Research Article

Empirical Evaluation of Computer-Adaptive Alternate Short Forms for the Assessment of Anomia Severity

William D. Hula,^{a,b} Gerasimos Fergadiotis,^c Alexander M. Swiderski,^{b,d}
JoAnn P. Silkes,^e and Stacey Kellough^d

Purpose: The purpose of this study was to verify the equivalence of 2 alternate test forms with nonoverlapping content generated by an item response theory (IRT)-based computer-adaptive test (CAT). The Philadelphia Naming Test (PNT; Roach, Schwartz, Martin, Grewal, & Brecher, 1996) was utilized as an item bank in a prospective, independent sample of persons with aphasia.

Method: Two alternate CAT short forms of the PNT were administered to a sample of 25 persons with aphasia who were at least 6 months postonset and received no treatment for 2 weeks before or during the study. The 1st session included administration of a 30-item PNT-CAT, and the 2nd session, conducted approximately 2 weeks later, included a variable-length PNT-CAT that excluded items administered in the 1st session and terminated when the modeled precision of the ability estimate was equal to or greater than the value obtained in the 1st

session. The ability estimates were analyzed in a Bayesian framework.

Results: The 2 test versions correlated highly ($r = .89$) and obtained means and standard deviations that were not credibly different from one another. The correlation and error variance between the 2 test versions were well predicted by the IRT measurement model.

Discussion: The results suggest that IRT-based CAT alternate forms may be productively used in the assessment of anomia. IRT methods offer advantages for the efficient and sensitive measurement of change over time. Future work should consider the potential impact of differential item functioning due to person factors and intervention-specific effects, as well as expanding the item bank to maximize the clinical utility of the test.

Supplemental Material: <https://doi.org/10.23641/asha.11368040>

The cardinal deficit of persons with aphasia (PWAs) is anomia, the difficulty with accessing and retrieving words during language production (Goodglass & Wingfield, 1997; Kohn & Goodglass, 1985; Nickels, 2002). Given its prevalence in the aphasic population, anomia is

typically a focal point of rehabilitation, and the development of new treatment approaches continues to receive substantial attention from researchers. Clinicians and researchers typically rely on confrontation picture-naming tests to quantify anomia severity and assess spontaneous or treatment-related change (e.g., Druks & Masterson, 2000; German, 1990; Kaplan, Goodglass, & Weintraub, 2001). The widespread usage of confrontation naming tests can be attributed to their straightforward administration and scoring procedures and because the ability to access and retrieve words during picture naming depends on cognitive processes that are also critical for successful language production in less constrained contexts (Fergadiotis, Kapantzoglou, Kintz, & Wright, 2018). They have also been found to be good indicators of overall aphasia severity (Schuell, Jenkins, & Jimenez-Pabon, 1964; Walker & Schwartz, 2012), and changes in the ability to name pictures have been linked to improvement in overall communicative functioning (Carragher, Conroy, Sage, & Wilkinson, 2012; Herbert, Hickin, Howard, Osborne, & Best, 2008).

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

^bDepartment of Communication Science and Disorders, University of Pittsburgh, PA

^cDepartment of Speech and Hearing Science, Portland State University, OR

^dResearch and Development Service, VA Pittsburgh Healthcare System, PA

^eDepartment of Speech and Hearing Sciences, University of Washington, Seattle

Correspondence to William D. Hula: William.Hula@va.gov

JoAnn P. Silkes is currently at San Diego State University, CA.

Editor-in-Chief: Sean Redmond

Editor: Sarah Wallace

Received May 10, 2019

Revision received August 12, 2019

Accepted August 22, 2019

https://doi.org/10.1044/2019_JSLHR-L-19-0213

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

Despite the widespread use and popularity of confrontation picture-naming tests, there is a need to develop more clinically useful tools for collecting impairment-based information from PWAs. To assess change, a test has to be administered on more than one occasion. However, repeated administration of the same items can threaten the validity of score estimates due to learning effects. To eliminate repeated exposure and potential learning effects, professionals often develop and use alternate test forms with nonoverlapping content. However, if such forms are not constructed based on a sound psychometric approach, drawing conclusions from scores on those tests can lead to inaccurate clinical conclusions. One reason is that the difficulty of the items has to be factored in the estimation of a person's ability level. Otherwise, direct comparisons of percentage points can be quite meaningless. For example, if a person is presented with a test form that includes easy items (e.g., cat, ear, key) and after treatment is presented with a test form that includes substantially more difficult items (e.g., binoculars, pyramid, volcano), then she may show a drop in her observed accuracy score even though she has improved in her underlying ability to access and retrieve words.

