ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Beyond Percent Correct: Measuring Change in Individual Picture Naming Ability

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

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

ARTICLE INFO

Article History:
Received April 24, 2020
Revision received September 16, 2020
Accepted August 20, 2021

Editor-in-Chief: Stephen M. Camarata
Editor: Mary Alt

https://doi.org/10.1044/2021_JSLHR-20-00205

ABSTRACT

Purpose: Meaningful changes in picture naming responses may be obscured when measuring accuracy instead of quality. A statistic that incorporates information about the severity and nature of impairments may be more sensitive to the effects of treatment.
Method: We analyzed data from repeated administrations of a naming test to 72 participants with stroke aphasia in a clinical trial for anomia therapy. Participants were divided into two groups for analysis to demonstrate replicability. We assessed reliability among response type scores from five raters. We then derived four summary statistics of naming ability and their changes over time for each participant: (a) the standard accuracy measure, (b) an accuracy measure adjusted for item difficulty, (c) an accuracy measure adjusted for item difficulty for specific response types, and (d) a distance measure adjusted for item difficulty for specific response types. While accuracy measures address the likelihood of a correct response, the distance measure reflects that different response types range in their similarity to the target. Model fit was assessed. The frequency of significant improvements and the average magnitude of improvements for each summary statistic were compared between treatment groups and a control group. Effect sizes for each model-based statistic were compared with the effect size for the standard accuracy measure.
Results: Interrater and intrarater reliability were near perfect, on average, though compromised somewhat by phonological-level errors. The effects of treatment were more evident, in terms of both frequency and magnitude, when using the distance measure versus the other accuracy statistics.
Conclusions: Consideration of item difficulty and response types revealed additional effects of treatment on naming scores beyond those observed for the standard accuracy measure. The results support theories that assume naming ability is decomposable into subabilities rather than being monolithic, suggesting new opportunities for measuring treatment outcomes.
Supplemental Material: https://doi.org/10.23641/asha.17019515

Picture naming tasks are frequently used to evaluate speech and language abilities in clinical populations, either as part of a larger language assessment battery or as a standalone assessment of word finding and speech production. Picture naming tasks are also frequently used to derive outcome measures for treatment monitoring or rehabilitation research. While naming accuracy scores have been reliable and useful in practice, informally, clinicians and researchers often note that clinical impressions of progress in verbal communication abilities are not fully captured by changes in naming accuracy scores. Although clients or participants may not necessarily produce more correct responses, the responses may appear to be of higher quality, that is, closer to the target. Formally measuring changes in the quality of naming behavior can open opportunities for intervention and study, but it also poses significant challenges. In this article, we present a statistical approach based on a cognitive

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

model that attempts to address these challenges, and we demonstrate that there are indeed measurable treatment gains that are not captured by accuracy scores.

What Is an Ability, and How Do We Know If It Changes?

A seminal treatise by Lord and Novick (1968), commissioned by the Educational Testing Service, presented a formal, statistical framework for working with test scores. They addressed fundamental questions, such as the following: What is a test score actually measuring? How do we know if it is reliable? How do we know if it is useful? By formally defining an *ability* as a propensity to answer test items correctly (i.e., a probability of success), methods and formal models that had been developed for statistical sampling and estimation could be brought to bear on relating a sample of data (test scores) to a probability estimate (ability). This allowed observed test scores to be decomposed into a "true ability" (or "true score") component and a "measurement noise" component. For any set of sampled data (test responses), we need to consider if the testing procedure is truly reflective of an underlying ability, because many factors can potentially affect test scores beyond the process of interest, such as vigilance level, time of day, item effects, and so on. If the ability *construct* is useful, it will lead to valid inferences based on the relative orderings of individuals; higher abilities should predict better outcomes outside of the specific testing situation.

A fundamental assumption of this framework is that, during a given test administration, ability levels remain approximately stable. If responses reflect a moving target, it becomes much harder to separate item-to-item ability changes from random fluctuations that are unrelated to the underlying ability. When the goal is to measure a change in ability, as is typically the case in rehabilitation research, abilities before and after an intervention are assumed to be approximately stable, thus defining a "true change" in ability, with measurement noise influencing observed test scores at each time point (Cronbach & Furby, 1970). This approach to quantifying abilities and their changes has been extremely popular and productive, but it has its limitations. If there are multiple ways to fail on a test item, then the simple probability of success (i.e., a single ability) does not reflect the full complexity of the underlying process (Embretson, 1997). Likewise, there may be multiple ways that the response process may change, and considering only the distal effects on the probability of correct responses might obscure other meaningful changes.

Picture Naming Behaviors in Aphasia Are Complex

Anomia (word-finding difficulty) is one of the most common long-term consequences of stroke. Clinical evaluations of language processing almost invariably include some form of object naming or picture naming task. However, naming is a deceptively complex process that can be disrupted in many ways, leading to multiple opportunities for different types of errors (Cuetos et al., 2000; Dell et al., 1997, 2004; Fromkin, 1971; Mitchum et al., 1990; Schwartz & Brecher, 2000). That is, because naming requires several different cognitive abilities (e.g., visual analysis, semantic analysis, lexical retrieval, phonological sequencing, and motor execution), problems at one or more processing stages can lead to different types of errors. Furthermore, multiple brain regions, widely distributed throughout the cerebrum, typically work in concert to produce fluent and accurate naming behavior (Giahi-Saravani et al., 2019). Cerebrovascular accidents like stroke do not respect the functional boundaries of the brain when causing damage and can therefore lead to highly diverse patterns of impairment in a clinical population, depending on the location and extent of strokes and the relevant premorbid abilities. On top of this, not all words offer the same error opportunities; some words may be harder to recall, and some words may be harder to pronounce. Thus, there will be an interaction between a given person's deficits and the properties of the items that the person is instructed to name, leading to further complexity in the patterns of test scores obtained during clinical evaluations.

Modeling Subcomponents of Naming

An alternative to using a simple accuracy measure on a naming task is to use error-*type* data to estimate the functional status of subcomponents of the naming process (Mitchum et al., 1990). For example, it is reasonable to assume that a participant who makes mostly semantic errors has a breakdown that is different from a participant who makes mostly phonological errors. Estimating abilities at these subcomponent levels is useful not only theoretically but also clinically, for example, in deciding on a treatment choice targeting semantic versus phonological levels (Abel et al., 2009; Boyle & Coelho, 1995; Leonard et al., 2008). A more detailed measurement of the subcomponents of the process could also be important for measuring change in recovery as well. Clinicians sometimes report that clients seem to be improving despite minimal changes on naming accuracy scores. This could be because improvement at one level of the process (avoiding semantic errors) is partially offset by a drop in performance at another level, which may be a result of increased load due to higher success rates at an early processing stage.

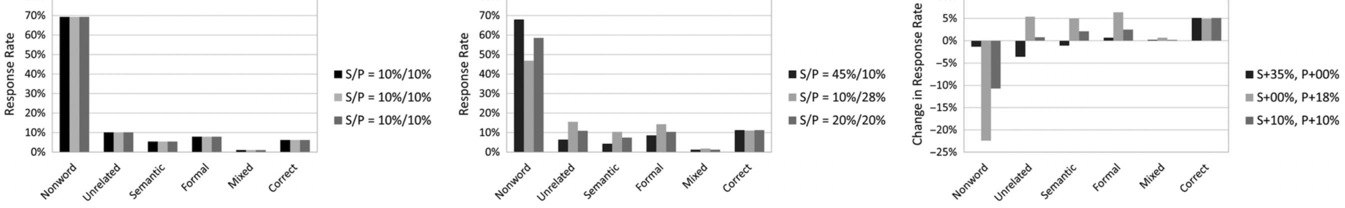
A Spreading Activation Model

Perhaps the most popular approach for estimating subcomponent abilities is based on the Dell et al. (1997) spreading activation model of lexical retrieval. This model consists of a three-layer network of units symbolizing semantic (meaning), lexical (word), and phonological (sound gesture) levels of mental word representations. During word production, activation spreads from the semantic units through the network, and after several timesteps, the most active word unit is selected, receiving its own activation boost. After several further timesteps, the most active phonological units are selected for production. Errors can occur at either selection stage (lexical or phonological) due to the decay of activation over time and the presence of intrinsic noise that is added to the activation on each timestep. In Foygel and Dell (2000), damage to the network was modeled by reducing the connection strengths between semantic and lexical units (S weights) or between lexical and phonological units (P weights), thereby inhibiting the flow of activation, increasing the influence of noise, and causing more errors. This model posits that two distinct categories of damage (S and P) can account for the rates of six different response types (correct and five error types). This supposition has gained support from analyses of patterns of naming errors in people with aphasia: The model provides a reasonable approximation to the observed response type frequencies using just these two parameters (Schwartz et al., 2006). This means that rather than using a single number (percent correct [%C]) to characterize the level of impairment, using the model, we can characterize impairment in two dimensions.

This simplified model can illustrate how considering only the correct response rate obscures the individual effects on subcomponents of naming. Consider three hypothetical participants that are simulated by the model who are identical before receiving any treatment, all having a relatively severe impairment (10% of maximum S and P weights) with a 6% correct response rate. The rates of other response types are shown in the bar graph in Figure 1A. After receiving different targeted therapies, the participants all improved their accuracy scores by 5%. The posttreatment response type rates are shown in Figure 1B, and the changes in each response type rate are shown in Figure 1C. Despite the nearly identical accuracy rates among the participants both before and after treatment, the error type rates reveal clear differences in the relative strengths of the subcomponents of each simulated participant. Posttreatment, the first participant's S weight is 4.5 times stronger than their P weight; the second participant's P weight is 2.8 times stronger than their S weight; and the third participant's S weight and P weight strengths are equal. All three participants increased their accuracy rate by the same amount (5%), but the changes in error types reveal how the individual responses to treatment differed, resulting in different patterns of performance posttreatment. The first participant's S weight increased 35%; the second participant's P weight increased 18%; and the third participant's S and P weights each increased 10%.

While the spreading activation model has illuminated a number of implications for multicomponent theories of naming (i.e., those that assume naming is decomposable into subprocesses rather than being monolithic), it also has its limitations. Because the model was designed for the purpose of theoretical inquiry, to test specific assumptions about interactivity among psycholinguistic representations, it lacks some features that can enhance the reliability of probability estimates and the accuracy of predictions for new data. First, the structure of the model's lexical neighborhood that defines the error opportunities for a given trial is manually hard-wired rather than being learned from real data. Second, differences between items are not considered; the error opportunities on each trial are assumed to be the same. Accounting for these sources of variance in test scores improves ability estimates and predictions of independent data (Walker et al., 2020). Finally, the multiple possibilities for changes in subcomponents still leaves open the question of how to compare the effects of treatment among participants (e.g., a relative ordering of participants' responses to treatment). A psychometric model can provide a complementary approach to modeling naming responses that addresses these limitations.

Figure 1. (A) The bar graph shows pretreatment response rates for three identical, hypothetical participants simulated by a connectionist model. (B) The bar graph shows posttreatment response rates for the same three participants. Although all three have the same correct response rate, their different error type rates reveal differences in their underlying connection strengths (S and P). (C) The bar graph shows changes in response rates from pretreatment to posttreatment for the same three participants. Although all three have the same change in correct response rate, changes in their error type rates reveal differences in their responses to treatment.



A Multinomial Processing Tree Model

Rather than simulating naming with a spreading activation model, Walker et al. (2018) modeled each naming attempt as a probabilistic sequence of successful or unsuccessful latent processes, with each sequence leading to a specific response type. The possible sequences of mental errors or successes leading to each response type on a naming trial can be represented as a binary-branching tree, like a flowchart, and probabilities can be assigned to each path through the tree. This type of statistical model sits on a spectrum between completely atheoretical models that are based on the expected sampling statistics for a given data type regardless of the source (e.g., analysis of variance or χ^2 test) and highly theoretical models that explain specific phenomena based on assumptions about the physical mechanisms that cause variance in the data (e.g., spreading activation). Multinomial processing tree (MPT) models are commonly used in psychological testing paradigms to tease apart latent mental processes (i.e., unobservable decisions, conscious or unconscious, in someone else's mind) that may be differentially contributing to the frequency of different response types (Batchelder & Riefer, 1999; Erdfelder et al., 2009). In addition to formalizing and testing theories of cognition, MPT models have also been useful for psychometric measurement in clinical populations (Batchelder et al., 1997; Batchelder, 1998; Batchelder & Riefer, 2007; Embretson & Yang, 2013; Yang & Embretson, 2007).

