improved in proportion of “pass” compared to other error responses, demonstrated modest improvements in naming efficiency, and participants 5 and 7 also demonstrated gains in functional communication per anecdotal family report (Supplementary materials S1). Together, these patterns suggest that future research should consider the distinct effects of restorative vs.

compensatory BEARS training, as PWA who are poor restorative treatment candidates may still be good candidates for improved adaptation and system calibration.

Study limitations:

There were limitations in this study regarding probe design, stimuli selection, stimuli scoring, participant selection, and dosing of treatment components that should be addressed in future work. Despite being matched for general production difficulty using an algorithm based on word frequency, age of acquisition, and word length in phonemes (Fergadiotis et al., 2015), untreated items had higher baseline performance than treated items, which may have negatively affected our ability to detect treatment-related generalization. We administered potential treatment items twice for selection (and selected items that were ≤ 50% accurate), but relied on difficulty matching to select untreated items in order to decrease testing burden (i.e., 224 instead of 672 words, administered twice). In retrospect, we realize that our approach assumed that if a participant could not name an easier potential treatment word, then they would be equally unlikely to name other words of the same difficulty level. However, 37% of the variance in naming accuracy was unexplained in Fergadiotis et al.'s (2015) prediction model, and our pre-selection assessment of treated but not untreated words in essence created a filter that allowed for "regression to the mean" on untrained words. As a result, when we selected easier words for treatment, participants were more likely to be able to name their matched pairs correctly. Predicted treatment generalization was still observed in terms of group effect sizes, but could not be distinguished from effects of repeated probe exposure; these effects may be partially attributable to this design limitation. As an additional limitation, naming and discourse probes were scored by the fourth author (who administered most of the assessment and treatment) and therefore were not blinded for time-point.

As discussed above, simple probe exposure can produce rising baseline effects (Creet et al., 2019). However, an additional potential limitation in this study was that PWA saw the written word-form for all treated and untreated items in a lexical decision task that was presented at each probe timepoint, which was administered for the purposes of secondary analyses not reported here. The lexical decision task was always provided after naming probes to minimize potential bias, and no feedback was provided in either task, but positive effects of this additional exposure cannot be ruled out. However, if simple repeated probe exposure in picture plus written form without feedback is indistinguishable from treatment-related generalization resulting from 25 hours of intensive SFV treatment, then the clinical efficiency of SFV and similar semantically-orient anomia treatments may require reevaluation. Given the pilot nature of this study, we chose not to recruit PWA with very severe anomia given their low treatment response to SFA (Quique et al., 2019). However, our participants who did not improve in terms of accuracy still improved in efficiency and pass rate. Therefore, the adaptive component of BEARS may be effective for individuals with more severe anomia and this should be examined in future work.

As noted, BEARS + SFV consists of two core components: computer-based semantic feature verification naming treatment and BEARS meta-cognitive system calibration training. While the dosage of meta-cognitive component was flexibly adjusted in the current study based on perceived participant need, the relative dosage of each component needs to be more carefully characterized in future comparative effectiveness work applying clearly established treatment fidelity procedures. Clinicians seeking to apply this approach should also weigh the relative benefits of each component for a given patient, based on their ongoing performance.

In aligning the current pilot with previous SFV work, several additional methodological differences should be noted when interpreting these findings. We trained a larger number of words than is typical (40 items split between 5 categories). While previous approaches have trained individual categories sequentially, we trained all five categories within each session, with practice blocked by category. However, these differences did not appear to be detrimental based on favorable treated effect sizes for trained words. In addition, the efficiency focus of the BEARS treatment component could have decreased depth of processing for the semantic feature questions. But if so, faster processing of questions would have decreased generalization effects to semantically-related untrained items, and the opposite pattern was found. Finally, treatments that rely on SFV also generally manipulate category typicality (e.g., Gilmore et al., 2020), which was not specifically addressed here. Given these differences, the current study findings reflect a modest extension of SFV-based anomia treatment.

Future directions:

In the current study, we relied on measures of efficiency (e.g., reward rate for naming probes) as proxy measures for system calibration, but future work could apply response time modeling techniques (Evans et al., 2019; Evans et al., 2020) to better assess changes in overly impulsive and overly conservative speed-accuracy tradeoffs and provide individualized computer-based feedback.

BEARS + SFV did not produce reliable changes in discourse informativeness or efficiency at the group level, suggesting that the word-level focus of this treatment was too far removed as a practice context to promote generalization to connected speech (Stokes & Baer, 1977; Thompson, 1989). However, there was a great deal of individual variability and several additional findings that suggest at least some changes may have occurred beyond the single-word level. These including anecdotal family reports of improved functional communication (participants 3, 5, and 7) and at least one instance of negative training transfer to conversation consisting of the overgeneralization of “pass” responses (participant 4, Supplementary materials S3). Therefore, it is worth exploring whether BEARS system calibration training might better improve communication efficiency at the discourse level (Whitney & Goldstein, 1989) if used to augment treatments which target discourse performance directly (e.g., BEARS + Attention Reading and Constrained Summarization, Rogalski & Edmonds, 2008)**.**

When providing BEARS meta-cognitive training and feedback, we paid a great deal of attention to participants’ body language and visible muscle tension as proxies for their frustration and overall effort. Electromyography or other appropriate biosignals may be sensitive to these observations, which could allow for the development of biofeedback-based BEARS training.

In motivating BEARS + SFV, we drew a distinction between the PWA’s core linguistic impairments and how adaptively they make use of their current system to optimize performance. It is likely that the ability to flexibly adjust to linguistic impairments and task constraints relies on domain-general cognitive abilities, which have been increasingly implicated as concomitant deficits in PWA (e.g., Gilmore et al., 2019; Murray, 2012). They also may depend on person-level factors such as linguistic anxiety (Cahana-Amitay et al., 2011). Therefore, cognitive and person-level predictors of treatment response should be explored in future work adequately powered to examine such effects.

In addition, we used the conceptualization of adaptation vs. impairment symptoms from Adaptation Theory (Kolk & Heeschen, 1990) to motivate this study's focus on adaptive system calibration. Original support for Adaptation Theory reported differences in self-awareness of grammatical output for people with fluent vs. nonfluent aphasia, with nonfluent individuals with Broca's aphasia argued to be more aware of the grammaticality of their output than individuals with Wernicke's aphasia, and therefore more able to strategically adapt their spoken output using telegraphic speech (Kolk & Heeschen, 1992). While research based on aphasia classification is outside the current scope of this pilot study (and inconsistent with our choice to characterize the language profiles of our participants using the CAT), future work could explore the relationship between aphasia syndromes, error awareness, and response to BEARS system calibration training. This would be especially important if BEARS was used to augment discourse-level treatments, which would be more dependent upon fluency considerations.