To overcome this barrier, Walker and Schwartz (2012) developed two nonoverlapping 30-item short forms of the Philadelphia Naming Test (PNT), a test that typically requires the administration of 175 items (Roach, Schwartz, Martin, Grewal, & Brecher, 1996). Walker and Schwartz used two guiding principles for the development of the forms. First, they included items that elicited error patterns similar to those expected based on the full test. Second, the distributions of the lexical properties of the items that were chosen—frequency, phoneme length, and semantic category—mirrored the distributions of the lexical properties of the full set of items. The performance of the two short forms was evaluated using data from 25 PWAs who were tested twice. Overall, the two forms demonstrated favorable psychometric properties as they were highly correlated ($r = .93$, 95% CI [.85, .97]), and there was no evidence of systematic bias based on the means of the two forms, $t(24) = 0.8$, $p = .43$.

However, these test forms have some notable limitations. First, because they are static short forms, that is, composed of a fixed set of items, they provide the greatest measurement precision over a relatively limited range of the naming ability continuum. Tests constructed according to classical methods typically provide maximum precision for respondents in the middle of the score distribution, with lower precision for those with higher and lower scores. The PNT short forms follow this pattern. Test construction and administration methods based on the item response theory (IRT) can be used to design adaptive short forms that more effectively minimize the loss of measurement precision relative to the full test, especially for individuals with more extreme high or low scores (Wainer et al., 2000). Second, consistent with most tests constructed using classical theory, the scoring methods and comparison tables published by Walker and Schwartz (2012) assume that the standard error of measurement, which reflects the precision

of score estimates, is constant regardless of the naming ability level of the PWA being tested. In fact, measurement error varies as a function of the degree to which the difficulty of the test targets the ability level of the person being tested (de Ayala, 2013). Therefore, confidence intervals and associated probabilities derived about the change score may be distorted. IRT also addresses this limitation by providing model-based estimates of precision as a function of ability (de Ayala, 2013).

IRT (de Ayala, 2013; Lord, 1980) is a psychometric approach that formalizes the relationship between latent abilities and test takers' observed responses to individual items using statistical models.¹ A major advantage of IRT is that it can be used to support item banking and computer-adaptive testing so that equivalent test forms with nonoverlapping items can be generated (Wainer et al., 2000). The item bank consists of items that measure a common latent ability and have been calibrated statistically so that their properties, in this case discrimination and difficulty, are known (Vale, 2006). Then, the IRT-based computer-adaptive test (CAT) utilizes an algorithm that selects and administers only those items from the item bank targeted to maximize statistical information at a particular ability level (Thompson, 2009). Furthermore, the algorithm can be programmed with additional constraints, such as avoiding the administration of an item if the item has been presented to a person on a previous testing occasion. CAT is an iterative process that typically proceeds in two steps. First, each person's ability is estimated based on their prior responses. Second, the CAT engine selects the next item to maximize information at the new ability estimate. This process continues until a stopping rule is satisfied. Depending on the testing purpose, the stopping rule may be a predetermined number of items, a certain level of measurement precision, or a combination of those criteria. Given an adequate item bank, the technique quickly converges into a sequence of items bracketing the test taker's ability while ignoring those that are too easy or too hard and efficiently shortening the test while maximizing its precision. Since IRT-based ability estimates are independent of the particular items administered, different sets of items may be given to the same person on different testing occasions with scores that are expressed on a single common scale.