The MPT-Naming model (Walker et al., 2018) estimates six abilities (labeled in italics) relevant for picture naming: the ability to initiate an attempt (*Attempt*); the ability to access the semantic neighborhood of the target and avoid totally unrelated words (*Sem*); the ability to avoid words that are either semantically related (*LexSem*), phonologically related (*LexPhon*), or both (*LexSel*); and an ability to avoid sublexical errors in phoneme sequencing and production (*Phon*; see Figure 2). Recall that, in this context, the term *ability* is referring to a participant's probability of taking a rightward branch in the tree diagram, independent of the item's propensity to influence the process one way or the other. The four different lexical abilities are motivated by the concept of *competitive selection,* whereby multiple candidate words are considered concurrently for selection during a naming attempt (Nozari & Hepner, 2019). The model assumes that the probability of a nontarget candidate being rejected depends on whether it is similar to the intended target word on semantic and/or phonological grounds; thus, the four possible relationships define four different abilities to reject four different types of nontarget competitors. Defining *Sem* and *LexSel* independently from *LexPhon* and *LexSem* allows for the possibility of an interaction, whereby both relationships being present (or not) leads to greater or lesser competition (i.e., a different probability of misselection) than would be expected from a simple combination of the competitions experienced when each relationship is present on its own.

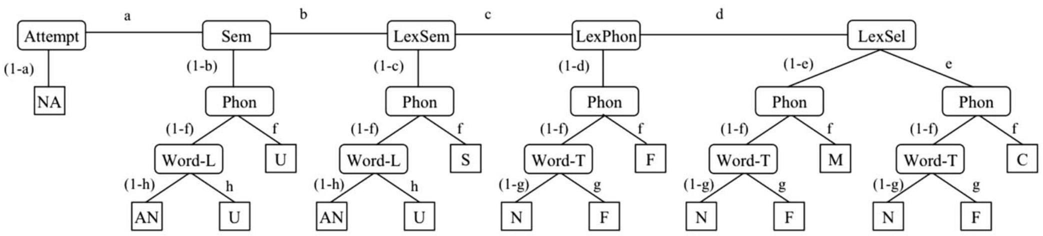
Together, these abilities govern the likelihood of eight possible response types: (a) correct – the response matches the target (i.e., the picture is named correctly), (b) mixed – the response is a real word that is both semantically and phonologically related to the target, (c) formal – the response is a real word that is phonologically related to the target, (d) semantic – the response is a real word that is semantically related to the target, (e) unrelated – the response is a real word that is neither semantically nor phonologically related to the target, (f) neologism – the response is a nonword that is phonologically related to the target, (g) abstruse neologism – the response is a nonword that is not phonologically related to the target, and (h) nonnaming attempt – a miscellaneous category including no responses, descriptions, picture parts, or sound effects. These categories follow the scoring and rationale for the Philadelphia Naming Test (PNT; Roach et al., 1996).

The relationships between response types and psycholinguistic processing levels in the MPT-Naming model are inspired by the two-step assumption found in many contemporary models of lexical access, which proposes an initial competition among word representations for production, followed by a subsequent competition among sublexical (i.e., morphological, syllabic, or phonological) segments for production (Dell et al., 1997; Foygel & Dell, 2000; Lambon Ralph et al., 2002; Levelt, 1989; Matti et al., 1998; Rapp & Goldrick, 2000; Roelofs, 2000). Furthermore, during processing at the lexical level, there is a progressive influence of different information sources, from early semantic influences to later phonological influences (Dell et al., 1997; Indefrey & Levelt, 2004; Mitchum et al., 1990). Finally, because the model is still fundamentally based on the formal concept of probabilities of success, item response theory (IRT) can be applied to account for item effects on each latent mental process; that is, items can be difficult in different ways by challenging different abilities (Embretson (Whitely), 1984; Embretson & Yang, 2006; Maris, 1995). By examining patterns of responses in a large cohort of people with aphasia, Walker et al. (2018, 2020) were able to disambiguate the psycholinguistic sources of errors and estimate the underlying abilities of participants and difficulties of test items, yielding valid inferences about participants (e.g., predicting other test scores) and about test items (e.g., predicting the number of similar-sounding words).

Deriving a Simple Measure From a Complex Model

While multidimensional characterizations of naming deficits clearly provide a richer measure of a participant's

Figure 2. The MPT-Naming model for a trial from a picture naming test, with equations for the probability of each response type (adapted with permission from Walker et al., 2018; Copyright © 2018, American Psychological Association). Leaf nodes represent response types. C = correct; M = mixed; F = formal; S = semantic; U = unrelated; N = neologism; AN = abstruse neologism; NA = nonnaming attempt. Lower case letters a–e represent probabilities of success associated with a latent mental process. These probabilities are determined by the interaction of a participant's ability and an item's difficulty via a logistic equation (not shown). Lower case letters g and h represent probabilities that account for the effects of the phonological neighborhood density of the target (i.e., the number of similar sounding words). The probability of each response type is calculated by multiplying the branches leading from the root node to the response type of interest and summing the results if there are multiple leaf nodes indicated.



$$
\begin{aligned}
\mathrm{Pr(C)} &= a*b*c*d*e*f \\
\mathrm{Pr(S)} &= a*b*(1\text{-}c)*f \\
\mathrm{Pr(F)} &= a*b*c*d*e*(1\text{-}f)*g + \\
&\quad a*b*c*d*(1\text{-}e)*(1\text{-}f)*g + \\
&\quad a*b*c*(1\text{-}d)*f + \\
&\quad a*b*c*(1\text{-}d)*(1\text{-}f)*g \\
\mathrm{Pr(M)} &= a*b*c*d*(1\text{-}e)*f \\
\mathrm{Pr(U)} &= a*b*(1\text{-}c)*(1\text{-}f)*h + \\
&\quad a*(1\text{-}b)*f + \\
&\quad a*(1\text{-}b)*(1\text{-}f)*h \\
\mathrm{Pr(N)} &= a*b*c*c*e*(1\text{-}f)*(1\text{-}g) + \\
&\quad a*b*c*d*(1\text{-}e)*(1\text{-}f)*(1\text{-}g) + \\
&\quad a*b*c*(1\text{-}d)*(1\text{-}f)*(1\text{-}g) \\
\mathrm{Pr(AN)} &= a*b*(1\text{-}c)*(1\text{-}f)*(1\text{-}h) + \\
&\quad a*(1\text{-}b)*(1\text{-}f)*(1\text{-}h) \\
\mathrm{Pr(NA)} &= 1\text{-}a
\end{aligned}
$$

abilities, the added complexity may be unwieldy or impractical for clinical application. For example, if naming requires multiple abilities and some improve while others decline, the overall effect of the changes may be difficult to interpret. A simple summary measure derived from a multidimensional model might, we hypothesized, be able to take advantage of the more detailed characterization of the participant's naming (sub)abilities while retaining the simplicity of a single summary statistic. Thus, we derived two simple and easily interpretable summary measures from the MPT-Naming model, which we then tested against the standard accuracy measure (%C) to determine whether they have improved sensitivity to treatment effects and whether inferences based on them have improved validity.

MPT-P(S): The Joint Probability of Avoiding All Error Types on a Typical Item

The abilities that are estimated using the MPT-Naming model pertain to the mental subprocesses that must be completed during a naming trial; there is no single parameter representing the probability of a correct response on a given item. Instead, this probability is calculated as the joint probability of avoiding all error types (e.g., semantic errors, phonological errors, mixed errors). If the ability scales are fixed relative to an item with average difficulty values for the test, then the resulting joint probability of avoiding all error types is numerically almost (due to stochastic model fitting) identical to %C (i.e., the maximum likelihood estimate of the average probability of a correct response for the test); we call this statistic MPT-P(S), where "MPT" indicates the model type, and "P(S)" indicates the overall probability of success (i.e., a correct response).

MPT-E(D): The Expected Proportion of Successful Latent Processes on a Typical Item

While the MPT-P(S) statistic incorporates error type information into its uncertainty about accuracy estimates and may therefore lead to different inferences than those based on %C, it is still fundamentally estimating a binary construct (i.e., whether a trial is correct or incorrect). However, therapy effects may be subtler than increasing correct responses; the responses after therapy may simply be closer to the target. However, what does it mean for a response to be closer to a target? Lexical measures such as semantic distance (Maki et al., 2004; Pennington et al., 2014; Vigliocco et al., 2002) or phonological distance (Sanders & Chin, 2009) between target and response might be used to address this question, but this approach relies on assumptions about the relative effects of semantic and phonological distances on psychological representation and communicative intelligibility. A functional approach, in which independent judges attempt to guess the intended target based on a response, could also provide a basis for a distance measure, but this approach would require substantially more data collection from the independent judges. Instead, we can leverage the natural ordering of response types provided by the MPT model's architecture.

In essence, we are proposing a type of latent partial credit model for naming. What is meant by *latent* partial credit? Typically, partial credit is assigned to each observed response type. For example, failure to respond is assigned 0 point, a semantic error is assigned 1 point, a minor phonological error is assigned 2 points, a correct response is assigned 3 points, and the final score is a sum of points over all items. This scoring rubric neglects the fact that a phonological error at a later processing stage could be masking a semantic error at an earlier processing stage, thus deserving fewer points. It also neglects the influence that different items have on the probabilities of different error types; for example, phonological errors on some items are less costly than on other items, at least for purposes of functional communication (e.g., *stethoscope* → *steposcope* vs. *cat* → *zat*). Rather than assigning points to each response type, we derive a latent partial credit score, MPT-E(D), by characterizing how far the participant's internal system gets toward producing a correct response on a typical item, based on the pattern of observed errors across different items. Specifically, because a correct response depends on six sequential latent processes, we calculate the expected value for the number of successful latent processes on a target item with average processing difficulties and then divide by six to obtain a proportion. This statistic, MPT-E(D), represents the *expected* (proportional) *distance* traveled down the processing tree toward a correct response.

It should be clear that this metric deviates from the typical use of MPT models, in which the probability parameters are the final estimates of interest. In particular, a claim is being made about the relative value of each error type (e.g., an unrelated word is more costly than a phonologically related nonword). While we do not claim that the relative ordering of response types in our current MPT-Naming model is the best or only way to operationalize a partial credit scale, we do believe that the architecture of the model is well justified in terms of the approximate correspondence between error type opportunities and the relative completeness of cognitive processing during a naming trial (Mitchum et al., 1990; Walker et al., 2018, 2020). To the extent that this belief is correct, and to the extent that standard behavioral therapy improves latent cognitive processing, we expect this measure to be more sensitive to therapy effects than measures that ignore error type data. Why? Consider that, according to two-step theories of lexical access, a hypothetical participant who successfully responds to a semantically focused (or phonologically focused) treatment but has an additional, persisting phonological (or semantic) deficit might have a more pronounced shift in error types than in correct responses. The MPT-E(D) statistic should be sensitive to these types of effects, despite the lack of accuracy changes, as well as being sensitive to combined effects of accuracy and error type improvement. Supplemental Material S1 illustrates the variables, dependencies, and prior assumptions of the longitudinal model used to calculate the MPT-P(S) and MPT-E(D) statistics as a directed acyclic graph. Both statistics can be derived "for free" when fitting the same model.

