Research Article

Enhancing the Efficiency of Confrontation Naming Assessment for Aphasia Using Computer Adaptive Testing

Gerasimos Fergadiotis,^a William D. Hula,^{b,c} Alexander M. Swiderski,^{c,d}
Chia-Ming Lei,^{c,d} and Stacey Kellough^d

Purpose: In this study, we investigated the agreement between the 175-item Philadelphia Naming Test (PNT; Roach, Schwartz, Martin, Grewal, & Brecher, 1996) and a 30-item computer adaptive PNT (PNT-CAT; Fergadiotis, Kellough, & Hula, 2015; Hula, Kellough, & Fergadiotis, 2015) created using item response theory (IRT) methods.

Method: The full PNT and the PNT-CAT were administered to 47 participants with aphasia in counterbalanced order. Latent trait-naming ability estimates for the 2 PNT versions were analyzed in a Bayesian framework, and the agreement between them was evaluated using correlation and measures of constant, variable, and total error. We also evaluated the extent to which individual pairwise differences were credibly greater than 0 and whether the IRT measurement model provided an adequate indication of the precision of individual score estimates.

Results: The agreement between the PNT and the PNT-CAT was strong, as indicated by high correlation ($r = .95$, 95% CI [.92, .97]), negligible bias, and low variable and total error. The number of statistically robust pairwise score differences did not credibly exceed the Type I error rate, and the precision of individual score estimates was reasonably well predicted by the IRT model.

Discussion: The strong agreement between the full PNT and the PNT-CAT suggests that the latter is a suitable measurement of anomia in group studies. The relatively robust estimates of score precision also suggest that the PNT-CAT can be useful for the clinical assessment of anomia in individual cases. Finally, the IRT methods used to construct the PNT-CAT provide a framework for additional development to further reduce measurement error.

Supplemental Material: <https://doi.org/10.23641/asha.8202176>

Naming impairment, or anomia, is a cardinal feature of aphasia (Goodglass & Wingfield, 1997; Kohn & Goodglass, 1985; Nickels, 2002), a disorder typically caused by stroke that affects approximately one in 250 people in the United States. The ability to access and retrieve words depends on cognitive processes that are critical for successful language production and has been found to be a good indicator of overall

aphasia severity (Schuell, Jenkins, & Jimenez-Pabon, 1964; Walker & Schwartz, 2012) and is predictive of the ability to convey information during discourse production (Fergadiotis & Wright, 2015). Furthermore, improvement in naming has been linked to improvement in overall communicative functioning (Carragher, Conroy, Sage, & Wilkinson, 2012; Herbert, Hickin, Howard, Osborne, & Best, 2008). Given the prevalence of anomia in the aphasic population and its tendency to persist even when other symptoms of aphasia remit (Goodglass & Wingfield, 1997), an accurate diagnosis of word retrieval deficits is essential.

Word production can be conceptualized as an emergent ability (McClelland, 2010) reflecting a constellation of cognitive processes that are activated for word access and phonological encoding. This skill is often modeled as a complex multistage process in which earlier stages are dominated by lexical-semantic processing and later stages by phonological processing (Caramazza, 1997; Foygel & Dell, 2000; Levelt, Roelofs, & Meyer, 1999). Furthermore, models typically involve a spreading activation mechanism

^aDepartment of Hearing & Speech Science, Portland State University, OR

^bGeriatric Research, Education and Clinical Center, VA Pittsburgh Healthcare System, PA

^cDepartment of Communication Science & Disorders, University of Pittsburgh, PA

^dVA Pittsburgh Healthcare System, PA

Correspondence to Gerasimos Fergadiotis: g3@pdx.edu

Editor-in-Chief: Sean Redmond

Editor: Charles Ellis

Received August 24, 2018

Revision received November 21, 2018

Accepted December 16, 2018

https://doi.org/10.1044/2018_JSLHR-L-18-0344

Disclosure: The authors have declared that no competing interests existed at the time of publication.

that acts across this sequence of stages (Foygel & Dell, 2000; Levelt et al., 1999). In the context of eliciting words using confrontation naming tests, recognition of a stimulus leads to the activation of its semantic features, which, in turn, activate a lexical representation that is specified for meaning and syntax. Then, the word's phonological form is retrieved, and its constituent phonemes are sequenced.

Confrontation picture-naming tests (e.g., Druks & Masterson, 2000; German, 1990; Kaplan, Goodglass, & Weintraub, 2001) are typically administered to assess anomia severity, build patients' profiles, and quantify change after treatment. Currently, confrontation picture-naming tests are utilized in standalone assessments, such as the Boston Naming Test (BNT; Kaplan et al., 2001) and the Philadelphia Naming Test (PNT; Roach et al., 1996), as well as aphasia assessments, such as the Western Aphasia Battery–Revised (Kertesz, 2007) and the Comprehensive Aphasia Test (Swinburn, Porter, & Howard, 2004). For example, the BNT, one of the most commonly used tests, includes 60 items, and on average, its administration requires approximately 35–45 min. Confrontation naming tests can be used to quantify a person's overall ability to access and retrieve words and, in many cases (e.g., BNT, PNT), offer the opportunity to categorize paraphasias into different types of errors based on which one can make inferences on the cognitive processes that subserve word finding. In this article, we are focused on developing more efficient ways of quantifying the overall ability to access and retrieve words. Extensions of the methods presented in this article to multidimensional models of multinomial error categories is a worthy topic of investigation (e.g., see Walker, Hickok, & Fridriksson, 2018), but it is not addressed here. In the remainder of the article, we use the term *naming ability* to refer to a unidimensional latent trait indexing the overall ability to retrieve object names. The term *anomia severity* is conceptually interchangeable, though oriented in the opposite direction.

Despite the widespread use of confrontation picture-naming tests, their utility for quantifying anomia is limited by a number of factors. First, currently available tests have been developed under classical test theory, which often depends on unrealistic assumptions (Embretson & Reise, 2000). For example, score precision for naming tests in current use is commonly expressed as a single standard error of measurement that is constant regardless of the ability level of the patients being tested. This assumption ignores that measurement error varies as a function of ability level and the difficulty of the items in a test; conceptually, closer alignment of items and patients in terms of difficulty and severity, respectively, results in higher measurement precision. To illustrate this point, consider a naming test that includes only easy items. If such a test were administered to a group of individuals with severe aphasia, their scores on the test would vary depending on their level of severity and rank ordering them would be possible. However, if the same test were administered to a group of mildly impaired individuals, it is likely that they would all receive perfect scores and rank ordering of their scores would not be possible.

None of the currently available anomia tests developed for clinical or research use in aphasiology can provide accurate standard errors of measurement for estimates across the ability continuum. The Test of Adolescent/Adult Word Finding (German, 1990) was developed using item response theory (IRT; described below), and its manual provides ability-dependent standard errors; however, its norms are based on a general population sample rather than on a sample of persons with aphasia (PWAs), thus limiting its validity and utility for measuring anomia in aphasia. In addition, del Toro et al. (2011) developed a short form of the BNT using IRT and data from PWAs but did not provide scoring tables or any user-friendly way of obtaining standard errors conditioned on ability.