Conclusion:

The purpose of this study was to develop and pilot a BEARS-augmented anomia treatment (BEARS + SFV) that combined computer-based feedback and metacognitive system calibration training. BEARS + SFV showed good feasibility as a pilot treatment. Results provided strong evidence for direct training effects on naming probe accuracy and some weaker evidence of generalization to semantically-related untrained words, consistent with previous semantically-oriented anomia SFA/ SFV research (Kiran & Roberts, 2010, Quique et al., 2019, Boyle, 2010). Results provided strong evidence for direct training effects on naming probe efficiency, and some weaker evidence that the BEARS treatment component improved naming efficiency for untrained words. There were no group-level improvements in measures of discourse performance, but participants did demonstrate improved system calibration based on their ability to shift the nature of their responses on inaccurate treatment trials, with an increasing proportion of “pass” responses compared to paraphasia or timeout nonresponses. In addition, computer-based feedback and BEARS training was designed to promote practice efficiency, and practice efficiency during treatment was positively correlated with treatment outcomes. Follow-up work will be necessary to replicate these effects and distinguish correlation from causation, but these findings are consistent with the claim that improving practice efficiency in SFV anomia treatment leads to greater treatment generalization and improved naming efficiency.

Overall, this study establishes the feasibility of BEARS + SFV and provides preliminary evidence that it improves naming efficiency, especially for trained words, and response adaptation for inaccurate trials. Therefore, future work should examine BEARS-augmented compared to standard aphasia interventions in well-powered comparative effectiveness research designed to characterize specific contributions of BEARS training on restorative and compensatory treatment outcomes.

On a final note, the current study also highlights the importance of considering processing speed in addition to accuracy in anomia treatment. People with very mild aphasia still report frustration over language efficiency (Cavanaugh & Haley, 2019), and need to improve in efficiency even when at ceiling for accuracy (Neto & Santos, 2012). On the other end of the proficiency spectrum, people with severe aphasia and non-responders may still have the potential to improve language efficiency, as suggested by the current findings. Inefficient communication is a major source of frustration, which makes it an important treatment target in its own right.

While the current study provides only preliminary support for BEARS, applying elements of this system calibration treatment framework may still be of interest to clinicians, and a summary of BEARS training and education materials have been made available in Appendix 1.

Acknowledgments and funding statement:

This research was funded by the VA Pittsburgh Healthcare System Geriatric Research Education and Clinical Center, the VA Healthcare Network- VISN 4 Competitive Career Development Fund, and the VA RR&D service (IK1 RX002475), with funds awarded to William S. Evans. Sincere thanks to Michael Walsh Dickey and Patrick Doyle for help in the initial conceptualization of this work, to Stacey Kellough for coding the treatment software, to Isabelle Hunt for help hand-scoring voice key reliability data, to Alyssa Autenreith and Angela Grzybowski for help with verification questions ratings, and to the following students for their efforts writing verification questions: Diana Shraybman, Anisah Rafi, Zheng Wang, Hannah Adams, Jocelyn Dunlap, Erin Womack, Sarah Juliani, Alexis Hensel, Kimberly Wong, Deanna Della Pia, Christine Lambrou.

References

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.

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/https://doi.org/10.1016/j.cortex.2013.01.005

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700-765.

Boyle, M. (2004). Semantic Feature Analysis Treatment for Anomia in Two Fluent Aphasia Syndromes. *American Journal of Speech-Language Pathology, 13*(3), 236-249. http://dx.doi.org/10.1044/1058-0360(2004/025)

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

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

Bray, A., Ismay, C., Chasnovski, E., Baumer, B., Cetinkaya-Rundel, M., Couch, S., Laderas, T., Solomon, N., Hardin, J., Kim, A. Y., Fultz, N., Friedman, D., Cotton, R., & Fannin, B. (2020). *infer Package.* In (Version 0.5.3) https://cran.r-project.org/web/packages/infer/infer.pdf

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

Brogan, E., Godecke, E., & Ciccone, N. (2020). Behind the therapy door: what is “usual care” aphasia therapy in acute stroke management? *Aphasiology*, 1-23. https://doi.org/10.1080/02687038.2020.1759268

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. (2017). Advanced Bayesian multilevel modeling with the R package brms. *arXiv preprint arXiv:1705.11123*.

Cahana-Amitay, D., Albert, M. L., Pyun, S. B., Westwood, A., Jenkins, T., Wolford, S., & Finley, M. (2011). Language as a Stressor in Aphasia. *Aphasiology, 25*(2), 593-614. https://doi.org/10.1080/02687038.2010.541469

Cameron, R. M., Wambaugh, J. L., & Mauszycki, S. C. (2010). Individual variability on discourse measures over repeated sampling times in persons with aphasia. *Aphasiology, 24*(6-8), 671-684. https://doi.org/10.1080/02687030903443813

Campanella, F., Skrap, M., & Vallesi, A. (2016). Speed-accuracy strategy regulations in prefrontal tumor patients. *Neuropsychologia, 82*, 1-10. https://doi.org/10.1016/j.neuropsychologia.2016.01.008

Carragher, M., Conroy, P., Sage, K., & Wilkinson, R. (2012). Can impairment-focused therapy change the everyday conversations of people with aphasia? A review of the literature and future directions. *Aphasiology, 26*(7), 895-916. https://doi.org/10.1080/02687038.2012.676164

Cavanaugh, R., & Haley, K. L. (2020). Subjective communication difficulties in very mild aphasia. *American Journal of Speech-Language Pathology*, *29*(1S), 437–448. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0222

Ceccarini, F., Guerra, S., Peressotti, A., Peressotti, F., Bulgheroni, M., Baccinelli, W., Bonato, B., & Castiello, U. (2020). Speed–accuracy trade-off in plants. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01753-4

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology, 14*(2), 133-142.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407.

Conroy, P., Sotiropoulou Drosopoulou, C., Humphreys, G. F., Halai, A. D., & Lambon Ralph, M. A. J. B. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *141*(6), 1815-1827.