IRT-P(S): The Probability of a Correct Response on a Typical Item

As a comparison model to further illustrate the impact of the MPT-Naming model's consideration of error type data, we also investigated an IRT model that only incorporates item response functions and Bayesian inference for accuracy estimates without considering error type information. In our Bayesian IRT model of naming accuracy, the probability of a correct response depends on the item being named. The difficulty of the items (i.e., the difficulty for producing a correct response) is estimated based on the naming data of a large, independent cohort of persons with aphasia. For our statistic of interest, IRT-P(S), we use the average difficulty over the test items to convert an estimate of a participant's overall naming ability (i.e., without considering subcomponent abilities) into estimates of a probability of a correct response on a typical item, using a simple IRT function known as a Rasch model or a one-parameter logistic model. This measurement construct is the same target that is represented by the %C and MPT-P(S) statistics, and we therefore expect point estimates of these statistics to be highly correlated; however, because IRT-P(S) incorporates item information into its uncertainty surrounding estimates, like MPT-P(S), the inferences that are made when using the Bayesian interval estimates may differ from inferences made with the other models. Supplemental Material S2 illustrates the variables, dependencies, and prior assumptions of the longitudinal model used to calculate the IRT-P(S) statistic, as a directed acyclic graph.

Purpose of the Current Study

The purpose of the current study was to compare inferences about the effects of standard behavioral therapy in individual participants with aphasia when using MPT model-based statistics, IRT-model-based statistics, or the overt %C rate statistic to measure change. We were interested in whether inferences for individual participants would be different for the three types of measures, and we predicted that the MPT-based measures would differ from the others by being more sensitive to change due to therapy.

In addition to providing validation evidence for our computational model's assumptions and parameter interpretations, the current work also advances the model's potential for individual clinical applications. Statistical models of group effects do not permit statistical conclusions to be made about the effect of therapy on any specific individual's ability. This is a notable concern given the heterogeneity observed in aphasic naming deficits (Nickels & Howard, 1995) and their patterns of recovery (Schwartz & Brecher, 2000): The group average may be a poor estimate of any individual. Additionally, in most nonresearch settings, clinicians must make decisions about whether their individual client is exhibiting progress and ideally support these decisions with objective measures. Thus, a statistical measure of individual change that is more sensitive than current outcome measures (i.e., those based only on correct responses) could provide a valuable tool for both researchers and clinicians.

Method

Data

Participants

We examined archived picture naming data from four separate cohorts of people with aphasia: a model calibration group, two treatment groups, and a control group. Inclusion criteria for all groups were left-hemisphere stroke and aphasia diagnosis without other neurological comorbidities. All participants signed an informed consent form approved by the institutional review boards at each study site. Demographic information for each group of participants is presented in Table 1.

The model calibration group was used to derive independent estimates of item difficulty. Data from this cohort and item difficulty estimates were reported by Walker et al. (2018), based on recorded responses from 365 people with aphasia.

The two treatment groups included a total of 72 participants with aphasia who completed a therapy program at the University of South Carolina and Medical University of South Carolina. The data were collected during an ongoing clinical trial, and we examined data from participants who had completed the initial (i.e., baseline) assessment and the 1-month posttherapy follow-up assessment at the time of analysis. Participants with lower than 2% correct naming at baseline were excluded. The motivation for this exclusion was comparison with the control group, which also excluded participants who were at or near floor performance levels (< 5%) at baseline. The rationale here is that participants who cannot access the lexicon at all may have disruptions that are unrelated to the lexical retrieval system *per se,* and thus any lack of observed change may not be informative about meaningful changes that occur within the lexical retrieval system. The treatment participants were split into two independent groups for analysis and replication based on a change in the assessment protocol, as described below; the first 38 included participants (Treatment Group 1) who performed the naming test 12 times over the course of the study, whereas the subsequent 34 included participants (Treatment Group 2) who performed the naming test 7 times. Almost all treatment participants were at least 12 months poststroke at their initial assessment (Treatment Group 1 median months poststroke = 32; Treatment Group 2 median months poststroke = 18); three participants in Treatment Group 2 were only 10 or 11 months poststroke at baseline assessment.

The clinical trial did not include a matched, untreated control group of people with aphasia; however,

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

Clinical and demographic information	Model calibration group	Treatment Group 1	Treatment Group 2	Control group
Participants (*N*)	365	38	34	24
Sex (F/M)	155/210	13/25	15/19	14/10
Age (yrs.)	60 (22–86)[a]	66 (29–76)	60 (38–80)	69 (37–80)
Education (yrs.)	12 (6–22)[b]	16 (12–20)	16 (12–20)	13 (7–20)
Months poststroke	13 (1–381)[a]	32 (12–241)	18 (10–99)	14 (8–29)
Handedness (A/L/R)	NA	1/4/33	1/3/30	NA
Race:				
African American	124 (34%)	10 (26%)	7 (21%)	7 (29%)
Asian	2 (0.5%)	0 (0%)	1 (3%)	0 (0%)
Caucasian	225 (62%)	28 (78%)	26 (76%)	17 (71%)
NA	14 (4%)	0 (0%)	0	0 (0%)
Speech motor deficit (*N*)	90 (25%)	25 (66%)	19 (63%)	1 (4%)
WAB-AQ	75 (16–98)[c]	62 (25–93)	65 (28–93)	NA
Aphasia type:				
Anomia	153 (42%)	9 (24%)	9 (26%)	5 (21%)
Broca's	94 (26%)	16 (42%)	12 (35%)	1 (4%)
Conduction	58 (16%)	8 (21%)	10 (29%)	9 (38%)
Global	8 (2%)	1 (3%)	1 (3%)	0 (0%)
Transcortical motor	3 (1%)	1 (3%)	1 (3%)	0 (0%)
Transcortical sensory	5 (1%)	0 (0%)	0 (0%)	1 (4%)
Wernicke's	44 (12%)	3 (8%)	1 (3%)	8 (33%)
Baseline PNT %C	62% (0%–98%)	51% (2%–97%)	51% (3%–98%)	78% (5%–98%)

Note. The median of continuous measures is reported with the range in parentheses. The proportion of the group is reported in parentheses for discrete measures. Treatment groups were reduced by excluding participants with less than 2% baseline PNT accuracy for comparison with the control group. F = female; M = male; yrs. = years; A = ambidextrous; L = left-handed; R = right-handed; WAB-AQ = Western Aphasia Battery–Aphasia Quotient (Kertesz, 2007); PNT = Philadelphia Naming Test; NA = not applicable.

[a] *N* = 351. [b] *N* = 240. [c] *N* = 271.

independent, longitudinal data from people with aphasia attempting to name the same test items were available from the Moss Aphasia Psycholinguistic Project Database (Mirman et al., 2010). While these data are different than the data used in the model calibration group, the participants represent a subset of this group (i.e., the data were collected at a different time). First, 28 participants with at least four or five PNTs in the database that were administered in consecutive, 12-week intervals were identified (see Supplemental Material S3 for database identifiers, demographic information, and naming data for individual participants). Data from 15 of these participants were reported in Schwartz and Brecher (2000). These participants were examined as part of a longitudinal investigation of spontaneous recovery during the first year-and-a-half poststroke, and each participant performed a single PNT every 3 months (12 weeks) for 1 year (five total PNTs) or until they withdrew from the study. Participants self-reported that they were not participating in therapy during the study. To ensure the relative stability of naming abilities in the control group over time, four participants with greater than 10% change in accuracy between their final two PNTs were excluded to yield a group-average change of 3.2%, which was not significantly different from zero (paired t test, two-tail p = .058), whereas a Bayes factor (BF_{10} = 1.6) indicated weak support for a mild, nonzero change over time (Rouder et al., 2009). The 12-week interval between test administrations in the control group approximated the 12-week interval encompassing naming assessments and therapy delivery in the treatment study. The months poststroke at the earlier PNT in the control group ranged 8–29 months, with a median of 14 months, meaning that the control participants were generally at an earlier point in their course of recovery than the treatment participants, though still considered to be in the chronic phase.

In Treatment Group 1, there were 25 participants (66%) with symptoms of a motor speech disorder (apraxia or dysarthria) affecting speech planning or articulator weakness or both; in Treatment Group 2, there were 19 participants (63%) with symptoms of a motor speech disorder. Notably, these are higher prevalences than in the item calibration group (25%) or the control group (4%).

Therapy Protocol

We refer to this study as the POLAR (Predicting Outcomes in Language Rehabilitation) trial. The purpose of this trial is to determine predictors of treated recovery in a large cohort of individuals with chronic aphasia (i.e., 12 months poststroke) following standard aphasia therapy. Participants each receive semantically focused and phonologically focused therapies in counterbalanced order, with each therapy regimen lasting 3 weeks. There is a 4-week interim between each therapy regimen. Semantically focused treatment regimens included (a) semantic feature analysis (Boyle, 2004; Boyle & Coelho, 1995; Coelho et al., 2000; Wambaugh & Ferguson, 2007), (b) the semantic barrier task (Davis, 2005, 2007; Pulvermüller et al., 2001), and (c) verb network strengthening treatment (Boo & Rose, 2011; Edmonds & Babb, 2011; Edmonds et al., 2014). Phonologically focused treatment regimens included (a) phonological components analysis (Leonard et al., 2008), (b) phonomotor treatment (Kendall et al., 2015), and (c) phonological judgment (Howard et al., 1985; Raymer et al., 1993). Details regarding the treatment protocol and fidelity assessments can be found in Spell et al. (2020).

Naming Assessment

The PNT includes 175 black-and-white line drawings of common nouns. Images are presented one at a time on a computer screen, and participants are instructed to use a single word to name the picture. A maximum of 30 s is allowed for responses. The same order of PNT items is presented to each participant; however, that order was randomized initially. Item difficulty estimates should account for influences of previous items on error probabilities because the difficulty estimates are derived from the same fixed order.

All language samples were recorded; the first complete naming attempt was transcribed; and responses were classified into the eight categories by graduate student research assistants in the Communication Sciences and Disorders Department at the University of South Carolina, who were supervised by an American Speech-Language-Hearing Association–certified speech-language pathologist. The optional lenient scoring rubric for motoric impairments was not applied (Schwartz et al., 2006). This scoring option allows for a single phoneme substitution, omission, or transposition when evaluating the phonological relatedness of responses from participants with motor speech disorders. The current MPT-Naming model does not distinguish between phonological and motor speech errors; they are both considered to be sublexical errors.

In the POLAR trial, assessment measures (including the PNT) are obtained at baseline, immediately following the first round of therapy, prior to initiating the second round of therapy, immediately following the second round of therapy, and at 1 and 6 months following the completion of therapy. Participants in Therapy Group 1 were administered two PNTs during each of these six testing sessions. Due to an assessment protocol change instituted to reduce the burden of scoring on examiners, participants in Therapy Group 2 were administered the PNT twice at the initial (baseline) testing session and then only once at the subsequent testing sessions. When testing sessions included pairs of naming tests, the goal was to have a span of approximately 24 hr between the test administrations. Despite administering nearly 700 PNTs to the treatment participants, for the purposes of this study, we were only interested in the single PNT immediately preceding and the single PNT immediately following the combined therapy phases (Weeks 1 and 12) for each participant (144 total PNTs), to match the amount and approximate timing of data collected from each control participant and to exclude novelty effects from the first test administration. In the control group, the final two PNTs in the longitudinal series were examined (i.e., the fourth and fifth tests), because these represented the plateau of spontaneous recovery as participants entered the chronic phase of the disorder.