Another notable limitation of classical test theory is that each test must typically be given in its entirety to obtain diagnostic information, thus contributing to increased testing burden and patients' fatigue and frustration. In turn, fatigue and frustration can contaminate performance, leading to inaccurate clinical conclusions. In some cases, short forms have been developed (del Toro et al., 2011; Walker & Schwartz, 2012) to address this limitation. However, all of the currently available short forms of anomia tests are static, as they include a priori selected items that are administered to all patients regardless of their level of impairment. For example, Walker and Schwartz (2012) developed and investigated the performance of two static, 30-item short forms of the PNT (Roach et al., 1996) that typically requires the administration of 175 items. Based on empirical data from 25 PWAs, Walker and Schwartz reported that the magnitude of the relationship between the two forms and the full PNT were .93 and .98. Furthermore, Walker and Schwartz reported that there was no significant difference between the variances associated with the short forms and the full PNT. However, they also reported a statistically significant difference in means between the logit-transformed accuracy scores from the short forms and the full PNT. Importantly, Walker and Schwartz provided recommendations for making statistical inferences based on the difference of two scores under the assumption of uniform measurement precision regardless of ability level, which, as discussed earlier, is an assumption that is unrealistic in most testing applications.

In practice, static tests almost always provide the highest precision for patients of moderate severity and increasingly poorer precision for patients with more extreme test scores. Furthermore, because short static forms include only a small number of items and those items are selected so that they are relevant to a wide spectrum of ability, static short forms typically do not reach high levels of precision for patients of any ability level. Some tests have basal or ceiling rules (e.g., discontinue after eight consecutive incorrect responses), but there is often a lack of documentation of the theoretical motivation and empirical support for such rules. In addition, basal and ceiling rules assume that items become progressively more difficult as a function of their presentation order in a test, but empirical studies

suggest that successive items on the BNT do not necessarily exhibit a monotonic increase in difficulty (Brookshire & Nicholas, 1995; Pedraza, Sachs, Ferman, Rush, & Lucas, 2011). As described below, IRT provides methods for both modeling precision as a function of ability and efficiently reducing the length of a test without sacrificing reliability.

IRT (De Ayala, 2013; Lord, 1980) is a psychometric approach that formalizes the relationship between latent abilities and test takers' observed patterns of behavior using statistical models. In contrast to observable behaviors such as confrontation naming responses, a hypothetical latent ability such as naming ability cannot be directly measured but must be inferred on the basis of the test taker's observed responses. Unlike a classical test theory, which focuses on test-level properties, IRT emphasizes item-level characteristics, such as item difficulty, and seeks to model the responses of individual persons to individual items. The one-parameter logistic IRT model, so named because it includes a single-item parameter, difficulty, predicts the probability of a correct response to an item as a logistic function of the difference between a person's ability and the item's difficulty. Given a string of responses by a particular person to a set of test items and difficulty estimates for those items calibrated to a common latent trait metric, the model can be used to estimate the person's ability on that metric. One notable property of IRT-based tests is that the testing instrument no longer rests on overall proportion correct as the measure of anomia severity nor does it require the administration of the same items. Different subsets of items can be administered on various testing occasions with ability scores provided on the same metric. In this context, the item difficulty estimates have typically been obtained by fitting the model to item response data provided by a sufficiently large sample of individuals from the population of interest and verifying the tenability of the model's assumptions given the data. For dichotomous responses, item difficulty is traditionally defined as the point on the latent trait scale where 50% of individuals with that ability estimate obtain a correct response. Fergadiotis et al. (2015) provide an introduction to IRT in the context of confrontation naming tests for assessing anomia. For a more general and complete presentation, see De Ayala (2013) and Embretson and Reise (2000).

A major advantage of IRT is that it can be used to support item banking and computer adaptive testing (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). An item bank is a set of items that measure a common latent ability and have been calibrated to a common scale jointly defined by person ability and item difficulty (Vale, 2006). An IRT-based computer adaptive test (CAT) is based on an algorithm that selects and administers only those items targeted to maximize statistical information at a particular ability level from an item bank (Thompson, 2009). Information is a central concept in IRT and formalizes the notion that items that are too easy or too hard for a given test taker provide very limited information for ability estimation. Information can be estimated conditional on ability and expressed as an *item information function*. The item

information function reflects the information (i.e., increase in precision) an item contributes at any level of the latent trait for the estimation of a person's ability and is denoted by $I_i(\theta)$, where i indexes the item. For the one-parameter logistic model (De Ayala, 2013; Rasch, 1993), which is implemented in the current study, item information is maximized at the location of the difficulty parameter δ_i of the item and is defined as

$$I_i(\theta) = P_i(\theta)Q_i(\theta) \quad (1)$$

where $P_i(\theta)$ is the probability of a correct response at a given θ level and $Q_i(\theta)$ is its complement. When multiple items are administered, information has an additive nature. The test information function represents the sum of the item information estimates of all the items at a given ability level:

$$I(\theta) = \sum_i I_i(\theta) \quad (2)$$

Importantly, information is the reciprocal of the error variance with which a participant's ability level can be estimated. The standard error of measurement (*SEM*) at a given θ level can be defined as

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (3)$$

The *SEM* can be used to build 95% confidence intervals (95% CI) around parameter estimates, given that it is expressed on the same metric as the latent trait level:

$$95\% \text{ CI} = \text{Ability Estimate}(\theta) \pm 1.96 \times SEM(\theta) \quad (4)$$

For a graphical representation of item and test information functions, as well as conditional *SEM*, see Hula et al. (2015). When a Bayesian ability estimation algorithm is used, as in the present investigation, the *SEM* is referred to as the posterior standard deviation (PSD), and the associated confidence interval can be interpreted as a Bayesian credible interval.

A basic CAT administration is an iterative process that includes two steps. For each iteration, first, the patient's ability is estimated based on the patient's prior responses. Then, given the calibrated item difficulty parameters, the CAT engine selects the next item to maximize information at the new ability estimate. Typically, this process reiterates until a stopping rule is satisfied (e.g., a specific number of items has been administered). The technique quickly converges into a sequence of items bracketing the test taker's ability while ignoring items that are too easy or too hard and thus efficiently shortening the test while maximizing its precision. Because IRT-based person score estimates are independent of the particular items administered, different

sets of items may be given to different examinees with scores that are expressed on a single common scale.

One obvious potential application of the aforementioned statistical machinery is to develop a CAT for assessing anomia in order to reduce response burden and administration time without sacrificing reliability. Furthermore, such a test would allow for the estimation of appropriate credible intervals tailored to the level of ability of the test taker. Using archived data (Mirman et al., 2010), Hula et al. (2015) used the calibrated PNT item difficulty estimates reported by Fergadiotis et al. (2015) to test a CAT algorithm for the PNT via simulations. The real response strings from the archived data, which were collected in standard administrations, were used to simulate the CAT procedure (Hula et al., 2015). The authors implemented a sampling algorithm that selected from the data set the response to a particular item whenever the CAT algorithm selected that item for administration. Thus, the CAT ability estimates were based on subsets of the responses given to the full test, and the simulation permitted the authors to estimate how a participant would have scored had they taken only the PNT-CAT. To reduce the dependency between the PNT-CAT and the full PNT when estimating the comparison metrics, the authors compared the short-form ability estimates to estimates obtained from the nonoverlapping items remaining in the data set after each short-form administration (referred to as the complement analysis). The results suggested that a 30-item PNT-CAT could provide estimates of naming ability that are more accurate and precise than existing static short versions. The estimates from the simulated administrations correlated highly with the scores on the full PNT (0.95–0.97).