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

Dietz, A., Thiessen, A., Griffith, J., Peterson, A., Sawyer, E., & McKelvey, M. (2013). The renegotiation of social roles in chronic aphasia: Finding a voice through AAC. *Aphasiology, 27*(3), 309-325. https://doi.org/10.1080/02687038.2012.725241

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine, 36*(14), 2187-2205. https://doi.org/10.1002/sim.7263

Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A Systematic Review of Semantic Feature Analysis Therapy Studies for Aphasia. *J Speech Lang Hear Res, 61*(5), 1261-1278. https://doi.org/10.1044/2018_jslhr-l-16-0330

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

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

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.

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2006). The treatment of anomia using errorless learning. *Neuropsychological Rehabilitation*, *16*(2), 129-154. https://doi-org.pitt.idm.oclc.org/10.1080/09602010443000254

Fridriksson, J., Holland, A. L., Beeson, P., & Morrow, L. (2005). Spaced retrieval treatment of anomia. *Aphasiology, 19*(2), 99-109. https://doi.org/10.1080/02687030444000660

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2019). Non-linguistic cognitive factors predict treatment-induced recovery in chronic post-stroke aphasia. *Archives of Physical Medicine and Rehabilitation, 100*(1), 1251-1258. https://doi.org/10.1016/j.apmr.2018.12.024

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2020). Typicality-based semantic treatment for anomia results in multiple levels of generalisation. *Neuropsychological Rehabilitation, 30*(5), 802-828. https://doi.org/10.1080/09602011.2018.1499533

Goodglass, H. (1980). Disorders of naming following brain injury. *American Scientist, 68*(6), 647-655.

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.

Hartig, F. (2020). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models* In

Harvey, S. R., Carragher, M., Dickey, M. W., Pierce, J. E., & Rose, M. L. (2020). Treatment dose in post-stroke aphasia: A systematic scoping review. *Neuropsychol Rehabil*, 1-32. https://doi.org/10.1080/09602011.2020.1786412

Heeschen, C., & Schegloff, E. A. (1999). Agrammatism, adaptation theory, conversation analysis: On the role of so-called telegraphic style in talk-in-interaction. *Aphasiology, 13*(4-5), 365-405.

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

Hunt, P., Soto, G., Maier, J., Müller, E., & Goetz, L. (2002). Collaborative teaming to support students with augmentative and alternative communication needs in general education classrooms. *Augmentative and Alternative Communication, 18*(1), 20-35. https://doi.org/10.1080/aac.18.1.20.35

Kay, J., Lesser, R., & Coltheart, M. (1996). Psycholinguistic assessments of language processing in aphasia (PALPA): An introduction. *Aphasiology, 10*(2), 159-180.

Kiran, S., & Roberts, P. M. (2010). Semantic feature analysis treatment in Spanish–English and French–English bilingual aphasia. *Aphasiology, 24*(2), 231-261.

Kolk, H., & Heeschen, C. (1990). Adaptation symptoms and impairment symptoms in Broca's aphasia. *Aphasiology, 4*(3), 221-231.

Kolk, H., & Heeschen, C. (1992). Agrammatism, paragrammatism and the management of language. *Language and Cognitive Processes*, *7*(2), 89–129. http://doi.org/10.1080/01690969208409381

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

Marshall, R. S., Laures-Gore, J., & Love, K. (2018). Brief mindfulness meditation group training in aphasia: exploring attention, language and psychophysiological outcomes. *International Journal of Language & Communication Disorders, 53*(1), 40-54. https://doi.org/10.1111/1460-6984.12325

Massaro, M., & Tompkins, C. A. (1994). Feature analysis for treatment of communication disorders in traumatically brain-injured patients: An efficacy study. *Clinical aphasiology, 22*, 245-256.

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. *J Speech Lang Hear Res, 59*(5), 1111-1122. https://doi.org/10.1044/2016_jslhr-l-15-0303

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology, 20*(6), 559-572.

Moineau, S., Dronkers, N. F., & Bates, E. (2005). Exploring the processing continuum of single-word comprehension in aphasia. *J Speech Lang Hear Res, 48*(4), 884-896. https://doi.org/10.1044/1092-4388(2005/061)

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

Murray, L. L. (2012). Attention and Other Cognitive Deficits in Aphasia: Presence and Relation to Language and Communication Measures. *American Journal of Speech-Language Pathology, 21*(2), S51-S64. https://doi.org/10.1044/1058-0360(2012/11-0067)

Neto, B., & Santos, M. E. (2012). Language after aphasia: Only a matter of speed processing? *Aphasiology, 26*(11), 1352-1361. https://doi.org/10.1080/02687038.2012.672023

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research, 36*(2), 338-350.

Oh, S. J., Eom, B., Park, C., & Sung, J. E. (2016). Treatment Efficacy of Semantic Feature Analyses for Persons with Aphasia: Evidence from Meta-Analyses. *Commun Sci Disord, 21*(2), 310-323. https://doi.org/https://doi.org/10.12963/csd.16312

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology, 24*(9), 971-990.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Prather, P., Zurif, E., Love, T., & Brownell, H. (1997). Speed of lexical activation in nonfluent Broca's aphasia and fluent Wernicke's aphasia. *Brain and Language, 59*(3), 391-411.

Quique, Y., Evans, W. S., & Dickey, M. W. (2018). 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

Rapp, B. (2011). Case series in cognitive neuropsychology: Promise, perils, and proper perspective. *Cognitive Neuropsychology*, *28*(7), 435–444. http://doi.org/10.1080/02643294.2012.697453

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108.

Rider, J. D., Wright, H. H., Marshall, R. C., & Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *American Journal of Speech-Language Pathology, 17*(2), 161-172.

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.

Rogalski, Y., & Edmonds, L. A. (2008). Attentive Reading and Constrained Summarisation (ARCS) treatment in primary progressive aphasia: A case study. *Aphasiology, 22*(7-8), 763-775. https://doi.org/10.1080/02687030701803796

Schuchard, J., & Middleton, E. L. (2018). The roles of retrieval practice versus errorless learning in strengthening lexical access in aphasia. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1700-1717. https://doi:10.1044/2018_JSLHR-L-17-0352

Silkes, J. P., Fergadiotis, G., Graue, K., & Kendall, D. L. (2020). Effects of Phonomotor Therapy and Semantic Feature Analysis on Discourse Production. *American Journal of Speech-Language Pathology*, 1-14. https://doi.org/https://doi.org/10.1044/2020_AJSLP-19-00111

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

Simmons-Mackie, N. (2018). *Aphasia in North America*. https://www.aphasiaaccess.org/state-of-aphasia

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

Starns, J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25*(2), 377-390.