Missing Data

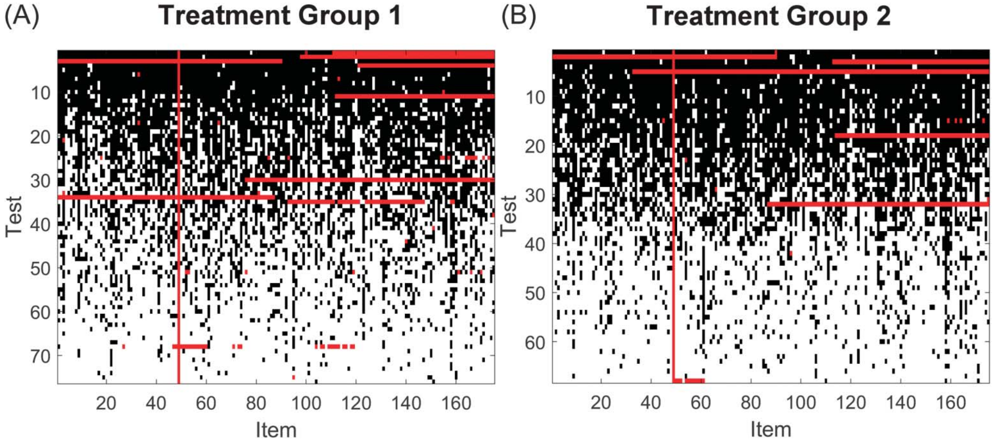
The correct (white), incorrect (black), and missing (red) trials are illustrated for Treatment Group 1 in Figure 3A and for Treatment Group 2 in Figure 3B. There are two tests represented consecutively for each participant, one before treatment and one after. Of the 13,300 total naming trials that were originally planned to be analyzed from Treatment Group 1, 729 trials (5.5%) were treated as missing. The minimum number of observed trials for a participant in this group was 74. There were several causes for missing data. (a) The picture naming test was revised from the version used in previous research, replacing the item *Eskimo* with the item *umbrella*; because item difficulty estimates from the independent cohort were not available for the item *umbrella*, it was excluded from analysis for all participants, accounting for 76 missing trials (10.4%). (b) The remaining 653 missing trials (89.6%) were due either to participant fatigue causing early termination of testing or to recording equipment and data storage malfunctions (the proportions for each of these separate causes were unavailable). Of the 11,900 total naming trials that were originally planned to be analyzed from Treatment Group 2, 534 trials (4.5%) were treated as missing. The minimum number of observed trials for a participant in this group was 32. Causes for missing data were similar to those in the first group. (a) The item *umbrella* was excluded, accounting for 68 missing trials (12.7%). (b) The remaining 466 missing trials (87.3%) were due either to participant fatigue causing early termination of testing or to recording equipment and data storage malfunctions.

The data were missing not at random. In both treatment groups, there was a mild, but significant, negative correlation between the proportion of observed trials that were correct on partial or complete tests and the number of missing trials on those tests, Group 1: $r(74) = -.34$, $p = .003$; Group 2: $r(66) = -.30$, $p = .01$, accounting for 11.5% and 9.0% of the variance in missing trials across tests in each group, respectively. Tests with many error responses typically took longer to administer, leading to greater pressure on local computer memory resources that sometimes failed to record the final portion of the test administration.

Inferences based on fewer observations are likely to be more conservative in identifying significant changes than inferences based on more observations, due to reduced confidence in the estimates (i.e., wider interval estimates). Values for missing data can be imputed as a byproduct of the Gibbs sampling procedure for estimating model parameters, but these imputed values do not affect the behavior of the Gibbs sampling procedure due to the fully specified conditional distributions from which it samples (i.e., the same point and interval estimates for parameters would be obtained if the data structure did not indicate any missing data and instead was simply reduced in size).

Crucially, item difficulty parameters were estimated from an independent calibration cohort without missing data. Using known difficulty values means that ability

Figure 3. A visual representation of the picture naming data, showing each trial as correct (white), incorrect (black), or missing (red), for (A) Treatment Group 1 and (B) Treatment Group 2. Each row represents an individual naming test, with two rows per participant. The rows are sorted in order of total accuracy on the test.



estimates in the treatment groups should be less affected by the specific items that are used for testing, which is a central tenet of IRT (Lord & Novick, 1968). For ability parameters, missing data would only bias estimates if the participant's response-generating mechanisms (i.e., whatever aspects of a participant that contribute to the probability of a response) worked differently on those missing trials. However, the mechanisms that created the majority of the missing data, such as recording equipment failures, are not expected to selectively impact trials where the participants' response-generating mechanisms deviate from the propensities estimated from the observed trials. For example, if a participant showed a propensity to make errors on items that were easy, one could infer that even more errors would be made on the missing items that were difficult, assuming the participant's abilities continued to be stable when naming these missing items. Regarding participant fatigue as a mechanism for missing data, a participant could have produced a low rate of correct responses on the initial trials and then terminated testing because the participant expected continued failures. In this case, we can simply assume we are estimating the participant's ability specifically when the participant feels energetic and confident enough to participate, in accordance with the model's assumption of stable ability levels over testing sessions. Recall that missing data are not the same as no response, one of the modeled response type categories, which would likely account for most strategic refusals to answer. Therefore, subsequent analyses do not include any additional corrections for missing data.

Reliability of PNT Scoring

Interrater and intrarater reliability were estimated for five raters who rescored two participants for interrater reliability and two others for intrarater reliability, yielding nine pairs of scored tests in each condition. These two groups of nine participants with aphasia were selected based on a previous scoring of the PNT by the supervisor in order to represent the full range of the severity scale (interrater group: %C average = 57.9%, *SD* = 39.8%, min. = 2.6%, max. = 97.7%; intrarater group: %C average = 31.2%, *SD* = 29.2%, min. = 0.6%, max. = 87.1%).

Because the data were nominal, interrater and intrarater reliability were quantified at the item level with Cohen's unweighted kappa (Cohen, 1960), examining agreement of polytomous scores (i.e., for all eight response types, assuming chance agreement of one out of eight, or 12.5%) and agreement of dichotomous scores (i.e., for each response type, assuming chance agreement of 50%). Kappa values are interpreted according to the scale presented in Ranganathan et al. (2017) and with respect to the recommended cutoff of .6 for identifying problematic agreement. Correlations between the rates of each response type combined across raters and kappa values for polytomous scores were examined to identify response types that are likely to compromise overall agreement within or between raters at the item level.

Additionally, confusion matrices and Jaccard indices were examined for agreement on each response type, as well as for the most likely response types to be confused. For agreements, the Jaccard index is the proportion of all trials scored with a given response type by one rater that were also scored with the same response type by the alternate rater. For confusions, the Jaccard index is the proportion of all trials scored with a given response type by one rater that were confused for another given response type by the alternate rater.

Participant-level agreement (summed over items) was evaluated by examining correlation coefficients of frequency counts for each response type, as well as means and standard deviations of difference scores, between different raters and within the same rater on different occasions.

Model Fitting

Under the null hypothesis testing framework, a statistical model assumes that a data set comes from an uninteresting random source, and we use the data to check if the model is wrong enough to reject it. Under the Bayesian framework, a statistical model assumes that a data set comes from a theoretically motivated process that includes randomness, and we use the data to check if the model is right enough to accept it (at least until a better model comes along). When we fit the model to data, we estimate the values for each participant ability that combine with the known item difficulty values to define a probability distribution over response types that, when sampled from, is most likely to generate the observed data. Formally, this estimation procedure is known as Gibbs sampling. Then, we check three different aspects of the ability estimation procedure. (a) The procedure to derive the estimates involve an element of chance. Do the ability estimates depend on the starting point of the random sampling procedure? This is called a *convergence* check. (b) The procedure to derive estimates also depends on somewhat arbitrary assumptions about the ability values that are expected to be encountered in the population before we observe any data. Do the ability estimates depend on arbitrary prior assumptions? This is called a *sensitivity* check. (c) Given the flexibility of possible model specifications, the estimation procedure is not guaranteed to capture relevant information about a given data set. Can the data be recovered from the ability estimates? This is called a *posterior prediction* or *model fit* check. If the model is deemed satisfactory after these checks, then we can be confident that the ability estimates are capturing reliable information about the data, allowing us to then make credible inferences based on the model.

Details regarding model fitting and model checking are provided in Supplemental Materials S4 and S5.

Making Inferences From Model-Based Measures of Change

With point and interval estimates for the measures of interest in hand, we are able to make inferences about which participants significantly responded to therapy. We wanted to know whether we would infer different numbers of participants with significant changes in the four summary statistics, for inferences made with a given confidence level. A difference in the estimated magnitude (point estimate) of observed change or a difference in our confidence in the estimated magnitude (interval estimate) of observed change can lead to different inferences for individual participants. However, if differences between measures are found, how would we know which outcome has higher validity given that there is no gold standard for identifying true change? Perhaps a measure's increased sensitivity in detecting the response to treatment is just a higher Type I error rate, for example. To answer this question, we employed a control group of similar people with aphasia who were tested multiple times on the same task but who did not have treatment. If a more "sensitive" measure is really just a measure with a higher rate of Type I error, the measure in question should detect "change," that is, exhibit Type I errors, equally in the treatment and control groups. Analogously, if differences in the magnitude of change statistics are due to increased bias rather than increased sensitivity, we expect the bias to be similar for those who did and did not receive therapy, leading to similar magnitudes of change statistics in these different groups (i.e., a loss of specificity). The ability to distinguish between participants who did and did not receive therapy based on change statistics provides evidence that these statistics are capturing real, systematic changes in naming responses that are attributable to the treatment intervention.

Statistical Analysis 1: Treatment Response Frequency

For each participant, we tested whether there was a significant change in each of the four summary statistics from baseline to immediately following therapy (i.e., during the 12-week interval encompassing both types of therapy). Inferences about change in %C were made using a classical null hypothesis significance testing approach, Fisher's exact test (Fisher, 1922), examining differences in proportions of correct responses before and after therapy. Inferences about change in IRT-P(S), MPT-P(S), and MPT-E(D) were based on whether the Bayesian 95% credible interval for the change variable included zero (Kruschke, 2013). We report the number of participants in each group with a significant change in each summary statistic. We compared the proportions of significant individual responders identified by each change statistic in each treatment group versus the control group, as well as in the aggregated treatment groups versus the control group, using Fisher's exact tests.

To apply Fisher's exact test to an individual participant's naming data, we assumed the data could be arranged in a 2 × 2 contingency table. The rows contained the frequencies of correct and incorrect picture naming responses, respectively; the columns contained the frequencies of responses before and after an interval of time, respectively. Fisher's exact test examines the null hypothesis that the underlying rate has not changed during the interval, yielding a *p* value. Performing statistical tests on the frequencies of correct and incorrect responses instead of a proportion correct statistic allows us to account for the effects of missing data on the confidence in our estimates and enables statistical inferences to be made about the effect of therapy on an individual participant. The *p* value for the Fisher exact test was computed for each participant using the MATLAB function *fexact* (Boedigheimer, 2021).

The assumptions of the Fisher exact test are that each observation comes from a pair of nominal variables (i.e., a naming trial is correct or incorrect, and comes before or after therapy), the observations are randomly sampled from the population, the observations are independent, and the row and column totals are fixed (i.e., both the total number of naming trials before and after therapy and the total numbers of items named correctly or incorrectly across both testing sessions are known). While only the first assumption is strictly true for our intended purposes, the remainder is reasonably approximated. When these assumptions are violated, the Fisher exact test is known to be overly conservative, meaning it is less likely to identify a real change when one exists. Nevertheless, it continues to be one of the most widely used tests in the field for examining differences in proportions, and we believe it serves as a reasonable benchmark for inference quality based on the expected distribution under a null hypothesis.

Statistical Analysis 2: Treatment Response Magnitude

Group-level comparisons of average changes in summary statistics between treatment groups and the control group were made using unpaired, two-sample *t* tests. While we refer to the conventional thresholds for statistical significance (α = .05) and statistical trends (α = .1), we also provide unstandardized effect sizes and other descriptive statistics to broaden the perspective beyond the dichotomous viewpoint of the significance test toward a more gradational evaluation of the evidence. Additionally, BFs (BF_{10}) are presented alongside *p* values, assuming a unit information prior with the default scale *r* on effect size (Rouder et al., 2009). The unit information prior was chosen because relatively small effect sizes are expected for group-level case–control comparisons in anomia therapy research. BF_{10} values greater than one favor the alternative hypothesis of a nonzero difference in means, whereas values less than one favor the null hypothesis, with 3.00 and 0.33 taken as thresholds for substantial support, respectively (Rouder et al., 2009).

Group-level overlap in change scores between therapy groups and the control group was assessed by examining the area under the receiver operating characteristic curve (AUC). The AUC represents the probability that a randomly selected individual from a treatment group will have greater improvement in a summary statistic than a randomly selected individual from the control group. The 95% confidence interval for the AUC of each summary statistic was estimated using 1,000 bootstrap samples of the data with replacement. Confidence intervals excluding .50 were taken as significant.