However, performance of the PNT-CAT has thus far only been assessed via simulations based on archival data, and therefore, estimates of its relative agreement with the full PNT may represent an upper bound. For example, simulations ignore error stemming from inaccurate online scoring of responses or error stemming from patients' inherent intrapersonal fluctuations in performance (e.g., attention lapses, fatigue). Furthermore, previously published work focused exclusively on group-level analyses. However, it is important to assess the ability of the PNT-CAT to generate precise scores at the individual level in order to make it useful in clinical settings. Therefore, a new study was conducted using live administrations and online scoring of the computer adaptive version of the PNT.

The main goal of this study is to verify that the PNT-CAT will provide precise and meaningful ability estimates with minimal response burden in a prospective, independent sample of PWAs. Specifically, at the group level, we hypothesized that (a) the PNT-CAT will be highly correlated with the full PNT scores. In addition, we hypothesized that (b) PNT-CAT scores would demonstrate a lack of bias corroborated by an absence of statistically significant differences between the means of the scores generated by the two versions of the test, and (c) there would not be systematic differences in the standard deviations of the

scores produced by the two versions of the test. We further hypothesized (d) that the total error between the two PNT versions, operationalized as the root mean square of the individual pairwise differences (RMSD), would be equal to or less than the value obtained in the Hula et al. (2015) simulation study. Also, given the predicted lack of bias, we hypothesized that (e) variable error, operationalized as the standard deviation of the individual difference scores, would account for the vast majority of the total error. At the individual level, we hypothesized that (f) the individual pairwise differences between the two versions of the test would exceed the measurement error predicted by the IRT model at a rate less than or equal to chance. We also hypothesized that (g) the variable error between the two versions would be well predicted by the conditional test information function derived from the IRT model.

Method

Participants

Currently, data from 47 monolingual PWAs have been collected and analyzed as a part of a larger ongoing project. Fifteen women and 32 men participated in the study. Forty-four participants self-identified as White Caucasian, and three as African American. These participants were recruited from the greater metropolitan regions of Portland, Oregon; Seattle, Washington; and Pittsburgh, Pennsylvania. The inclusion criteria were (a) single or multiple strokes restricted to the left hemisphere ≥ 6 months postonset, resulting in aphasia; (b) performance below the normal cutoff for ≥ 2 modality scores on the Comprehensive Aphasia Test (Swinburn, Porter, & Howard, 2004); (c) no speech-language treatment for 2 weeks prior to or during the study; (d) English as a first language; and (e) corrected near visual acuity of $\geq 6/12$ on the Tumbling E Chart (Lotery et al., 2000). The exclusion criteria were (a) history of head trauma resulting in loss of consciousness with significant cognitive sequelae, (b) history of psychotic or neurological disorder other than stroke or right hemisphere stroke, (c) chronic conditions likely to impair cognition (e.g., renal or hepatic failure), (d) Comprehensive Aphasia Test Recognition Memory *T* score of < 32 (raw score $< 4/10$), and (e) severe apraxia of speech as measured by the Apraxia of Speech Rating Scale (Strand, Duffy, Clark, & Josephs, 2014). Descriptive data are presented in Table 1.

Procedure

Test Procedures

Both the full PNT and the PNT-CAT were administered using a Java-based software application written by the fifth author. The software used the PNT item parameter estimates reported by Fergadiotis et al. (2015) and the Bayesian expected a posteriori scoring (Bock & Mislevy, 1982; De Ayala, 2013) to generate latent trait-naming ability estimates from the dichotomously scored responses.

Table 1. Sample descriptive statistics.

	Education, years	Age, years	Months postonset	CAT modality mean
<i>M</i> (<i>SD</i>)	15.0 (3.3)	65.6 (12.5)	88.7 (90.3)	50.4 (8.7)
Min–max	6–23	29–93	10–500	27.8–65.5

Note. CAT = Comprehensive Aphasia Test.

Based on Hula et al. (2015), the population distribution was assumed to have a mean of 0 and an *SD* of 1.48.

For the full PNT, items were administered in random order. This decision was motivated by the fact that the IRT model assumes that item presentation order does not affect performance. Fergadiotis et al. (2015) found no effect of item position on item difficulty in the archived data (Mirman et al., 2010) collected according to standard administration procedures, suggesting that this assumption was tenable for that data set. Random administration order for the full PNT was selected in the present investigation to begin building a database that could be compared to the Mirman et al. (2010) data as a more robust test of this assumption.

For the PNT-CAT, the test began with a provisional ability estimate of zero and presented the item (“pumpkin”) that provided the most information at that level. Following the response, the examiner inconspicuously entered a dichotomous correct/incorrect judgment via keyboard. The software then updated the ability estimate and its PSD, selected the item that provided the most information at this updated ability level, and presented it as the next item. This procedure was reiterated until 30 items had been administered. At the conclusion of each PNT test administration, the software generated a text file containing the stimuli, the dichotomous response codes, time stamps for each item, interim and final ability score estimates, and the PSDs of the score estimates. As noted before, PSDs are Bayesian analogues of frequentist standard errors and can be used to develop credible intervals about score estimates and their differences.

Both versions of the PNT were administered according to the published guidelines of the test (Roach et al., 1996) with two exceptions. The first, noted above, was that the full PNT was administered in random order to each participant. Second, to minimize learning effects, no feedback was provided after a participant had finished responding to an item. When judging the accuracy of a response, mild distortions of phonemes that did not cross phonemic boundaries (e.g., due to apraxia of speech or dysarthria) were ignored, and the lenient scoring rule for participants with apraxia of speech was not applied. For 35 participants, two trained examiners were present during PNT-CAT administrations and participated in scoring decisions. When the two examiners disagreed on a scoring decision, the test administration was briefly halted while they reviewed the audio recording and reached consensus. The PNT-CAT administrations for the remaining participants were scored by a single examiner, as were full PNT administrations to all

participants. The examiners included three students enrolled in the graduate program at Portland State University (PSU), three American Speech-Language-Hearing Association (ASHA)–certified speech-language pathologists, and one trained research associate who conducted the data collection at the VA Pittsburgh Healthcare System (VAPHS), and one ASHA-certified speech-language pathologist at the University of Washington.

All PNT test responses were audio-recorded, and these recordings were used for offline error coding and reliability checking. The current analyses were performed on the accuracy data entered online by the examiner.