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

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

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

Touron, D. R., Swaim, E. T., & Hertzog, C. (2007). Moderation of Older Adults' Retrieval Reluctance Through Task Instructions and Monetary Incentives. *The Journals of Gerontology: Series B, 62*(3), P149-P155. https://doi.org/10.1093/geronb/62.3.P149

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A Diffusion Model Account of Criterion Shifts in the Lexical Decision Task. *J Mem Lang, 58*(1), 140-159. https://doi.org/10.1016/j.jml.2007.04.006

Wallace, S. E., & Kimelman, M. D. (2013). Generalization of word retrieval following semantic feature treatment. *Neurorehabilitation, 32*(4), 899-913.

Wambaugh, J. L., & Ferguson, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research & Development, 44*(3).

Whitney, J. L., & Goldstein, H. (1989). Using self-monitoring to reduce disfluencies in speakers with mild aphasia. *J Speech Hear Disord, 54*(4), 576-586. https://doi.org/10.1044/jshd.5404.576

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67-85. https://doi.org/10.1016/0001-6918(77)90012-9

Description of appendices:

Appendix 1. Detailed description of clinician-provided BEARS metacognitive training

Appendix 2. Bayesian model fitting procedures and full model results

Description of supplementary materials:

S1. by-participant detailed description of individualized BEARS training.

S2. additional language testing results

S3a. Pass rate (proportion of "pass" responses relative to paraphasia and nonresponse error types) by session. P1, P5, and P7 were the three participants who responded least to the treatment in terms of naming probe accuracy.

S3b. Practice efficiency during treatment over time, as measured by cumulative feedback points by session.

Appendix 1: Detailed description of clinician-provided BEARS metacognitive training

1. General education regarding balancing effort, accuracy, and response speed. To begin, all participants were educated on the speed-accuracy tradeoff at the beginning of treatment. A figure similar to Figure 2 (but abstracted with the specific numbers removed) was used to supplement the discussion of speed and accuracy. Education, explaining the tradeoff between speed and accuracy and the eventual plateau of accuracy even with increasing processing time, was provided. Participants were introduced to the possibility of adjusting word-finding speed based on perceived difficulty/accessibility of each target. Participants were also taught to say "pass" or "move on" once they decided that the target word was not accessible in the moment. Additionally, they were instructed to only respond with one word in a single attempt. Cues to provide only one attempt were provided throughout treatment. Proper use of speed-accuracy tradeoff was explicitly reinforced for participants throughout treatment. Each participant received an individualized meta-cognitive portion of the treatment, in which both speed and effort were addressed. Clinician judgement was used to determine which meta-cognitive strategies in each category were most useful for an individual's treatment.

2. Self-monitoring tension and frustration and learning when to relax. The clinician subjectively monitored participants' muscle tension, body language, vocal quality, and prosody as proxy measures of retrieval effort, especially when noted in conjunction with unsuccessful word retrieval. If a participant appeared to be using too much effort in an attempt to access the target word, the clinician would draw the participant's attention to the feeling of effort and tension and recommended strategies to reduce the tension and frustration. The clinician would encourage the participant to notice when they were starting to get tense and frustrated, and to take a breath and/or deliberately relax tension in their neck and shoulders.

3. Analogies to understand anomia and speed-accuracy tradeoffs.

• ***Example analogy 1***. Imagine a junk drawer in your garage that is filled to the brim with tools. If you want to get the screwdriver out of the drawer and just grab for something without taking your time, you will likely select a different tool like a hammer or a wrench. Alternatively, if you take just a little bit more time, you could easily select the screwdriver. Similarly, if you move too quickly to select a target word, it is more likely that you will select the wrong word or sounds. You need to take your time so that you are sure you have "the right tool".

• ***Example analogy 2***. Tension and processing interference were likened to creating radio static in the head. Participants were told to "adjust the radio dial" to make the single clearer by relaxing and allowing for more efficient access.

• ***Example analogy 3***. Building a case vs. making a snap judgement. The clinician highlighted the differences of building a case and making an immediate decision when you're in the moment. Naming a picture would be an immediate decision. If long delays were observed, the clinician would remind the participant that it is not about building a case, but rather making that quick decision in the moment.

4. Not waiting too long. If a participant was waiting more than 15 seconds, the clinician would provide education that increased wait time beyond a certain point will likely not result in successfully access of the target word. Additionally, the clinician would highlight the relationship between increased wait times and increased tension and frustration. The participant would be encouraged to "move on" as soon as they felt that they were having trouble accessing the target word. The clinician would lead the participant through discussion and reflection on what it feels like for them when they cannot access the target word in that moment. This feeling would be contrasted with scenarios where the participant was able to access the target word within 1-10 seconds so they could learn when to "move on."

5. Modulating speed based on retrieval difficulty. If a participant began to overgeneralize speed (always slowing down or speeding up regardless of accuracy performance), the clinician would encourage them to modulate their speed based on perceived difficulty. For example, they may be encouraged to give themselves additional time to access a polysyllabic word such as "sledgehammer" as compared to a monosyllabic word such as "bit." Additionally, in these instances, they would be encouraged to slow down their articulation rate and give themselves the time to say the word accurately.

6. Identifying the *right* amount of effort. Some participants were engaged in a discussion about using the "*right* amount of effort." "Too much effort" can result in visible tension and participants may feel as though they are trying to "push out" a word. Alternatively, participants were encouraged to be sure they were engaged with enough effort so as not to "operate on cruise control." After a large number of trials, participants would occasionally repeat the last word they heard in the semantic feature verification question or the previous target. They were encouraged to monitor the amount of cognitive resources they were devoting to the task.

7. Working *with* your aphasia. Education was provided regarding the concept of "working *with* your aphasia" instead of against it. Discussion surrounded the idea that the participant will likely still experience word finding difficulty following treatment. However, responding adaptively will help avoid exacerbating those difficulties. They were encouraged to use strategies to manage their frustration, effort, and timing to increase the likelihood of accessing the target word.

8. Distinguishing initiation speed vs. articulation rate. Participants were encouraged to name pictures quickly, which was occasionally led to increased articulation rate as well. Since increasing articulation rate often resulted in additional sound errors, participants were instructed to give themselves time to say the word accurately once they had accessed it and to not overly speed up their articulation rate.