Following McHorney et al. (1997), we directly compared the change statistics with one another by examining their relative precision (RP) in distinguishing treatment groups from the control group. While the *p* value s and BF_{10} values compare the change observed in one statistic to a hypothetical null change, the RP values compare the change observed in one statistic to the change observed in another statistic. We defined RP as the ratio (fraction) of the *t* statistics from the comparisons of the treatment and control groups' average magnitudes of change (i.e., the ratio of standardized effect sizes). The %C effect size was assigned as the relative standard (i.e., the denominator of the RP ratio). The 95% confidence interval for the RP of each summary statistic was estimated using 1,000 bootstrap samples of the data with replacement. Confidence intervals excluding 1.00 were taken as significant.

Results

Reliability of PNT Scoring

The kappa values for interrater and intrarater reliability, along with their correlation with the rate of each response type, are presented in Supplemental Material S6. Kappa values for interrater agreement of polytomous scores (i.e., considering all response types) ranged from .52 (moderate agreement) to 1 (perfect agreement), with an average of .81 (near-perfect agreement). There was one participant (of nine) for whom kappa indicated problematic interrater agreement in the polytomous scores (kappa = .52), although this is still considered to be moderate agreement. There were very strong correlations between interrater kappa for polytomous scores and the rate of correct (r = .76, p = 1.70×10^{-2}), formal (r = −.95, $p < 1.00 \times 10^{-4}$), abstruse neologism (r = −.86, p = 2.70×10^{-3}), and unrelated response rates (r = −.84, p = 4.20×10^{-3}), and a moderate correlation with nonword response rates (r = −.66, p = 6.60×10^{-2}). Correlations between interrater kappa for polytomous scores and semantic response rates ($r = -2.60 \times 10^{-3}$, p = .99) or mixed response rates (r = .36, p = .34) were negligible to weak. These results indicate that lower accuracy due to more nonword errors and word errors without a semantic relation to the target (i.e., phonological-level errors, in psycholinguistic terms) can compromise agreement at the item level between different raters.

Kappa values for intrarater agreement of polytomous scoring of all response types ranged from .67 (substantial agreement) to 1.00 (perfect agreement), with an average of .86 (near-perfect agreement). There were no participants for whom kappa indicated problematic intrarater agreement in the polytomous scores (all kappa > .60). There was a very strong correlation between intrarater kappa for polytomous scores and the rate of correct responses (r = .74, p = 2.30×10^{-2}), and moderate correlations with nonnaming attempt (r = −.63, p = 6.90×10^{-2}) and formal response rates (r = −.58, p = .10). Correlation strengths between interrater kappa for polytomous scores and rates of other response types were negligible to weak (maximum absolute r = .29, minimum p = .45, for semantic errors). These results indicate that lower accuracy due to nonnaming attempts and formal errors can compromise agreement somewhat within the same rater who scores the same test items at different times, though the minimum intrarater agreement was still substantial.

The confusion matrices and Jaccard indices for interrater and intrarater agreement are also included in Supplemental Material S6. For interrater comparisons, correct responses had the highest agreement (Jaccard index = 90%), followed by nonnaming attempt (70%), semantic (57%), mixed (54%), neologism (49%), abstruse neologism (43%), formal (32%), and unrelated responses (26%). A plurality of trials scored with a given response type by any rater agreed with the alternate rater's scored response type, for all response types except unrelated errors. Trials that were scored as unrelated errors by one rater were most often scored as nonnaming attempts by the alternate rater (33%), possibly due to misidentification of the first complete response. Unrelated errors were confused with formal errors (13%), and formal errors were confused with unrelated errors (6%), reflecting a difference in judgment of the phonological relationship between the target and the response. Abstruse neologism errors were most often confused with neologism errors (30%), and neologism errors were confused with abstruse neologism errors (14%), again reflecting a difference in judgment of the phonological relationship between the target and the response. Mixed errors were most often confused with semantic errors (19%), and semantic errors were also confused with mixed errors (9%), again reflecting a difference in judgment of the phonological relationship between the target and the response. Unrelated errors were confused with abstruse neologism errors (11%), and abstruse neologism errors were confused with unrelated errors (6%), reflecting a difference in judgment of the lexical status of the response. Formal errors were most often confused with neologism errors (25%), and neologism errors were also confused with formal errors (12%), again reflecting a difference in judgment of the lexical status of the response. Formal errors were confused with correct responses (22%), as were neologism errors (18%), likely reflecting differences in interpreting dialectal or articulatory-phonological influences on utterances. Semantic errors (15%) and mixed errors (12%) were confused with nonnaming attempts, likely reflecting differences in the categorization of verbs and adjectives as single-word descriptions versus naming attempts. Differences in judgment of the semantic relationship between the target and the response were rare. Notably, most of the disagreements at the item level occurred between response types that were relatively close in their ordering of processing completion assigned by the MPT model (i.e., unrelated and no response errors are similar indicators of less successful cognitive processing, whereas formal and neologism errors are similar indicators of more successful cognitive processing).

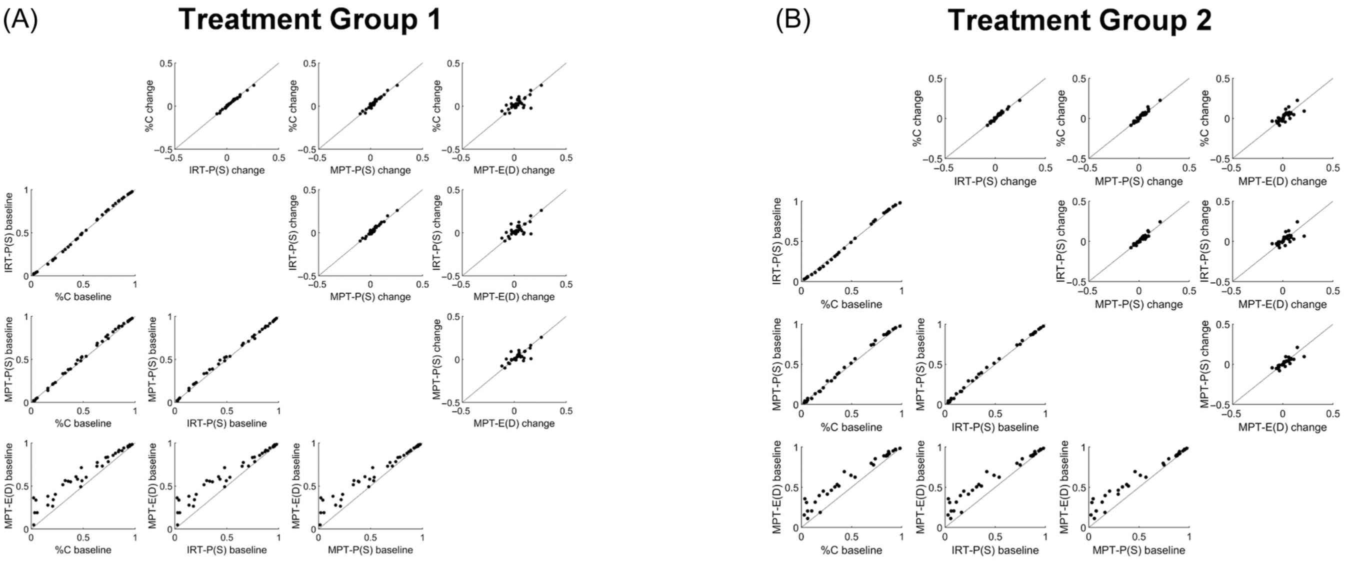
Despite the presence of item-level discrepancies in scoring, participant-level frequency counts were highly consistent both within and between raters for all response types. Among different raters, the average difference score was less than a single item for all response types except correct responses, which had an average difference of two items. The minimum correlation between frequency counts from different raters was r = .83 for neologism errors; all other r > .96. Within the same raters, the minimum correlation between frequency counts on different scoring occasions was r = .95, for mixed errors.

Model Fitting

Figure 4 shows the correlations between the point estimates for the four summary statistics at baseline and the point estimates of changes in these summary statistics after therapy. Of note, in both treatment groups, %C, IRT-P(S), and MPT-P(S) all have extremely high correlations with each other for both baseline and change point estimates, essentially identical, whereas MPT-E(D) is strongly related to the other summary statistics, but not identical. At baseline, MPT-E(D) is always greater than or equal to the other accuracy statistics, meaning that it technically has less room to improve over time; however, this discrepancy only becomes relevant for comparison with the other statistics if participants are reaching the improvement ceiling (no participants achieved perfect accuracy).

All model checks were satisfactory. We demonstrated convergence of the estimates for the variables of interest, meaning that the obtained estimates are reproducible despite the element of chance in the estimation procedure. The posterior point and interval estimates of abilities were highly consistent across different justified prior distribution specifications, suggesting that our posterior estimates appropriately depend on the data and are not unduly influenced by arbitrary prior beliefs. Finally, the vast majority of the observed individual response type frequencies as well as the overall distributions of response type

Figure 4. Scatter plots showing the relationships between point estimates of the four summary statistics at baseline (bottom left) and the relationships between point estimates of the changes in the summary statistics after treatment (top right). The dotted, diagonal line represents the identity line. %C = percent correct; MPT = multinomial processing tree; IRT = item response theory.



frequencies generated by each participant could be adequately recovered from the estimated ability parameters. Given the established reliability of the ability estimates, we can examine whether they are useful for making inferences.

Statistical Analysis 1: Treatment Response Frequency

Figure 5 illustrates the relationships between baseline measures of each summary statistic and changes in each summary statistic for the control group and treatment groups. Participants exhibiting a significant change (positive or negative) are shaded in black and circled, whereas participants exhibiting a nonsignificant change are shaded in gray. Although nontherapy-related changes (improvement or decline) may be present in each of our samples, under the assumption that treatment was at least partially effective, we expected a significantly higher frequency of performance improvements in treatment groups relative to the control group, and we expected nonsignificant differences in frequencies of performance declines. These expectations were confirmed. Table 2 shows the frequencies of

Figure 5. Scatter plots showing the relationships between baseline measurements and estimated changes for each of the summary statistics after treatment. Points representing participants with significant change (two-tail p or posterior $p < .05$) in either direction, improvement or deterioration, are shaded black and circled; points representing participants with nonsignificant change are shaded gray. %C = percent correct; MPT = multinomial processing tree; IRT = item response theory.