Study Design

Data were collected over two sessions separated by 9–84 days¹ (*Mdn* = 14). Participants were pseudorandomized to receive either the full PNT or the PNT-CAT in their first session. In the sample of participants reported here, 20 participants received the full PNT first, and 27 received the PNT-CAT first.

Analysis

The data were prepared for statistical analysis following guidelines from Kline (2011) and Tabachnick and Fidell (2007) in R (R Core Team, 2018) and SPSS (IBM Corp., 2017). Ability estimates from the PNT-CAT and the full PNT were assessed in terms of the assumptions of normality, homoscedasticity, linearity, and the presence of outliers.

Three additional sets of preliminary analyses were performed. First, we evaluated interrater agreement for the dichotomous item-level scores. Groups of at least three trained research assistants at the PSU and VAPHS sites independently listened to recordings of the PNT administrations and scored each item. Disagreements among offline scorers were resolved by consensus, except in cases where they hinged on differences in phonetic transcription, in which case they were resolved by majority rule based on their independent transcriptions. Three cases from each of the three data collection sites were randomly selected, and the offline consensus scoring for each PNT administration was compared to the online scoring using Cohen’s kappa, as implemented in the R package ICC (Wolak & Wolak, 2015). The purpose of the second set of preliminary

¹Excluding the single participant with 84 days elapsed between testing sessions, the maximum time elapsed was 31 days.

analyses was to determine whether there were any random or fixed effects of data collection site that needed to be included in the statistical modeling of agreement between the two PNT versions. These analyses consisted of fitting and comparing single-level and multilevel regression models using the lme4 (Bates, Maechler, Bolker, & Walker, 2015) and AICcmodavg (Mazerolle & Mazerolle, 2017) packages in R. Finally, we analyzed the average administration time of the two versions of the PNT to quantify the gains associated with administering the PNT-CAT over the full PNT.

The primary analysis was conducted in a Bayesian framework using Stan (Carpenter et al., 2017) and rStan (Stan Development Team, 2017). The model took input from the paired full PNT and PNT-CAT score estimates provided by the administration software. It assumed that the score estimates were distributed multivariate normally, with five estimated parameters (two means, two standard deviations, and one correlation). Each parameter was given dispersed starting values and a vague prior, thus allowing the Bayesian estimation process to explore the full parameter space and provide conservative estimates of posterior distributions (McElreath, 2018). The Stan code, including the prior specifications, is provided in Supplemental Material S1. Four Hamiltonian Markov chain Monte Carlo (MCMC) chains were run with 3,000 iterations discarded as warm-up and 5,000 iterations monitored for convergence and parameter estimation. MCMC convergence was assessed graphically by inspection of the autocorrelation and trace plots and statistically using the Gelman–Rubin potential scale reduction statistic ($\hat{R}$) and the number of effective samples.

The $\hat{R}$ statistic is a ratio of the variance within each chain to the variance pooled across chains, with values close to 1 indicating satisfactory convergence of the chains to a stable distribution (Gelman et al., 2013). The number of effective samples factors out the autocorrelation in the observed MCMC chains and estimates the number of independent samples that would provide the same degree of precision for the parameter estimates (Carpenter et al., 2017). A large number of effective samples suggest convergence.

The posterior distributions for all estimated and derived parameters were summarized by the mean, median, mode, and 95% highest posterior density interval (95% HDI). The posterior median was taken as the point estimate for each parameter. The HDI is comparable to the frequentist confidence interval and is determined as the narrowest interval containing the assigned proportion of the posterior distribution's probability mass within which all values have a higher probability density than any values outside the interval.

In addition to the correlation, several parameters were generated from the Bayesian model to answer questions of interest. These parameters included total error (RMSD), constant error or bias, and variable error. RMSD is an index of absolute agreement and is calculated as

$$\text{RMSD}(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^n (\hat{\theta}_j - \theta_j)^2}{n}} \quad (5)$$

where $\hat{\theta}_j$ is the ability estimate from the PNT-CAT for person j , θ_j the estimate from the full PNT, and n is the number of participants ($n = 47$). Low RMSD values indicate more accurate measurement based on agreement with the full PNT estimates. RMSD yields meaningful results when the errors ($\hat{\theta}_j - \theta_j$) follow a normal distribution (Chai & Draxler, 2014), an assumption that we tested using the Anderson–Darling omnibus test for the composite hypothesis of normality (Thode, 2002). RMSD represents the total error between the two PNT versions and can be decomposed into constant error (bias), which is the difference in means calculated as the average signed difference score, and variable error, which is the standard deviation of the difference scores (Schmidt & Lee, 2005). In the present Bayesian model, we estimated variable error as

$$\text{VE}(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta}) + \text{var}(\theta) - 2\text{cov}(\hat{\theta}, \theta)} \quad (6)$$

where $\text{var}(\hat{\theta})$ and $\text{var}(\theta)$ are the variances of the PNT-CAT and full PNT ability estimates, respectively, and $\text{cov}(\hat{\theta}, \theta)$ is their covariance. RMSD was estimated in the Bayesian model as the square root of the sum of the squared constant error and bias. Finally, we derived the difference between the estimated standard deviations of the two test versions.

A second set of analyses was undertaken to further evaluate agreement between the two test versions at the individual level. In order to evaluate whether disagreement in individual cases exceeded that predicted by the measurement model, we computed a 95% credible interval for each of the 47 observed pairwise difference scores using the conditional PSDs of the individual score estimates. The formula for this credible interval was

$$95\% \text{ CI} = 1.96 \pm \sqrt{\text{PSD}_{\text{PNT}}^2 + \text{PSD}_{\text{PNT-CAT}}^2} \quad (7)$$

where PSD_{PNT} and $\text{PSD}_{\text{PNT-CAT}}$ are the conditional PSDs for the full PNT and PNT-CAT score estimates, respectively. The square root term in this equation represents the conditional PSD about the pairwise difference score, and the average conditional PSD about the pairwise differences represents a model prediction for variable error, the standard deviation of the observed pairwise differences.

Finally, we computed the empirical marginal reliability for the full PNT and the PNT-CAT. Marginal reliability

is conceptually similar to Cronbach's alpha and is estimated as a function of the total variance of the ability estimates and their error variance (i.e., the average squared PSDs of the ability estimates) based on the following formula:

$$r = 1 - \frac{\sigma_e^2}{\sigma_0^2} \text{ (Thissen, 2000). To create the 95\% HDI of}$$

the marginal reliability estimates, we used Bayesian bootstrapping as implemented in the bayesboot R package (Bååth, 2018). Furthermore, we plotted the conditional *SEM* (i.e., PSDs of the ability estimates) that can be used to obtain unique 95% CIs for each ability estimate. Given that the precision of the test varies as a function of ability, we view the conditional *SEM* as a more informative quantity for test score interpretation. Marginal reliability is provided primarily for comparing the IRT versions of the PNT presented in this article to other confrontation naming tests developed under classical test theory.