9. Noting differences between conversation and the treatment task. Some participants questioned the benefits of the training to "pass" or "move on" or were noted to be using this approach to some extent during conversation. Here, education focused on the benefits of identifying when word retrieval was going to be unsuccessful during both treatment and in conversation. During anomia treatment, participants should "move on" to keep practicing, while during conversation, this could be the same point to shift towards alternative communication strategies (e.g., writing the word instead of saying it).

10. Wait until you have one word. Encouraging speed was often met with semantic paraphasias after a number of successive trials. When this occurred, participants were encouraged to slow down and only produce one word to name the target. This helped to reduce the number of immediate self-corrections.

Table A1. Summary of the BEARS metacognitive training strategies used with each participant.

	p1	p2	p3	p4	p5	p6	p7	p8	p9
1. General education regarding BEARS	x	x	x	x	x	x	x	x	x
2. Self-monitoring tension/frustration and learning to relax	x	x	x	x		x	x	x	x
3. Analogies for anomia and speed-accuracy tradeoffs	x		x	x		x	x		
4. Not waiting too long					x		x		
5. Modulating speed based on retrieval difficulty			x	x		x		x	x
6. Identifying the *right* amount of effort	x						x	x	x
7. Working *with* your aphasia	x	x	x			x	x	x	x
8. Distinguishing initiation speed vs. articulation rate.	x	x				x		x	x
9. Differences between conversation and treatment.				x			x		x
10. Wait until you have one word					x	x		x	x

Appendix 2: Model Fitting Procedures and Full Model Results

For predictions 1, 2, and 3, group-level performance was evaluated using Bayesian generalized linear mixed-effect models using the R package BRMS (Bürkner, 2017) following the interrupted time series approach described by Huitema and McKean (2000) and Moeyaert et al. (2017). Models were implemented separately for each item condition (i.e., treated, related untreated, unrelated untreated). The proportion of correct naming attempts was modeled using a binomial distribution via a logistic link function. Reward rate and "pass" rate were modeled using a lognormal distribution with a gaussian link function.

For each dependent variable (i.e., accuracy, reward rate, and "pass" rate) and item condition, a maximal model with random intercepts by participant and correlated random slopes for each predictor variable was initially fit. Each model was then reduced iteratively on the basis of posterior predictive checks, leave-one-out cross validation, and model convergence (in this case eliminating the number of divergent transitions). Because visualization of the aggregate data indicated that response to treatment may be nonlinear, a quadratic term for baseline session and slope change were also evaluated and only maintained in a given model if they significantly improved model fit. Differences between item conditions were determined by evaluating whether 90% credible intervals overlapped between conditions.

As reward rate is a combination of a count variable (the number of correct responses) and response time, the model was fit in an iterative fashion with multiple probability distributions, including the lognormal, truncated normal, and gamma distributions, with the lognormal distribution showing the best fit. Model fit for each of these probability distributions was compared via posterior predictive checks and leave-one-out cross validation.

For naming accuracy, models were evaluated for overdispersion via simulation using the R package DHARMa (Hartig, 2020); overdispersion was not present ($p > .05$). 4000 iterations were run for each of four independent Hamiltonian Markov Chain Monte Carlo; the initial 2000 chains were discarded and not included in the estimation of each parameter. Models were run with weakly informative priors: normal distributions with a mean of 0 and standard deviation of 10 for beta coefficients and a half-cauchy distribution with a mean of 0 and standard deviation of 5. Models were assessed for convergence using the split-half potential scale reduction factor (Gelman et al., 2013) and the effective sample size. In all the models the estimated split-half potential scale reduction factor values were less than 1.01, and the number of effective sample sizes exceeded 1000 for all parameters. Posterior predictive checks confirmed the models adequately fit the data.

Table A1. Naming Probe Accuracy Bayesian General Mixed-Effect Model Results

Model	Parameter	Mean	Std. Error	Lower CI	Upper CI
Treated	*Population level effects*				
	Intercept	-1.89	0.26	-2.32	-1.48
	Baseline Slope	0.2	0.06	0.10	0.29
	Level Change	0.56	0.26	0.15	0.98
	Slope Change	0.03	0.08	-0.10	0.15
	Slope $Change^2$	-0.02	0.01	-0.03	-0.01
	Group level effects				
	sd: Intercept	0.56	0.24	0.27	0.98
	sd: Baseline Slope	0.08	0.04	0.02	0.15
	sd: Level Change	0.51	0.26	0.16	0.97
	sd: Slope Change	0.05	0.04	0.00	0.13
Related	*Population level effects*				
	Intercept	-0.84	0.34	-1.40	-0.30
	Baseline Slope	0.06	0.07	-0.05	0.16
	Level Change	0.08	0.22	-0.28	0.45
	Slope Change	-0.01	0.07	-0.12	0.10
	Group level effects				
	sd: Intercept	0.84	0.29	0.48	1.38
	sd: Baseline Slope	0.04	0.03	0.00	0.09
	sd: Level Change	0.17	0.14	0.01	0.45
	sd: Slope Change	0.04	0.03	0.00	0.10
Unrelated	*Population level effects*				
	Intercept	-1.34	0.33	-1.89	-0.82
	Baseline Slope	0.19	0.07	0.08	0.30
	Level Change	-0.28	0.23	-0.64	0.10
	Slope Change	-0.17	0.07	-0.28	-0.05
	Group level effects				
	sd: Intercept	0.79	0.29	0.44	1.33
	sd: Baseline Slope	0.05	0.03	0.01	0.10
	sd: Level Change	0.21	0.17	0.02	0.54
	sd: Slope Change	0.04	0.03	0.00	0.10

Note: Beta-coefficients are presented as log odds. CI refers to the 90% credible interval.