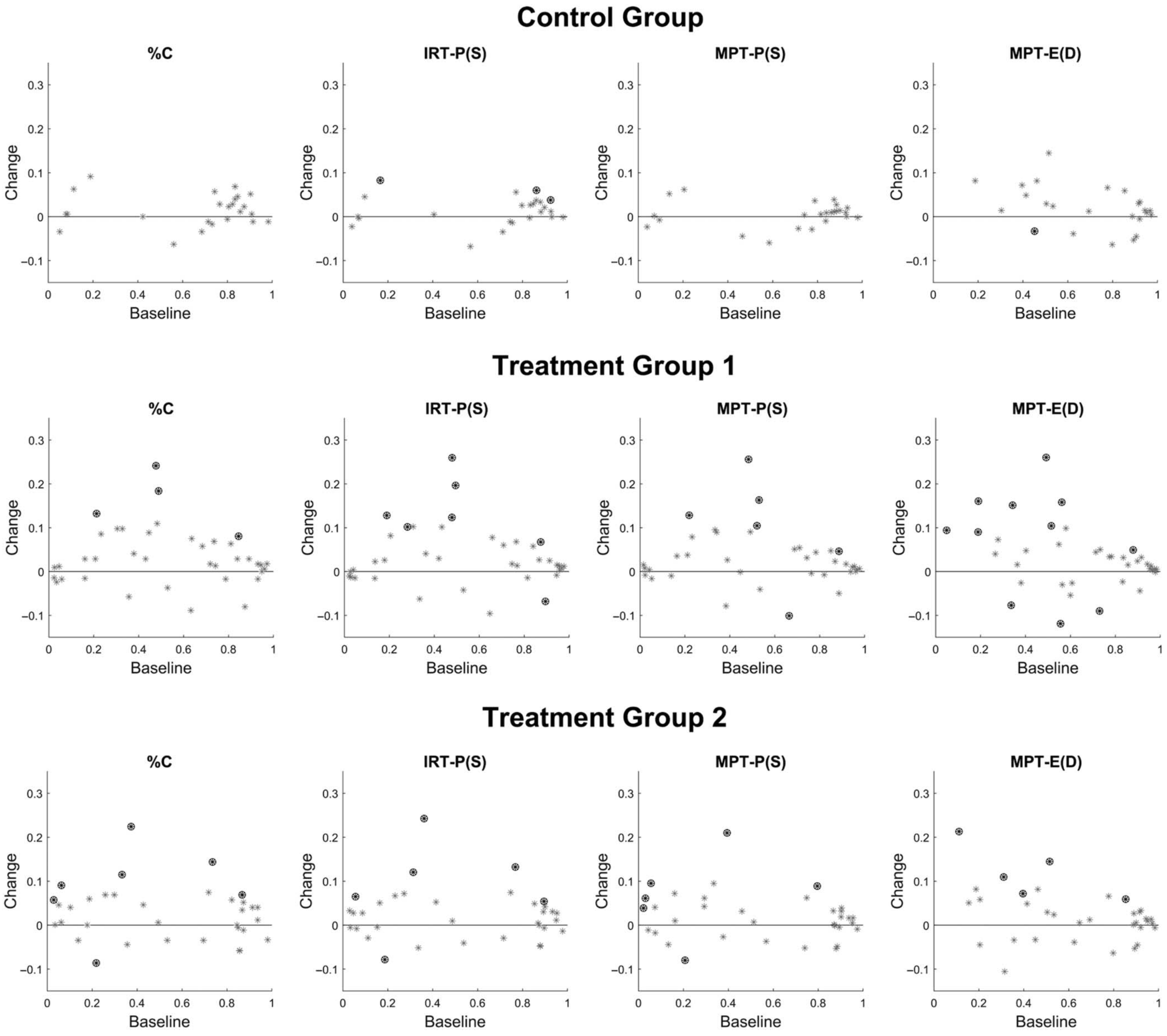


Table 2. Frequency of significant improvement.

Group	%C	IRT-P(S)	MPT-P(S)	MPT-E(D)
Control group	0	3	0	0
	0%	13%	0%	0%
Treatment Group 1	4	6	5	8**
	11%	16%	13%	21%
Treatment Group 2	6**	5	5*	5*
	18%	15%	15%	15%
Treatment Groups	10*	11	10*	13**
1 + 2	14%	15%	14%	18%

Note. Treatment versus control frequency. %C = percent correct; IRT = item response theory; MPT = multinomial processing tree.
*$p < .1$. **$p < .05$.

significant improvements, and Table 3 shows the frequencies of significant declines. Comparing the summary statistics to one another, the MPT-E(D) statistic identified the maximum number of significant improvements across treatment groups and the minimum number of significant improvements in the control group. It also identified the maximum number of significant declines in the treatment groups and control group, suggesting that this statistic is sensitive to changes in both directions; however, the frequency of significant declines in the treatment groups was not significantly greater than in the control group, in accordance with expectations about treatment efficacy. That is, while we did find evidence that treatment leads to significantly more improvements of MPT-E(D), we did not find evidence that treatment leads to significantly more declines.

Inferences about significant improvement for individual participants based on different measures sometimes-resulted in different conclusions. In general, the MPT model-based distance measure, MPT-E(D), tended to identify more significant responders than accuracy measures, MPT-P(S), IRT-P(S), or %C. Three of the four statistics, %C, MPT-P(S), and MPT-E(D), reflected significant effects of treatment on the frequency of individual improvements relative to the control group, whereas the IRT-P(S) statistic identified too many improvements in the control group to differentiate them from the treatment groups.

Table 3. Frequency of significant decline.

Group	%C	IRT-P(S)	MPT-P(S)	MPT-E(D)
Control group	0	0	0	1
	0%	0%	0%	4%
Treatment Group 1	0	1	1	3
	0%	3%	3%	8%
Treatment Group 2	1	1	1	0
	0%	3%	3%	0%
Treatment Groups	1	2	2	3
1 + 2	1%	3%	3%	4%

Note. Minimum p = .27, treatment versus control frequency. %C = percent correct; IRT = item response theory MPT = multinomial processing tree.

Statistical Analysis 2: Treatment Response Magnitude

A summary of the results is presented in Table 4. The only summary statistic for which the alternative hypothesis (i.e., greater average improvement for treatment versus control) was supported was MPT-E(D) (Treatment Group 1: $p = .01$, $BF_{10} = 5.85$; Treatment Group 2: $p = .03$, $BF_{10} = 3.25$). Recall that the BF_{10} value indicates how many times more likely the alternative hypothesis is than the null hypothesis. It was 5.9 times more likely that there was greater improvement for Treatment Group 1 versus control than that these groups had equal improvement. It was 3.3 times more likely that there was greater improvement for Treatment Group 2 versus control than that these groups had equal improvement. With regard to RP, the IRT-P(S) and MPT-P(S) statistics were not significantly different from %C at detecting the group-level effects of therapy in either treatment group (max. RP = 1.44); however, the MPT-E(D) statistic trended differently than %C in Treatment Group 1 (RP = 2.00; 97% of bootstrap samples > 1.00) and was significantly different in Treatment Group 2 (RP = 2.51; 98% of bootstrap samples > 1.00).

Figure 6 presents the AUC analysis of change magnitude as bar graphs. In both treatment groups, MPT-E(D) was the only statistic that was able to discriminate treatment participants from control participants at rates significantly better than chance (Treatment Group 1: AUC = .71, CI [.57, .83]; Treatment Group 2: AUC = .67, CI [.52, .79]).

Differences in the magnitude of improvement between the treatment groups and the control group were significantly larger for the MPT-E(D) statistic than the other measures. These results support the assumptions of the MPT-Naming model (a) that multiple latent components are contributing to picture naming responses and (b) that targeted therapies may have systematic effects that are measurable despite the persistence of impairments in other cognitive components.

Summary of Results

Taken together, these analyses provide strong evidence that the MPT-E(D) statistic is more sensitive to treatment effects than statistics based on accuracy alone. Inferences based on the MPT-E(D) statistic identified more significant responders in the treatment groups overall and significantly more responders compared with the control group. In contrast to the other accuracy statistics, participants in the treatment groups could be distinguished from participants in the control group based on the magnitude of changes in the MPT-E(D) statistic at rates greater than chance. The group-level effects of treatment versus control were 2.0–2.5 times stronger for the MPT-E(D) statistic than the %C statistic, and this effect held across both bootstrap resampling of the data and independent replication of the

Table 4. Summary of results comparing change magnitude in each statistic between the treatment groups and the control group.

Treatment group	Variable	%C	IRT-P(S)	MPT-P(S)	MPT-E(D)
Treatment Group 1	Mean change difference	.02	.02	.03	.04
	Pooled standard deviation	.06	.06	.05	.06
	t	1.30	1.43	1.86	2.59
	Unpaired, two-tail *p*	.20	.16	.07	.01
	BF_{10}	0.73	0.85	1.55	5.85
	RP [95% CI]	1.00 [1.00, 1.00]	1.10 [0.06, 2.13]	1.44 [0.51, 2.52]	2.00 [0.99, 3.14]
	AUC [95% CI]	.59 [.43, .73]	.58 [.44, .73]	.63 [.48, .76]	.71 [.57, .83]
Treatment Group 2	Mean change difference	.01	.01	.02	.03
	Pooled standard deviation	.05	.05	.05	.05
	t	0.91	0.86	1.26	2.29
	Unpaired, two-tail *p*	.37	.39	.21	.03
	BF_{10}	0.51	0.49	0.71	3.25
	RP [95% CI]	1.00 [1.00, 1.00]	0.94 [0.66, 2.30]	1.39 [0.29, 2.90]	2.51 [1.11, 4.19]
	AUC [95% CI]	.57 [.41, .70]	.54 [.38, .69]	.59 [.43, .74]	.67 [.52, .79]

Note. %C = percent correct; IRT = item response theory; MPT = multinomial processing tree; BF = Bayes factor; RP = relative precision; CI = confidence interval; AUC = area under the receiver operating characteristic curve.

experiment. The response type data, assessed through the lens of a latent cognitive model, provided an expanded view of systematic and measurable treatment effects.

General Discussion

Systematic Effects of Therapy on Error Types

In this study, we examined changes in four different summary statistics derived from picture naming response type data. We found that accounting for error types can increase confidence in observed changes in accuracy. We also found that a scale for completeness of cognitive processing can be operationalized with respect to error type data, which can capture treatment effects that improve cognitive processing without necessarily improving accuracy. There were seven participants (out of 72) who exhibited a significant change in MPT-E(D) without a significant change in %C, four participants showing the reverse pattern, and six participants with significant changes in both statistics. The changes in response type rates for each of these participants are shown in Figure 7. While changes in response type rates are a main component of the inference outcome, it is important to remember that the number of observed trials and the items on which errors occurred also affect the inference, and these aspects are not pictured. When MPT-E(D) changed without significant changes in %C, generally the most

Figure 6. Bar graphs showing the pairwise classification accuracy (i.e., the area under the receiver operating characteristic curve [AUC]) for distinguishing treatment participants from control participants based on the magnitude of change in each summary statistic. Error bars representing 95% confidence intervals for AUC were estimated using 1,000 bootstrap samples with replacement. MPT = multinomial processing tree; IRT = item response theory; %C = percent correct.