Results

Preliminary Analyses

Visual inspection and examination of skewness and kurtosis statistics for the PNT and PNT-CAT score estimates revealed that both variables were moderately skewed and minimally to moderately kurtotic, but no gross violations were noted. Close inspection of the bivariate scatter plot (see Figure 1) did not suggest violations of the assumptions of homoscedasticity or linearity. Furthermore, no univariate or multivariate outliers were identified using *z* scores and Mahalanobis' distance. The Anderson–Darling test, used to assess whether the differences between the two versions of the test were normally distributed, was not statistically significant, $A = 0.496$, $p = .203$.

Cohen's kappa coefficients and point-to-point agreement are presented in Table 2. Overall, interrater agreement for item-level dichotomous scoring ranged from strong to excellent based on McHugh (2012) with one exception. For one participant, interrater agreement was only moderately strong for the PNT-CAT, as indexed by a kappa estimate of 0.60 and point-to-point agreement of 80% resulting from disagreements on six items. In this case, all six disagreements were attributable either to self-corrected responses for which it was difficult to identify the first complete attempt, a vowel distortion versus a substitution error that increased the difficulty of the perceptual judgment, or both. Despite the lower-than-optimal point-to-point agreement, both sets of raters produced the same overall score (16 of 30 correct) and thus the same ability estimates. This fact, combined with the excellent targeting of the CAT items in this particular case (an optimally targeted CAT would produce 50% correct), suggests that these scoring disagreements had negligible impact on the overall score estimate.

Chi-square tests comparing a single-level model regressing full PNT scores on PNT-CAT scores with multilevel models including (a) intercepts varying by data collection site and (b) intercepts and slopes varying by site yielded *p* values of $> .99$. The single-level model obtained a lower sample

Figure 1. Scatter plot of naming ability score estimates from the full Philadelphia Naming Test (PNT) over estimates from the PNT-CAT30. The shaded region represents the 95% confidence interval about the hypothesis that the two PNT versions would provide equal score estimates for each individual case, which is represented by the dashed identity line. The three cases displayed in blue obtained differences whose 95% confidence intervals excluded zero. CAT = computer adaptive test.

size-corrected AIC value (58.88) than either the varying intercept model (69.11) or the varying intercept and slope model (74.08). Furthermore, in the varying intercept model, the proportion of random variance in full PNT scores accounted for by data collection site was 0.03, and the proportion of total variance in full PNT scores accounted for by site was 0.003. These results permitted us to conduct the primary analyses in a Bayesian framework using a straightforward model of the means, standard deviations, and correlation for the two variables without the need to specify hyperpriors in terms of parameters for priors at the site level.

Finally, we analyzed the average administration time of the two versions of the test. On average, the PNT-CAT required 8.26 min ($SD = 3.62$) to administer, with approximately 75% of the participants completing the test in 10 min or less; only three cases required between 15 and 18 min to complete the test. As expected, the full PNT was associated with substantially longer administration times (on average 31.13 min, $SD = 11.72$) that were evenly distributed around the mean.

Main Analyses

Group Level

The trace plots for all parameters, provided in Supplemental Material S2, demonstrated rapid convergence,

Table 2. Interrater reliability for item-level dichotomous scores for three randomly chosen participants from each of the three data collection sites.

Participant	PNT-CAT30			Full PNT		
	% Agreement	Kappa	95% CI	% Agreement	Kappa	95% CI
PSU1	90	0.80	[0.58, 1.00]	94	0.85	[0.77, 0.94]
PSU2	90	0.78	[0.56, 1.00]	91	0.81	[0.72, 0.90]
PSU3	97	0.93	[0.80, 1.00]	95	0.91	[0.86, 0.97]
UW1	93	0.85	[0.65, 1.00]	94	0.86	[0.78, 0.94]
UW2	100	1.00	[1.00, 1.00]	98	0.90	[0.79, 1.00]
UW3	97	0.89	[0.68, 1.00]	95	0.83	[0.73, 0.94]
PGH1	98	0.93	[0.79, 1.00]	98	0.94	[0.88, 1.00]
PGH2	97	0.93	[0.79, 1.00]	94	0.86	[0.78, 0.95]
PGH3	80	0.60	[0.30, 0.90]	91	0.72	[0.57, 0.86]

Note. PNT = Philadelphia Naming Test; CAT = computer adaptive test; CI = confidence interval; PSU = Portland State University; UW = University of Washington; PGH = VA Pittsburgh Healthcare System.

stationarity relative to their means, and good mixing. The autocorrelation plots, also provided in Supplemental Material S2, corroborated this assessment and showed minimal autocorrelation beyond lag 5. The ($\hat{R}$) statistic and number of effective samples for each parameter are provided in Table 3. These also indicated satisfactory convergence and mixing.

Table 3 also provides the point estimates and 95% HDIs for each parameter. Histograms of the posterior distributions for the PNT-CAT and full PNT means and standard deviations are provided in Figure 2. The posterior distributions for the correlation, bias, variable error, and total error are displayed in Figure 3. As expected, the two versions of the PNT were strongly correlated ($r = .95$, 95% HDI [.92, .97]), with negligible difference in their means (bias = 0.06, 95% HDI [−0.08, 0.18]) or standard deviations (−0.07, 95% HDI [−0.20, 0.06]). Furthermore, the total error or RMSD between the score estimates provided by the two versions was consistent with predictions at 0.46 (95% HDI [0.36, 0.56]). Also, given the negligible bias, the total error was composed almost entirely of variable error, estimated as the standard deviation of the individual difference scores (0.45, 95% HDI [0.36, 0.56]).

Individual-Level Analyses

Of the 47 cases, three obtained pairwise differences whose 95% CI excluded zero. These results are displayed in Figure 1. In this scatter plot, the dashed line is the identity line, representing the hypotheses that full PNT and PNT-CAT scores are equal. The shaded region represents the 95% CI about this hypothesis, developed from the pairwise observations and their conditional PSDs according to procedures described by Wright and Stone (1979). The widening of the shaded area toward the margins of the plot reflects the fact that the PNT and the CAT item bank constructed from it provide less information for PWAs with more extreme high or low naming ability. The three observations that obtained differences credibly larger than zero are outside the shaded 95% CI region and displayed in blue. These three cases represent 6.3% of the sample (95% HDI [1.6%, 15.8%]), and this proportion is not credibly larger than the 5% of cases that would be expected to show differences this large or larger.

We also compared the average of the conditional PSDs for the individual pairwise differences derived from the IRT model test information function to the posterior estimate of variable error, the standard deviation of the

Table 3. The posterior mean, standard error (SE) of the mean, standard deviation (SD), median (*Mdn*), 95% highest posterior density interval (HDI), $\hat{R}$ statistic, and number of effective samples for each parameter in the Bayesian analysis model.