Table A2. Naming Probe Reward Rate Bayesian General Mixed-Effect Model Results

Model	Parameter	Mean	Std. Error	Lower CI	Upper CI
Treated	*Population level effects*				
	Intercept	-0.17	0.27	-0.60	0.26
	Baseline Slope	0.22	0.04	0.15	0.29
	Level Change	0.49	0.21	0.17	0.83
	Slope Change	-0.15	0.04	-0.22	-0.09
	Group level effects				
	sd: Intercept	0.72	0.25	0.43	1.17
	sd: Baseline Slope	0.04	0.03	0.01	0.09
	sd: Level Change	0.48	0.21	0.20	0.85
	sd: Slope Change	0.03	0.03	0.00	0.08
Related	*Population level effects*				
	Intercept	0.53	0.36	-0.05	1.10
	Baseline Slope	0.16	0.04	0.08	0.23
	Level Change	-0.03	0.15	-0.27	0.22
	Slope Change	-0.09	0.05	-0.17	-0.02
	Group level effects				
	sd: Intercept	0.95	0.32	0.59	1.53
	sd: Baseline Slope	0.03	0.02	0.00	0.07
	sd: Level Change	0.14	0.11	0.01	0.35
	sd: Slope Change	0.03	0.02	0.00	0.07
Unrelated	*Population level effects*				
	Intercept	0.21	0.32	-0.30	0.72
	Baseline Slope	0.27	0.04	0.20	0.33
	Level Change	-0.32	0.17	-0.60	-0.04
	Slope Change	-0.22	0.04	-0.29	-0.15
	Group level effects				
	sd: Intercept	0.86	0.28	0.52	1.40
	sd: Baseline Slope	0.03	0.02	0.00	0.06
	sd: Level Change	0.32	0.16	0.08	0.61
	sd: Slope Change	0.03	0.02	0.00	0.07

Table A3. Naming Probe "Pass" Rate Bayesian General Mixed-Effect Model Results

Model	Parameter	Mean	Std. Error	Lower CI	Upper CI
Treated	*Population level effects*				
	Intercept	-3.01	0.62	-4.06	-2.04
	Baseline Slope	-0.14	0.13	-0.37	0.07
	Level Change	1.73	0.78	0.46	3.01
	Slope Change	0.55	0.18	0.25	0.85
	Slope $Change^2$	-0.03	0.02	-0.06	-0.01
	Group level effects				
	sd: Intercept	1.38	0.53	0.72	2.35
	sd: Baseline Slope	0.08	0.07	0.01	0.22
	sd: Level Change	1.85	0.63	1.05	2.99
	sd: Slope Change	0.20	0.13	0.03	0.44
	sd: Slope $Change^2$	0.03	0.02	0.01	0.06
Related	*Population level effects*				
	Intercept	-2.36	0.50	-3.20	-1.58
	Baseline Slope	0.01	0.13	-0.20	0.22
	Level Change	1.13	0.53	0.28	2.02
	Slope Change	0.14	0.17	-0.14	0.41
	Slope $Change^2$	-0.01	0.01	-0.03	0.01
	Group level effects				
	sd: Intercept	0.88	0.41	0.34	1.64
	sd: Baseline Slope	0.06	0.05	0.00	0.15
	sd: Level Change	0.92	0.46	0.31	1.77
	sd: Slope Change	0.07	0.06	0.01	0.19
	sd: Slope $Change^2$	0.01	0.01	0.00	0.03
Unrelated	*Population level effects*				
	Intercept	-3.46	0.66	-4.59	-2.45
	Baseline Slope	-0.06	0.20	-0.39	0.27
	Level Change	1.70	0.73	0.55	2.94
	Slope Change	0.51	0.24	0.12	0.91
	Slope $Change^2$	-0.04	0.02	-0.07	-0.02
	Group level effects				
	sd: Intercept	0.79	0.51	0.11	1.71
	sd: Baseline Slope	0.10	0.08	0.01	0.24
	sd: Level Change	0.90	0.51	0.18	1.84
	sd: Slope Change	0.14	0.12	0.01	0.35
	sd: Slope $Change^2$	0.02	0.01	0.01	0.05

Supplementary materials S1: by-participant description of individualized BEARS training.

Participant 1 consistently benefited from cues to "Let the dust settle" (i.e., to allow for overactive lexical information in short-term memory to decay). This was because after multiple trials, he was noted to become perseverative, especially when he responded too quickly. Discussed "tool drawer" analogy (see Appendix 1). He was also observed to either build up too much tension prior to retrieving the target or move on too quickly without enough time for an attempt. Therefore, additional education was provided regarding balancing the right amount of time with the right amount of effort. He was cued to "breathe and slow down" as opposed to always saying "move on." Participant 1 also benefited from discussion about understanding and working *with* your aphasia. Discussed his apparent maladaptive habits of trying to force a word out through increased tension or trying to "sneak up on it" by trying to produce it as quickly as possible (which often resulted in errors), as opposed to a more adaptive approach of easing into word retrieval. Additional feedback regarding his level of initiation speed and apparent tension was provided as needed.

Participant 2 had a tendency to make impulsive errors and benefited from cues to "breathe and slow down." Frequent tension feedback was provided when he became tense after missing multiple targets consecutively. Tension and frustration appeared to be his biggest battle. He benefited from discussion about understanding and working *with* your aphasia. Also discussed easing into retrieving a word rather than trying to "spit it out."

Participant 3 benefited from education regarding how to modulate his naming speed. He was encouraged to slow down when he saw a picture that he knew gave him more difficulty. Discussed putting in the right amount of effort and being aware of his output (tool drawer analogy). He was encouraged to slow down if he found himself coming up with other treatment targets (i.e. visor for eggplant; both treated items). He was noted to become more aware of his output in the treatment game but not as much in conversation with the clinician. As a treatment session went on, he benefited from additional cues to breathe and relax, as tension would build up if he began missing treatment targets he could typically access. Of note, when increased difficulty was noted in a treatment session, it was frequently in conjunction with self-reports of poor sleep or stress in his personal life. Upon final follow up, participant's wife reported increased confidence. Participant reported that he was now ordering his own food at the deli counter without difficulty. Per wife, both of his children also reported noticeable improvements in his language.

Participant 4 was good at providing his own life examples and analogies in relation to his experience of aphasia. He required a lot of cues and education to learn when to move on when a retrieval attempt was not going to be successful and when to give himself more time when he had a good chance of succeeding. Increased naming difficulty was noted during step 3 of each trial (after questions) compared to other participants. He was noted to be impulsive during his second attempts, frequently producing a word within the category but not the target item. He benefited from frequent cues to relax, as frustration was a barrier for his naming. Education was provided regarding the differences between intensive drill-based practice to improve his language system vs. typical total communication. He was encouraged to continue to use circumlocution strategies in his day-to-day life, but to only provide one word during the treatment game. Upon exit, participant reported that he did not feel like he improved much in his speech/language but really enjoyed the treatment sessions.