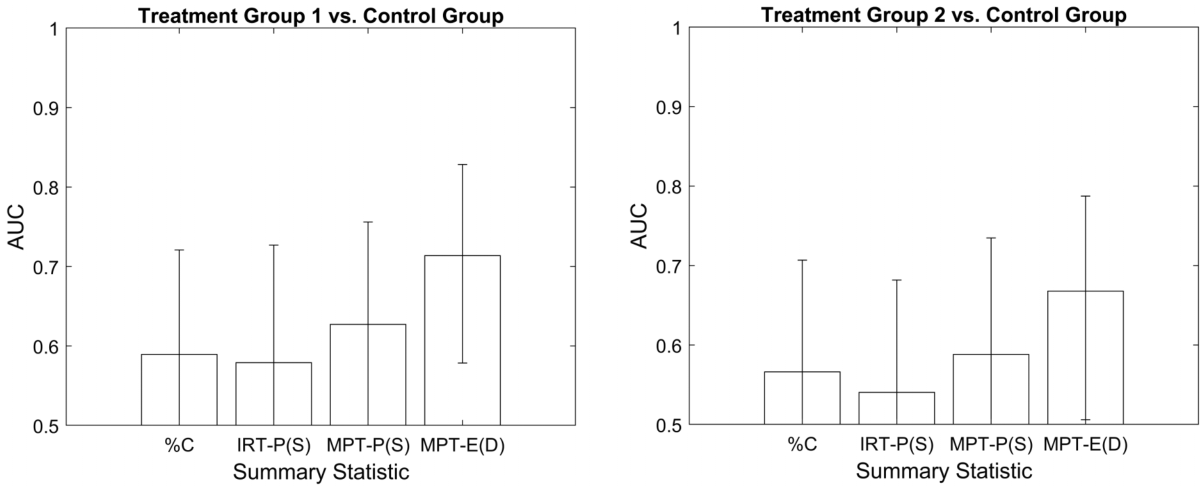
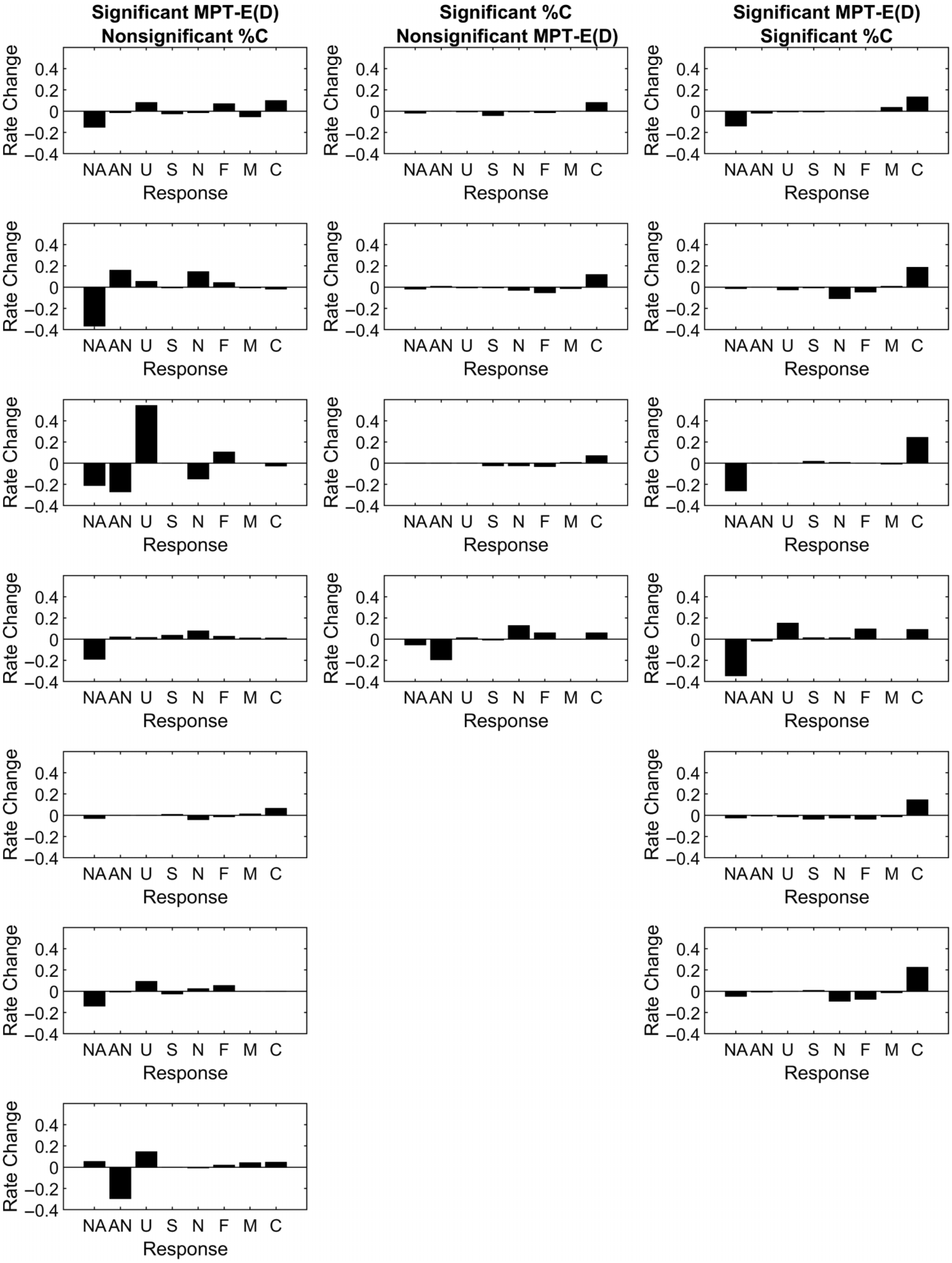


Figure 7. Bar graphs showing the changes in response type rates for seven participants with significant MPT-E(D) change and nonsignificant %C change (left), four participants with nonsignificant MPT-E(D) change and significant %C change (middle), and six participants with significant MPT-E(D) change and significant %C change (right). %C = percent correct; MPT = multinomial processing tree; NA = no attempt; AN = abstruse neologism; U = unrelated; S = semantic; N = neologism; F = formal; M = mixed; C = correct.



severe error types (i.e., nonnaming attempts and abstruse neologisms) were converted into less severe error types. When %C changed without significant changes in MPT-E(D), the changes in %C were relatively small and, for MPT-E(D), these changes were moderated by the knowledge that certain, infrequent error types are likely to occur on certain items due to chance rather than a true latent change. That is, for someone without any changes in ability, it is still reasonable to expect that a phonological error, for example, might occur on a phonologically difficult item before treatment but not after treatment simply due to chance, although this should happen infrequently. Without considering the specific difficulty of the items and the error types that are committed, one might consider a small improvement in accuracy to be more significant than is warranted. Agreement between the statistics typically involved relatively large increases in correct response rates along with decreases in nonnaming attempts.

These incremental improvements that are reflected in the MPT-E(D) statistic are measurable and distinct from changes that occur during spontaneous recovery in the chronic stage, as revealed by our case–control comparison of change magnitude. They are also consistent with theoretical proposals regarding the nature of speech and language deficits in aphasia. The continuity thesis (Freud, 1953) asserts that the causes of speech errors observed in aphasia lie on a continuum ranging from the healthy (but still occasionally error-prone) symbolic production system to a completely unregulated, random symbolic production system, which implies that the observed error types offer clues regarding the extent to which the system is damaged. This has been claimed to be the most important feature for any model seeking to explain rates of different speech errors in aphasia (Dell et al., 1997; Schwartz et al., 2006). Neurocomputational models of aphasia that adopt the parallel distributed processing approach (Ueno et al., 2011) also instantiate gradients of performance that can be localized to specific layers and connections mediating a series of representational transformations from visual input to spoken output. The MPT model's assumptions are therefore better aligned with current theories of word retrieval than the all-or-nothing quality represented by accuracy statistics. Calibrating our measurement scales to be more sensitive to these incremental gains in cognitive processing is an important advancement for the POLAR study, because predicting outcomes presumes that we can accurately measure outcomes (and not just the proverbial tip of the iceberg).

Systematic Effects of Test Items on Error Types

Different estimates of item difficulty will affect how correction of a certain error on a certain item is interpreted. To further illustrate the differences between item difficulty estimates based on correct or incorrect responses and item difficulty estimates based on multiple error types, we examined the correlations between these types of estimates. First, we compared our new estimates of the difficulty of producing an accurate response with an extant analysis in the literature. Fergadiotis et al. (2015) described an IRT analysis for PNT accuracy based on a subset of 251 of the 365 participants included in the calibration cohort for the current study. We examined the correlation between our Bayesian point estimates of item difficulty and the estimates reported by Fergadiotis et al. (2015), finding a very strong correlation (r = .99). The relationship for higher difficulty items veers off the identity line somewhat (see Supplemental Material S7), due to the inclusion of participants with more severe impairment, on average, in the current calibration cohort; more importantly, however, the relative ordering of items is essentially preserved. Next, we compared our estimates of the difficulty of producing an accurate response with our estimates of the difficulty of avoiding a particular error type. The correlations of accuracy difficulty with the MPT difficulties for all 175 items were as follows: *Attempt* (r = .84), *LexSem* (r = .56), *LexPhon* (r = .61), *LexSel* (r = .76), and *Phon* (r = .84). Given relatively less correspondence between the *LexSem* difficulty and the accuracy difficulty, for example, remediation of semantic errors on semantically challenging targets would not receive due credit using the accuracy difficulty scale. Thus, the difficulty scales used by the MPT model provide insight into a wider variety of impairment changes. Inferences based on the incorrect assumption of unidimensional item difficulty may be more misleading than inferences based on classical test theory that ignore item heterogeneity.

Just as accounting for item difficulty without accounting for response types was insufficient for detecting the case–control differences, accounting for response types without accounting for item difficulty or the sequential nature of cognitive processing also failed to discriminate between treatment and control groups. In a post hoc analysis of the standard partial credit score (i.e., assigning points to each item based on response types and summing over items for a total score), we found that the standard partial credit scores behaved like the standard accuracy scores when making group-level inferences based on change magnitude (see Supplemental Material S8). These results are reassuring that we are not simply "reinventing the wheel" with the MPT-E(D) statistic. The cognitive model is necessary for revealing additional, systematic effects of treatment on response types.

Limitations and Future Directions

Sampling Limitations

As with most aphasia research, our data are not collected from a random sample of persons with aphasia, nor are they collected from a random sample of picturable nouns in English. The characteristics of the participants and items, therefore, should be considered carefully before extending our estimates and conclusions beyond the current study cohorts.

Cohort differences. Because we were limited to analyzing samples of convenience, there were several important differences between the participant cohorts that could have affected the current study. The most notable concern is the difference between treatment and control groups in terms of their baseline accuracy distributions. Generally, the control participants had higher baseline accuracy than treatment participants. If we had included more severe control participants, we might have observed more spontaneous improvement of MPT-E(D) and less contrast between the groups. Even if this were the case, it would imply that severe and mild aphasias at baseline are distinguishable by changes in response types over time, but not by changes in accuracy over time. So, the MPT-E(D) statistic would still be informative.

Another potential concern is the difference between treatment groups and the control and calibration groups in terms of motor speech impairment. Perhaps the observed group differences relate to estimates of item difficulties being more valid for the control group than for the treatment groups, because their specific impairments are more similar. The issue of item calibration based on a "representative" sample can be quite challenging in the context of a heterogeneous condition like aphasia. In the current work, we sought to be as inclusive as possible, relying on estimates that are supported by external validation experiments (Walker et al., 2018, 2020), but future work should investigate the impact of item calibration procedures.

The control and treatment groups also differed in terms of the chronicity of their conditions. Generally, the control participants were earlier in their course of recovery than treatment participants. This difference leads to a conservative stance on the hypothesis that treatment improves performance, because the control group may still be exhibiting some spontaneous recovery. The fact that changes in behavior are more noticeable in a group that has had the condition longer supports the claim that the treatment caused the changes rather than the mere passage of time.

The control and treatment groups also differed in terms of the number of times they were assessed during the study. Perhaps the repeated administration of the naming assessment, without the treatment, might have yielded the same results. While this is possible, the expectation of group differences remains, and only the MPT-E(D) statistic was sensitive to these differences.

Refractory effects. The PNT items were presented to each participant in the same, fixed, random order. This raises the concern that items presented previously may have affected the probabilities of response types on items presented later via refractory effects (i.e., cumulative semantic interference). This is a minor concern for the current study because, given the same, fixed order of item presentation at each assessment, refractory effects are expected to be the same across assessments and should therefore cancel out when differences in performance are examined for significant change in ability between tests. Furthermore, the MPT-Naming model's item difficulty estimates should account for order effects on the probability of a given error type, because the same presentation order was used for estimating item difficulties. This issue does raise concerns for the future development of short forms or adaptive tests that are presented in a different order, and thus may engender different item difficulties associated with order effects. These deviations in item difficulty due to order effects are expected to be small in the general population, with minimal impact on ability estimates when aggregated over items from multiple categories. This is because refractory effects are observed in a specific subset of people with aphasia, those with semantic control impairments (Jefferies et al., 2007). While these effects are robustly observed in latency data, in naming accuracy data, they are either nonsignificant (Belke, 2013; Howard et al., 2006; Riès et al., 2015; Runnqvist et al., 2012; Schnur, 2014) or very small, with each previously named item from a given category increasing the probability of a semantic error by less than 1% (Harvey et al., 2019). When these effects are aggregated over participants and items from multiple categories, they are negligible (i.e., comparable to error from random sampling).

Transcription and Scoring Limitations

Data collection and scoring is time-consuming and difficult, making the current methods unfeasible for many clinical settings. Our reliability analysis revealed that, although scoring agreement is quite high for a random pair of scorers scoring a random test from the full range of ability, there is clearly room for improvement in the context of more severe phonological-level impairment. Transcription and identification of phonological relations to the target appear to be particularly disagreement-prone steps of the data collection procedure. Automated scoring protocols are in development that can aid in the collection and standardization of response type data (Fergadiotis et al., 2016). Furthermore, knowledge about variance among scorers can be incorporated into future statistical models of the data-generating processes.