Parameter	<i>M</i>	<i>SE mean</i>	<i>SD</i>	<i>Mdn</i>	95% HDI	$\hat{R}$	<i>n</i> Effective samples
PNT-CAT mean	0.61	< 0.005	0.20	0.61	[0.22, 1.01]	1.00	8,764
Full PNT mean	0.55	< 0.005	0.21	0.55	[0.14, 0.97]	1.00	8,611
PNT-CAT SD	1.38	< 0.005	0.14	1.37	[1.12, 1.68]	1.00	7,485
Full PNT SD	1.45	< 0.005	0.15	1.44	[1.18, 1.77]	1.00	7,208
PNT-CAT SD – full PNT SD	−0.07	< 0.005	0.07	−0.07	[−0.20, 0.06]	1.00	20,000
Correlation	0.95	< 0.005	0.02	0.95	[0.92, 0.97]	1.00	8,820
Bias	0.06	< 0.005	0.07	0.06	[−0.08, 0.18]	1.00	20,000
Variable error	0.45	< 0.005	0.05	0.45	[0.36, 0.56]	1.00	14,469
Total error	0.46	< 0.005	0.05	0.46	[0.36, 0.56]	1.00	13,982

Note. PNT = Philadelphia Naming Test; CAT = computer adaptive test.

Figure 2. Histograms of the posterior distributions for the full Philadelphia Naming Test (PNT) and PNT-CAT means and standard deviations. HDI = highest posterior density interval; CAT = computer adaptive test.

observed difference scores derived from the Bayesian analysis model. As noted above, the former represents the IRT measurement model's prediction of the standard deviation of the differences between the two PNT versions, and the latter represents the same quantity estimated from the observed score distributions. The average conditional PSD was 0.38, and the point estimate for the standard deviation of the observed differences was larger, though not robustly so, at 0.45 (95% HDI [0.36, 0.56]).

Finally, Figure 4 shows the conditional PSDs for the full PNT (red) and the PNT-CAT (black). The trendlines were estimated using fifth degree polynomials fitted to the data from the 47 cases reported in this manuscript. As expected, the full PNT was associated with smaller PSDs across the ability continuum compared to the PNT-CAT given that the former was based on the full item bank. Furthermore, the PNT-CAT exhibited almost uniform PSDs for the range of ability estimates between -1.5 and 1.2 , which reflects the availability of several items in the item bank for estimating ability in that region. Beyond

this region, precision decreased monotonically for participants the more extreme their ability estimates were. Marginal reliability for the PNT-CAT was .94 (95% HDI [.91, .96]); for the full PNT, marginal reliability was .98 (95% HDI [.97, .99]).

Discussion

The purpose of this study was to evaluate the performance of an IRT-based CAT version of the PNT. This test is often used in research applications, but its utility, especially in fast-paced clinical settings, is limited because it requires the administration of the large number of items. Prior real-data simulations had suggested that a 30-item PNT-CAT may provide accurate estimates of naming ability, but the results were not verified empirically. To this end, we conducted a new study with data obtained from a new independent sample of 47 PWAs. We evaluated the performance of the CAT version of the PNT and its agreement with the full PNT based on multiple criteria, including relative agreement, bias, variable error, total

Figure 3. Posterior distributions for the correlation, bias, variable error, and total error. HDI = highest posterior density interval.

error, homogeneity of variance, and precision for measuring differences at the individual level as predicted by the measurement model.

First, the PNT-CAT appears to be a strong, viable candidate for use in group studies that include anomia severity either as a descriptive variable or as a predictor or outcome variable in the main analyses. Based on previous simulation work, we had hypothesized that the PNT-CAT would be highly correlated with the full PNT scores. The high relative agreement observed in this study, as indicated by the Pearson product-moment correlation, is evidence that scores from either the full PNT or the PNT-CAT could be used practically interchangeably in group studies that model variance/covariance matrices or derive the strength and nature of the relationships of variables directly from the data (e.g., canonical correlation and its special cases, or structural equation modeling). Furthermore, we had predicted the absence of statistically significant differences between the means of the scores generated by the full PNT and the PNT-CAT and that there would not be systematic differences in the standard deviations of the

scores produced by the two versions of the test. Based on our results, the absence of bias and the homogeneity of variance across the two versions of the PNT, the PNT-CAT suggests that the PNT-CAT is adequate for group studies that focus on group mean comparisons, such as analysis of variance, multivariate analysis of variance, and structural equation models that supplement the variance/covariance matrix with a mean structure.

In addition, the current findings agree with the real-data simulation results published in Hula et al. (2015), who conducted the simulation study based on archival PNT response data from 251 participants from the Moss Aphasia Psycholinguistics Project Database (Mirman et al., 2010). Hula et al. reported a frequentist correlation of .950 (95% CI [.937, .960]) and an RMSD of 0.478 (95% CI [0.433, 0.535]). In the current study and consistent with our predictions, our Bayesian analysis yielded a correlation of .951 (95% HDI [.918, .974]) and an RMSD of 0.458 (95% HDI [0.364, 0.563]). The differences in the confidence/credible intervals can be attributed most likely to the difference in sample size based on which the

Figure 4. Scatter plot of posterior standard deviation as a function of estimated ability from the full Philadelphia Naming Test (PNT) and the PNT-CAT. The trend lines were generated by fitting fifth degree polynomials to the data. CAT = computer adaptive test.

analyses were conducted (251 participants vs. 47 participants). Taken together, these findings add converging evidence regarding the expected performance of the PNT-CAT algorithm as a viable substitute of the full PNT for estimating anomia severity. Furthermore, they provide support for the practical utility of real-data simulations for predicting with precision the behavior of CAT algorithms.

Our results are also comparable to the findings published in Walker and Schwartz (2012), who studied the performance of the two static, 30-item short forms of the PNT. Walker and Schwartz reported that the magnitude of the relationship between the two forms and the full PNT were .93 and .98. No confidence intervals were provided, but we estimated them based on the magnitude of the correlation reported, the sample size, and a *r*-to-*z* transformation according to which the 95% CIs were [.846, .969] and [.955, .991]. These results regarding relative agreement are similar to the results presented in the current study. Furthermore, similar to the PNT-CAT, Walker and Schwartz reported that there was not a significant difference between the variances associated with the short forms and the full PNT. However, whereas our IRT CAT algorithm yielded PNT-CAT scores with no bias, Walker and Schwartz reported a small yet statistically significant difference in means between the logit-transformed accuracy scores from the short forms and the full PNT. Furthermore, in part because of the statistically significant mean differences, Walker and Schwartz estimated a regression equation to convert a static PNT30 score to a full PNT score and provided recommendations on how to interpret converted scores to the percentile norms derived from Moss Aphasia Psycholinguistics Project Database. One notable advantage of our approach is that IRT scores, as well as their credible

intervals, both from the PNT-CAT and the full PNT are by design on the same scale facilitating interpretation and obviating the need of converting scores using a predictive model.

Importantly, the CAT algorithmic approach has additional advantages compared to static testing instruments developed with classical test theory. First, such static forms include a priori determined items that typically maximize precision for a particular region of the ability spectrum. Therefore, increasing measurement precision for other regions of ability requires the development of new forms. Alternatively, one could create forms that include items that span the full difficulty spectrum and are thus informative for PWAs with different levels of severity. However, the latter approach increases the likelihood that PWAs will be presented with a large number of items that are not well tailored to their level of ability and could render the assessment instrument inefficient. On the other hand, once a CAT engine is established, it is possible to enrich the item bank from which the CAT algorithm selects items. An enriched item bank would allow for higher precision estimates for PWAs with fewer items delivered, because more items would be available for specific ability regions. This is of specific importance for future assessment of PWAs with relatively mild and more severe naming deficits as the PNT does not have a substantial number of items for these particular ability regions. This can be seen in Figure 4 that demonstrates how measurement precision drops when testing individuals at the two ends of the ability distribution (also see Fergadiotis et al., 2015, for the full test information function). Given that information is a well-defined concept and property of items within the IRT framework, IRT provides the means to engineer an item bank with

desirable characteristics that would maximize the clinical utility of the tool by introducing items primarily for the regions of ability, which currently demonstrate low item coverage.