Participant 5 required multiple and frequent cues to provide a single word to name the item, as he provided constant commentary about his performance and was frequently tangential (e.g., sharing stories about his own life). Commentary and tangential stories were considerably reduced over the course of treatment. Frequent encouragement was provided to increase speed for semantic feature questions, as he often got distracted by commenting on the question or by responding quickly verbally but not selecting the button. Commentary was reduced with training, but speed did not seem to improve. He was encouraged to remain silent until he decided to attempt the word or move on. Noted improvement was made on this skill throughout treatment. He had a tendency to attempt to name a target word a number of times, which appeared to increase his chance of subsequent perseverative errors. He was instructed to name the picture only one time, and his ability to do so improved over the course of treatment. He appeared to benefit more from concrete examples and practice rather than from abstract analogies. He seemed to improve his ability to wait until he was ready before attempting to say the word. Tangential behavior (commentary and stories) were noted to increase with fatigue. Per wife, both his sister and brother-in-law commented on how much his language improved. Wife reported that his language was the best it has been since his stroke, and that he started to initiate more conversation in the car as opposed to sitting quietly.

Participant 6 benefitted significantly from cues to "relax" (i.e. to self-monitor level of tension in her shoulders). She was encouraged to slow down, the difference between feeling rushed vs. not rushed as she attempted to retrieve a word. Overall, she was very motivated to earn feedback points and get through a many treatment trials. She was encouraged to self-monitor her output for sound errors. She was encouraged to use the "most efficient road" for word retrieval (natural access vs. compensatory strategies) during the treatment game. This was because she occasionally attempted to visualize and then read the spelling of a word in her head, which took longer and resulted in sound errors for irregularly spelled words. She was educated about modulating her speed as needed, and that she did not need to slow down for every picture, just when she encountered retrieval difficulty. She was also trained to only provide a single naming attempt once she was ready, as she often attempted to retrieve a word too quickly and then produced multiple self-corrected errors.

Participant 7 was encouraged frequently to speed up his overall responses. He became extremely frustrated during more difficult tasks and would often take the maximum amount of time to respond. Used an education strategy related to his former law enforcement experience to illustrate the need to make a quick "gut decision" instead of seeking to "build a case" like a detective. Trained on ways to "clear the static" (i.e. taking a breath when the task became difficult, slowing down, counting to 10 to try to clear instances of perseveration). Encouraged him to be aware of output to reduce perseverations or unrelated responses. Both his wife and participant reported that his day-to-day language performance improved after treatment. Improved speed was evident in overall duration of exit testing.

Participant 8 benefited from education regarding modulating his naming speed. He was encouraged to slow down when he saw a picture that he knew gave him more difficulty.

Discussed putting in the right amount of effort and self-monitoring output. He was encouraged to slow down if he found himself coming up with unrelated treatment targets (i.e. “visor” for “eggplant” which were both treated items). He implemented several strategies to relax and slow down without clinician prompting. Given comprehension impairments, he benefited most from training using direct strategies as opposed to abstract scenarios or analogies. He was very motivated by treatment game feedback. As treatment progressed and his performance accuracy increased, he was instructed to continue to balance speed and accuracy, performing at a faster rate that occasionally resulted in an error instead of slowing down significantly to achieve 100% performance accuracy.

Participant 9 required significant encouragement throughout treatment. Following a few sessions of treatment, his increased awareness of errors and choosing to “move on” was observed to over-generalize to conversation, resulting in reduced output. Therefore, a distinction was drawn between the treatment game and general conversation, and he was encouraged to continue to focus on getting his point across despite sound errors (and to shift to alternative communication strategies, if necessary). Discussed putting in “the right amount” of effort and time before moving past a picture. Obvious tension was observed when he attempted a word before he was “ready.” He was encouraged to “break the tension” by taking a deep breath and sitting back in his chair before trying again. With moderate cueing to complete these techniques, he had good success retrieving his target in instances of increased difficulty.

Supplementary materials S2: Additional language testing results

Additional participant testing: Philadelphia Naming Test

	Correct	Semantic Errors	Formal Errors	Mixed Errors	Unrelated Errors	Nonword Errors	Non-naming Response	S-Weight	P-Weight
Participant 1	72	23	6	4	6	35	29	0.0182	0.0619
Participant 2	121	10	4	4	0	27	9	0.0294	0.0169
Participant 3	151	6	2	5	0	6	5	0.0313	0.025
Participant 4	129	18	5	12	1	5	5	0.02	0.0288
Participant 5	63	8	9	8	3	7	77	0.0176	0.025
Participant 6	117	0	13	1	1	30	13	0.0325	0.0151
Participant 7	62	16	2	3	10	1	81	0.0151	0.0325
Participant 8	103	6	17	3	3	23	20	0.0213	0.02
Participant 9	120	5	14	3	1	22	10	0.0269	0.0182

Additional participant testing: Nicholas and Brookshire, Camel and Cactus, PALPA 25 and PALPA 36

Assessments	Cactus and Camel	PALPA 25		PALPA 36
	Percentage Correct	Words	Non-Words	
Participant 1	87.50%	70%	95%	0
Participant 2	75.00%	100%	83.33%	7
Participant 3	79.69%	96.67%	95%	12
Participant 4	71.89%	78.33%	96.67%	0
Participant 5	75.00%	85%	86.67%	13
Participant 6	92.19%	100%	96.67%	21
Participant 7	64.06%	25%	88.30%	1
Participant 8	84.40%	93.33%	98.33%	18
Participant 9	81.25%	93.33%	91.67%	4