Short Forms and Adaptive Testing

The IRT formalisms encapsulated within the MPT-Naming model offer the potential for development of short forms and computer-adaptive tests for more efficient measurement of summary statistics; however, in the current study, we estimated abilities and their longitudinal changes using as much of the available data as possible (while maintaining comparability between groups). While the MPT model can naturally accommodate data from abbreviated tests without compromising its logic of inference, it remains to be seen how many and which items are necessary to infer whether a change has occurred with a desired level of confidence. Furthermore, testing specific psychometric properties of latent components in a multicomponent model—such as dimensionality, discriminability, item invariance, or local independence—can be quite challenging, as many of the standard assessments assume a more direct relationship exists between model parameters and data. To the extent that model assumptions are not met, inferences are expected to be suboptimal. Thus, in our investigations to date, rather than attempting to determine the truth of these assumptions individually, we have attempted to determine whether the assumptions, taken together, are useful compared with standard approaches (Box, 1979). We hope that the statistical model might ultimately inform the development of computerized adaptive naming tests, such as the one that already exists for naming accuracy on the PNT (Hula et al., 2015, 2019).

Modeling Correlations Between Abilities at the Population Level

We know that naming abilities, particularly lexical abilities (i.e., *LexSem, LexPhon,* and *LexSel*), are generally correlated with each other in a cross-sectional sample of participants with aphasia, and longitudinal changes in abilities might also be expected to exhibit significant correlations. Currently, our prior assumptions stipulate that these are all independent from one another, so incorporating our knowledge of these correlations into the model could be a logical next step. The latent-trait approach proposed by Klauer (2010) that handles estimation of correlated parameters in MPT models was adapted to also account for item heterogeneity by Matzke et al. (2015). Reformulating the MPT-Naming model in this framework is a natural progression that could constrain parameter estimates in informative ways. However, there are potential pitfalls in terms of appropriately defining the variety of correlated abilities and changes in abilities that one might find in different variants and intensities of aphasia. For example, within the calibration cohort, we find significantly different correlations between point estimates of the *LexPhon* and *Phon* parameters for subsets of participants with or without apraxia of speech, a motor speech planning disorder (Z test; $r1$ [with apraxia] = .52, $r2$ [without apraxia] = .74, $n1$ = 90, $n2$ = 275; two-tailed p = .002). That is, the severity of sublexical (*Phon*) deficits in participants without apraxia are strongly related to the severity of lexical–phonological (*LexPhon*) deficits, but this is less true of participants with apraxia, where the severity of the sublexical deficit tends to better explain the relative frequencies of word and nonword errors on its own. Further research will need to identify appropriate calibration cohorts to inform prior expectations about correlated parameters.

Modeling Individual Responses, Processes, and Changes

The cognitive psychometric model of picture naming abilities and changes over time can be developed further. We used the scoring protocol for naming data outlined by Roach et al. (1996), but there are other ways of categorizing or quantifying characteristics of speech errors. For example, the current scoring protocol does not consider differences in articulatory phonetic disturbances versus phonological substitutions. Alternative scoring protocols may reveal different aspects or substages of cognitive processing. Alternative tree structures may also lead to different estimates of abilities. Finally, the temporal structure of the model can be updated. For example, changes over time could be modeled as continuous curves, given continuous observations over time (Evans et al., 2021). The temporal structure of each trial could also be modeled to accommodate inferences about reaction times (Evans et al., 2020). The current results can provide a new benchmark for a model's capability to measure the specific effects of treatment.

Clinical Significance

Finally, while we have found clear evidence for statistically significant effects of therapy, we still do not know how these effects translate into clinical significance. That is, we do not know if an improvement of these summary statistics relates to an improvement in quality of life or success in communication. This is true of all the measures studied here, including %C. Nevertheless, benchmarks for clinical significance can and should be related to concepts of statistical significance, appropriately framed with respect to the ultimate quantities of interest (i.e., those related to client satisfaction). The current study provides new, potentially relevant metrics that warrant further investigation in this regard.

Data Availability

All naming data, model code, analysis code, and posterior samples generated from this study are provided in Supplemental Material S9. A smaller directory without posterior samples is also provided in Supplemental Material S10.

Author Contributions

Grant M. Walker: Conceptualization (Lead), Data curation (Equal), Formal analysis (Lead), Methodology (Equal), Software (Lead), Validation (Lead), Visualization (Lead), Writing – original draft (Lead), Writing – review & editing (Lead). **Alexandra Basilakos:** Data curation (Equal), Investigation (Lead), Project administration (Lead), Writing – review & editing (Supporting). **Julius Fridriksson:** Funding acquisition (Lead), Methodology (Equal), Project administration (Lead), Resources (Lead), Supervision (Lead), Writing – review & editing (Supporting). **Gregory Hickok:** Conceptualization (Supporting), Funding acquisition (Supporting), Writing – review & editing (Supporting).

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders Grant P50 DC014664, awarded to J. Fridriksson. Leigh Ann Spell supervised graduate student raters and collected reliability data.

References

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

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10*(4), 331–344. https://doi.org/10.1037/1040-3590.10.4.331

Batchelder, W. H., Chosak-Reiter, J., Shankle, W. R., & Dick, M. B. (1997). A multinomial modeling analysis of memory deficits in Alzheimer's disease and vascular dementia. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 52*(5), P206–P215. https://doi.org/10.1093/geronb/52b.5.p206

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*(1), 57–86. https://doi.org/10.3758/BF03210812

Batchelder, W. H., & Riefer, D. M. (2007). *Using multinomial processing tree models to measure cognitive deficits in clinical populations*. https://doi.org/10.1037/11556-001

Belke, E. (2013). Long-lasting inhibitory semantic context effects on object naming are necessarily conceptually mediated: Implications for models of lexical-semantic encoding. *Journal of Memory and Language, 69*(3), 228–256. https://doi.org/10.1016/j.jml.2013.05.008

Boedigheimer, M. (2021). *fexact* (varargin). MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/22550-fexact-varargin

Boo, M., & Rose, M. L. (2011). The efficacy of repetition, semantic, and gesture treatments for verb retrieval and use in Broca's aphasia. *Aphasiology, 25*(2), 154–175. https://doi.org/10.1080/02687031003743789

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Boyle, M. (2004). Semantic feature analysis treatment for anomia in two fluent aphasia syndromes. *American Journal of Speech-Language Pathology, 13*(3), 236–249. https://doi.org/10.1044/1058-0360(2004/025)

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. https://doi.org/10.1044/1058-0360.0404.94

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. https://doi.org/10.1080/026870300401513

Cohen, J. (1960). Kappa: Coefficient of concordance. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we. *Psychological Bulletin, 74*(1), 68–80. https://doi.org/10.1037/h0029382

Cuetos, F., Aguado, G., & Caramazza, A. (2000). Dissociation of semantic and phonological errors in naming. *Brain and Language, 75*(3), 451–460. https://doi.org/10.1006/brln.2000.2383

Davis, G. A. (2005). PACE revisited. *Aphasiology, 19*(1), 21–38. https://doi.org/10.1080/02687030444000598

Davis, G. A. (2007). *Aphasiology: Disorders and clinical practice*. Pearson/Allyn & Bacon.

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology, 21*(2–4), 125–145. https://doi.org/10.1080/02643290342000320

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

Edmonds, L. A., & Babb, M. (2011). Effect of verb network strengthening treatment in moderate-to-severe aphasia. *American Journal of Speech-Language Pathology, 20*(2), 131–145. https://doi.org/10.1044/1058-0360(2011/10-0036)

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

Embretson (Whitely), S. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175–186. https://doi.org/10.1007/BF02294171

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). Springer. https://doi.org/10.1007/978-1-4757-2691-6_18

Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement, 7*(3), 335–350.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*(1), 14–36. https://doi.org/10.1007/s11336-012-9296-y

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift Für Psychologie / Journal of Psychology, 217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

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

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

Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology, 25*(4S), S776–S787. https://doi.org/10.1044/2016_AJSLP-15-0147

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

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87. https://doi.org/10.2307/2340521

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

Freud, S. (1953). *On aphasia: A critical study*. International University Press.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*(1), 27–52. https://doi.org/10.2307/412187

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

Harvey, D. Y., Traut, H. J., & Middleton, E. L. (2019). Semantic interference in speech error production in a randomised continuous naming task: Evidence from aphasia. *Language, Cognition and Neuroscience, 34*(1), 69–86. https://doi.org/10.1080/23273798.2018.1501500

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition, 100*(3), 464–482. https://doi.org/10.1016/j.cognition.2005.02.006

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

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

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

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Jefferies, E., Baker, S. S., Doran, M., & Lambon Ralph, M. A. (2007). Refractory effects in stroke aphasia: A consequence of poor semantic control. *Neuropsychologia, 45*(5), 1065–1079. https://doi.org/10.1016/j.neuropsychologia.2006.09.009

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

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

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146

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

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

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

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

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers, 36*(3), 421–431. https://doi.org/10.3758/BF03195590

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523–547. https://doi.org/10.1007/BF02294327

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

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

McHorney, C. A., Haley, S. M., & Ware, J. E., Jr. (1997). Evaluation of the MOS SF-36 physical functioning scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology, 50*(4), 451–461. https://doi.org/10.1016/S0895-4356(96)00424-6

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

Mitchum, C. C., Ritgert, B. A., Sandson, J., & Berndt, R. S. (1990). The use of response analysis in confrontation naming. *Aphasiology, 4*(3), 261–279. https://doi.org/10.1080/02687039008249079

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters. *Neuropsychologia, 33*(10), 1281–1303. https://doi.org/10.1016/0028-3932(95)00102-9

Nozari, N., & Hepner, C. R. (2019). To select or to wait? The importance of criterion setting in debates of competitive lexical selection. *Cognitive Neuropsychology, 36*(5–6), 193–207. https://doi.org/10.1080/02643294.2018.1476335

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Pulvermüller, F., Neininger, B., Elbert, T., Mohr, B., Rockstroh, B., Koebbel, P., & Taub, E. (2001). Constraint-induced therapy of chronic aphasia after stroke. *Stroke, 32*(7), 1621–1626. https://doi.org/10.1161/01.STR.32.7.1621

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Measures of agreement. *Perspectives in Clinical Research, 8*(4), 187–191. https://doi.org/10.4103/picr.PICR_123_17

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

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

Riès, S. K., Karzmark, C. R., Navarrete, E., Knight, R. T., & Dronkers, N. F. (2015). Specifying the role of the left prefrontal cortex in word selection. *Brain and Language, 149*, 135–147. https://doi.org/10.1016/j.bandl.2015.07.007

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. http://eprints-prod-05.library.pitt.edu/215/1/24-09.pdf

Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. Wheeldon (Ed.), *Aspects of language production* (pp. 71–114). Psychology Press.

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

Runnqvist, E., Strijkers, K., Alario, F. X., & Costa, A. (2012). Cumulative semantic interference is blind to language: Implications for models of bilingual speech production. *Journal of Memory and Language, 66*(4), 850–869. https://doi.org/10.1016/j.jml.2012.02.007

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures*. *Journal of Quantitative Linguistics, 16*(1), 96–114. https://doi.org/10.1080/09296170802514138

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

Schwartz, M. F., & Brecher, A. R. (2000). A model-driven analysis of severity, response characteristics, and partial recovery in aphasics' picture naming. *Brain and Language, 73*(1), 62–91. https://doi.org/10.1006/brln.2000.2310

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming⋆. *Journal of Memory and Language, 54*(2), 228–264. https://doi.org/10.1016/J.JML.2005.10.001

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

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron, 72*(2), 385–396. https://doi.org/10.1016/j.neuron.2011.09.013

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition, 85*(3), B61–B69. https://doi.org/10.1016/S0010-0277(02)00107-5

Walker, G. M., Fridriksson, J., & Hickok, G. (2020). Connections and selections: Comparing multivariate predictions and parameter associations from latent variable models of picture naming. *Cognitive Neuropsychology, 38*(1), 50–71. https://doi.org/10.1080/02643294.2020.1837092

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

Wambaugh, J. L., & Ferguson, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research and Development, 44*(3), 381–394. https://doi.org/10.1682/JRRD.2006.05.0038

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications (pp. 119–145)*. Cambridge University Press.