In order to estimate item difficulty for further items to be included in the item bank, they must first be calibrated within the IRT framework, which requires sample sizes exceeding 200 participants (De Ayala, 2013). An alternate option is the exploration of predictive modeling for item difficulty estimation (Gierl & Haladyna, 2012) using readily available lexical variables. To this end, we (Fergadiotis, Swiderski, & Hula, 2018) have developed a regression model accounting for a substantial proportion of the variance of the difficulty parameters of the items on the PNT ($\text{adj. } R^2 = .692$). Future directions for both the algorithmic CAT and this predictive model should be focused on the estimation of item difficulty parameters for a novel picture-naming task with the predictive model and assessing their ability to reliably estimate naming ability on a large independent sample of PWAs with a CAT. In addition to creating a reliable short form, the development of a large item bank would also support the creation of multiple nonoverlapping forms for repeated assessments. The utility of administering repeated, but directly comparable, nonoverlapping assessments would eliminate learning effects while mitigating fatigue when assessing response to treatment in both research and clinical settings while explicitly measuring patients' latent naming ability in the statistically robust IRT framework.

We also conducted a post hoc analysis to study the performance of the CAT algorithm for increasing number of items. First, we focused on the relative agreement of the CAT-IRT ability estimates with the ability estimates derived from the full PNT as a function of increasingly longer PNT-CAT administrations; then, we focused on the average precision of the PNT-CAT test scores as a function of the number of items administered during the CAT. The first set of results for this analysis are presented on Figure 5 (left), which shows the correlations between ability estimates from the PNT-CAT of various lengths and the full PNT. As expected, the relative agreement between the two sets of scores increases monotonically as a function of the number of items administered. Interestingly, after the administration of only nine items, the correlation between the PNT-CAT and the full PNT ability estimates exceeds .90, which is considerably high and suggests that the CAT algorithm could potentially yield useful scores for screening purposes using only a handful of the PNT items. After 20 up to 30 items, the gains in relative agreement are substantially smaller compared to the initial rate of improvement. Furthermore, Figure 4 (right) shows the average PSD for PNT-CAT lengths 1–30. As expected, they decrease monotonically as information accumulates after the administration of each item. The CAT algorithm makes it possible for the professional end user of an assessment tool to flexibly balance the desired precision of a score estimate with test length, the time available, and the patient burden.

In addition to evaluating group-level agreement between the PNT-CAT and the full PNT, we also assessed the level of agreement for individual cases. We had hypothesized that the individual pairwise differences between the two versions of the test would exceed the measurement error predicted by the IRT model at a rate less than or equal to chance. The analysis in this article revealed that the percentage of cases obtaining statistically robust differences (defined by a 95% credible interval excluding zero), at 6.3%, was well predicted by the IRT model. As an alternative perspective on the same issue, we had hypothesized that the conditional test information function derived from the IRT model would accurately predict the variable error between the two versions of the test. Thus, we compared the average model-predicted PSD of the pairwise differences (0.38) to the analogous quantity estimated from the observed differences, the standard deviation of the observed differences: 0.45 (95% HDI [0.36, 0.56]). These results suggest that the model adequately estimates the measurement error inherent in the score estimates and thus provides clinicians and researchers with an appropriate assessment of their uncertainty, enabling them to draw appropriate conclusions about score differences in individual cases. However, it is also the case that some of the individual differences were large in absolute terms, irrespective of measurement error. The largest pairwise differences were in the range of 0.6–0.8, on a scale where the expected population standard deviation is approximately 1.4. Although there is a nonlinear relationship between the model-estimated logit-scaled scores and the proportion correct, these largest differences generally corresponded to implied proportion correct score differences of 0.15–0.2 and number-correct differences for the full PNT in the range of 25–35. Given that some of these differences occurred for participants with score estimates outside the range optimally targeted by the item bank (i.e., the full PNT), adding items as discussed above may help to reduce the size of these differences, while simultaneously increasing the precision of their measurement.

Another aspect of this study that bears discussion is the variety and qualifications of the individuals who administered and scored the PNT. The examiners included three master's-level speech-language pathology students enrolled in the graduate program at PSU, two ASHA-certified speech-language pathologists and one trained research associate at the VAPHS, and one postdoctoral speech-language pathologist at the University of Washington. Although all examiners were certainly well trained in the administration and scoring of the PNT, particularly with respect to identification of the first complete attempt and of correct responses, there was marked variability in their background and level of experience. Despite this variability, the study obtained strong interrater reliability for item-level scoring decisions and agreement between the two test versions that was consistent with well-informed predictions. We believe that the variety of backgrounds and experience levels among the examiners enhances the potential for these results to generalize to a wide variety of clinical and research settings.

Figure 5. Correlation of the computer adaptive test ability estimates with the ability estimates derived from the full Philadelphia Naming Test (left) and their precision (right) for increasing number of items.

Limitations and Future Directions

One advantage of the full PNT at this time is that, by administering all 175 items of the test, professionals can develop stable anomia profiles with diagnostic utility and implications for prognosis and treatment planning. Although the PNT-CAT can provide a reliable anomia severity estimate in a fraction of the time required for administering the full test, it cannot provide a usefully stable diagnostic anomia profile automatically. One reason for this is that the IRT model upon which it is based assumes that naming ability is a unidimensional trait that varies only along a single continuum from less to more. By contrast, the currently favored theoretical and computational models on which anomia profile may be based assume at least two distinct underlying components to naming behavior (i.e., semantic and phonological). Second, the IRT model takes as its input dichotomously scored responses and thus does not have available to it the information provided by coding of error types. It is also unlikely that a handful of items (e.g., 30) would contain sufficient information to replicate the error profiles generated based on the full test. One direction that we are currently exploring is a modified CAT-IRT algorithm designed to, first, establish a patient's severity level and, then, select and administer items that have a high probability of eliciting errors.

In order to make a computer adaptive naming test useful for generating anomia profiles, it will be necessary to generalize the IRT model in both directions by employing a multidimensional latent trait model that predicts not

only correct versus incorrect responses but also specific error types. Walker et al. (2018) have recently taken a significant step in this direction by fitting a large sample of PNT data to a multinomial processing tree model informed by current cognitive models of word production. However, implementation of CAT driven by a multidimensional model for multinomial data will be significantly more complicated than the relatively simple application in this study. Furthermore, it remains to be seen whether a relatively brief assessment such as the one used in this study can provide enough information to a multidimensional model to permit stable estimation of multiple underlying traits.