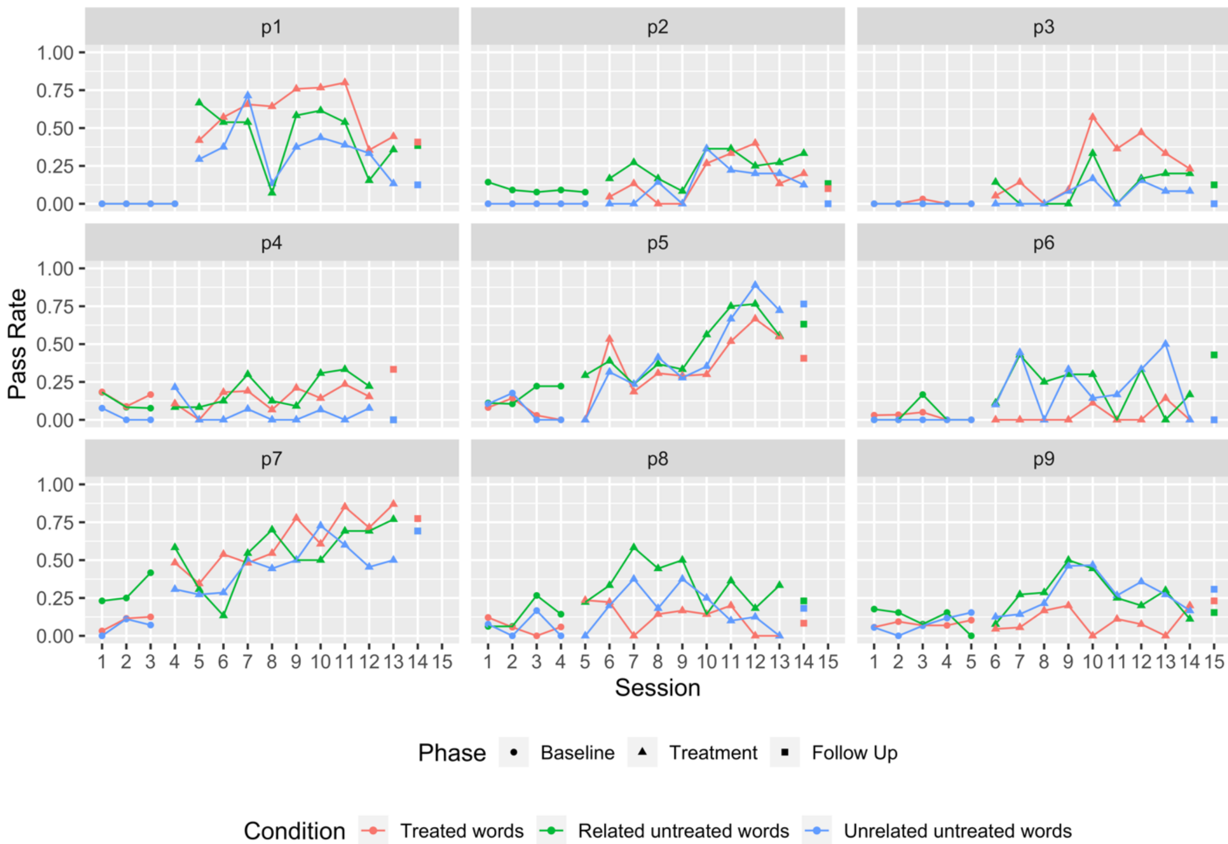

Figure S3a. Pass rate (proportion of "pass" responses relative to paraphasia and nonresponse error types) by session. P1, P5, and P7 were the three participants who responded least to the treatment in terms of naming probe accuracy.

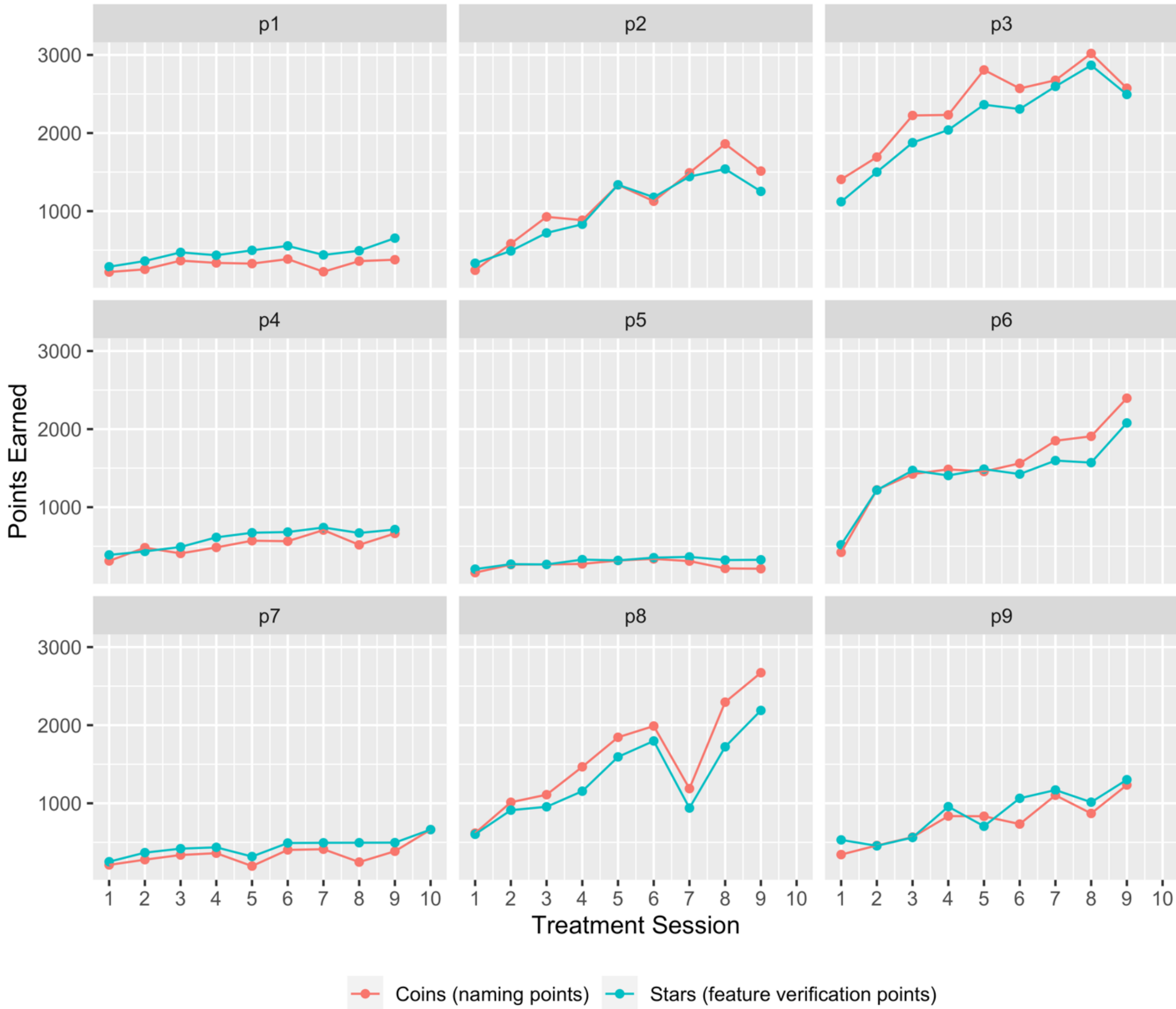

Figure S3b. Practice efficiency during treatment over time, as measured by cumulative feedback points by session.