Acknowledgments

This study was supported by National Institute on Deafness and Other Communication Disorders Award R03DC014556 (awarded to PI: Fergadiotis). The authors thank the study participants, the Western Pennsylvania Patient Registry, and the following individuals who contributed to data collection, interrater reliability, or both: Emily Boss, Haley Dresang, Maria Fenner, Michelle Gravier, Angela Grzybowski, Brooke Lang, Joann Silkes, Alyssa Verlinich, Lauren Weerts, Hattie Olson, and Kasey Graue.

References

- Bååth, R. (2018). Bayesboot: An implementation of Rubin's (1981) Bayesian bootstrap. *R Package Version 0.2.2*. Retrieved from <https://github.com/rasmusab/bayesboot>

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444. <https://doi.org/10.1177/014662168200600405>
- Brookshire, R. H., & Nicholas, L. E. (1995). Relationship of word frequency in printed materials and judgements of word frequency in daily life to Boston Naming Test of performance of aphasic adults. *Clinical Aphasiology*, 23, 107–119. Retrieved from <http://aphasiology.pitt.edu/id/eprint/194>
- Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14(1), 177–208. <https://doi.org/10.1080/026432997381664>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), <https://doi.org/10.18637/jss.v076.i01>
- Carragher, M., Conroy, P., Sage, K., & Wilkinson, R. (2012). Can impairment-focused therapy change the everyday conversations of people with aphasia? A review of the literature and future directions. *Aphasiology*, 26(7), 895–916. <https://doi.org/10.1080/02687038.2012.676164>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- De Ayala, R. J. (2013). *The theory and practice of item response theory* (2nd ed.). New York, NY: Guilford.
- del Toro, C. M., Bislick, L. P., Comer, M., Velozo, C., Romero, S., Rothi, L. J. G., & Kendall, D. L. (2011). Development of a short form of the Boston Naming Test for individuals with aphasia. *Journal of Speech, Language, and Hearing Research*, 54(4), 1089–1100. [https://doi.org/10.1044/1092-4388\(2010/09-0119\)](https://doi.org/10.1044/1092-4388(2010/09-0119))
- Druks, D. J., & Masterson, J. (2000). *An object and action naming battery*. United Kingdom: Psychology Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, 58(3), 865–877. https://doi.org/10.1044/2015_JSLHR-L-14-0249
- Fergadiotis, G., Swiderski, A. M., & Hula, W. D. (2018). Predicting confrontation naming item difficulty. *Aphasiology*, 33(6), 689–709. <https://doi.org/10.1080/02687038.2018.1495310>
- Fergadiotis, G., & Wright, H. H. (2015). Modelling confrontation naming and discourse performance in aphasia. *Aphasiology*, 30(4), 1–17. <https://doi.org/10.1080/02687038.2015.1067288>
- Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, 43(2), 182–216. <https://doi.org/10.1006/jmla.2000.2716>
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- German, D. J. (1990). *Test of Adolescent/Adult Word Finding*. Allen, TX: DLM.
- Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. New York: Routledge.
- Goodglass, H., & Wingfield, A. (1997). *Anomia: Neuroanatomical and cognitive correlates*. San Diego, CA: Academic Press.
- Herbert, R., Hickin, J., Howard, D., Osborne, F., & Best, W. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology*, 22(2), 184–203. <https://doi.org/10.1080/02687030701262613>
- 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
- IBM Corp. (2017). *IBM SPSS statistics for Windows* (Version 25.0). Armonk, NY: Author.
- Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test* (2nd ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.
- Kertesz, A. (2007). *Western Aphasia Battery—Revised*. New York, NY: Grune & Stratton.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.
- Kohn, S. E., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and Language*, 24(2), 266–283. [https://doi.org/10.1016/0093-934X\(85\)90135-X](https://doi.org/10.1016/0093-934X(85)90135-X)
- Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–38. <https://doi.org/10.3115/992628.992631>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lotery, A. J., Wiggam, M. I., Jackson, A. J., Refson, K., Fullerton, K. J., Gilmore, D. H., & Beringer, T. R. (2000). Correctable visual impairment in stroke rehabilitation patients. *Age and Ageing*, 29(3), 221–222.
- Mazerolle, M. J., & Mazerolle, M. M. J. (2017). *AICcmadv: Model selection and multimodel inference based on (Q)AIC(c). R package Version 2.1-1*. Retrieved from <https://cran.r-project.org/package=AICcmadv>
- McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, 2(4), 751–770. <https://doi.org/10.1111/j.1756-8765.2010.01116.x>
- McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- 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>
- Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10–11), 935–979. <https://doi.org/10.1080/02687030244000563>
- Pedraza, O., Sachs, B. C., Ferman, T. J., Rush, B. K., & Lucas, J. A. (2011). Difficulty and discrimination parameters of Boston Naming Test items in a consecutive clinical series. *Archives of Clinical Neuropsychology*, 26(5), 434–444. <https://doi.org/10.1093/arclin/acr042>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.
- 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. Retrieved from <http://aphasiology.pitt.edu/id/eprint/215>
- Schmidt, R. A., & Lee, T. (2005). *Motor control and learning: A behavioral emphasis* (4th ed.). Champaign, IL: Human Kinetics.
- Schuell, H., Jenkins, J. J., & Jimenez-Pabon, E. (1964). *Aphasia in adults*. New York, NY: Harper & Row.

-
- Stan Development Team.** (2017). *RStan: The R interface to Stan*. (Version 2.16.2). Retrieved from <http://mc-stan.org>
- Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K.** (2014). The Apraxia of Speech Rating Scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders, 51*, 43–50. <https://doi.org/10.1016/j.jcomdis.2014.06.008>
- Swinburn, K., Porter, G., & Howard, D.** (2004). *Comprehensive Aphasia Test*. New York, NY: Psychology Press.
- Tabachnick, B. G., & Fidell, L. S.** (2007). *Using multivariate statistics* (6th ed.). Boston: Pearson/Allyn & Bacon.
- Thissen, D.** (2000). Reliability and measurement precision. In *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.
- Thode, H. C.** (2002). *Testing for normality*. New York, NY: Marcel Dekker.
- Thompson, N. A.** (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778–793. <https://doi.org/10.1177/0013164408324460>
- Vale, C. D.** (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Routledge.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J.** (2000). *Computerized adaptive testing: A primer*. New York: Routledge.
- Walker, G. M., Hickok, G., & Fridriksson, J.** (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological Assessment, 30*(6), 809–826. <https://doi.org/10.1037/pas0000529>
- Walker, G. M., & Schwartz, M. F.** (2012). Short-form Philadelphia Naming Test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology, 21*(2), S140–S153. [https://doi.org/10.1044/1058-0360\(2012/11-0089\)](https://doi.org/10.1044/1058-0360(2012/11-0089))
- Wolak, M., & Wolak, M. M.** (2015). *Package “ICC.” Facilitating estimation of the intraclass correlation coefficient*. Retrieved from <https://github.com/matthewwolak/ICC>
- Wright, B. D., & Stone, M. H.** (1979). *Best test design*. Chicago, IL: MESA Press.