Using archival PNT data from 251 PWAs (Mirman et al., 2010), Fergadiotis, Kellough, and Hula (2015) assessed whether dichotomous (correct/incorrect) PNT response data could be adequately fit to a unidimensional IRT model. They concluded that fit to a one-parameter logistic IRT model was adequate and estimated the difficulty parameters for the PNT items. Then, Hula, Kellough, and Fergadiotis (2015) conducted a simulation study to test the ability of the CAT algorithm to generate two equivalent

¹Fergadiotis et al. (2015) and Hula 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).

alternate test forms with nonoverlapping items from the PNT. Specifically, Hula and colleagues simulated a 30-item CAT short form administration (PNT-CAT30) to each PWA in the data set. Then, for the same simulated participants, they simulated a variable-length CAT short form (PNT-CATVL) for which the CAT algorithm was configured to avoid administering the items that were already used for the simulation of the CAT30. In addition, for the CATVL, the stopping rule was set as precision equal to or lower than that obtained by the CAT30, or a maximum of 100 items. Then, the ability estimates on the two forms were compared across participants. According to the results, ability estimates on the 30-item computer-adaptive version of the PNT correlated highly with the ability estimates derived from the CATVL ($r = .937$, 95% CI [.922, .949]).

Despite the high agreement, the ability of the CAT engine to generate two alternate forms using items from the PNT has, thus far, only been assessed via simulations based on archival data. Simulations can be useful for establishing proof of concept and deriving initial parameter estimates of interest, but they ignore a number of factors associated with live administrations that could affect the agreement between the scores of the two forms (Makransky, Dale, Havmose, & Bleses, 2016; Ware, Gandek, Sinclair, & Bjorner, 2005). For example, simulations ignore error stemming from inaccurate online scoring of responses. Importantly, the simulations conducted by Hula et al. (2015) were based on responses to the full PNT that were collected in a single testing session and thus cannot account for error stemming from within-person fluctuations in performance over time. Therefore, a new study was conducted using live administrations and online scoring of the two computer-adaptive versions of the PNT.

The main goal of the current study is to verify the equivalence of two alternate test forms with nonoverlapping content generated by an IRT CAT engine using the PNT items as an item bank in a prospective, independent sample of PWAs. Specifically, we hypothesized that (a) the ability estimates from the two forms would be highly correlated. In addition, we hypothesized that (b) 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 short forms and (c) there would not be systematic differences in the standard deviations of the scores produced by the two short forms of the test. We further hypothesized that (d) the total error between the two PNT short forms, operationalized as the root-mean-square of the individual pairwise differences (RMSD), would be similar to the value obtained in Hula et al.'s (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.

Method

This study was approved by the following institutional review boards: Portland State University, University of

Pittsburgh, University of Washington, and the VA Pittsburgh Healthcare System. All participants gave written informed consent prior to participation.

Participants

Data from 25 PWAs (16 males and nine females) were collected and analyzed as a part of a larger ongoing project. PWAs were recruited from the greater metropolitan regions of Portland, Oregon; Seattle, Washington; and Pittsburgh, Pennsylvania. The inclusion criteria were as follows: (a) aphasia due to single or multiple strokes restricted to the left hemisphere, (b) ≥ 6 months postonset, (c) performance below the normal cutoff for ≥ 2 modality scores on the Comprehensive Aphasia Test (Swinburn, Porter, & Howard, 2004), (d) no speech-language treatment for 2 weeks prior to or during the study, (e) monolingual English speaker, and (f) corrected near visual acuity of $\geq 6/12$ on the Tumbling E Chart (Lotery et al., 2000). Participants were excluded if they had (a) a history of head trauma resulting in loss of consciousness with significant cognitive sequelae, (b) a history of psychiatric or neurological disorder other than stroke, (c) chronic conditions likely to impair cognition (e.g., renal or hepatic failure), (d) a 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

The PNT-CAT30 and the PNT-CATVL were administered using a Java-based software application written by the fifth author. The items used by the software included all of the PNT items along with their parameter estimates as reported by Fergadiotis et al. (2015). The software was configured for Bayesian-expected a posteriori scoring (Bock & Mislevy, 1982; De Ayala, 2013) to generate latent trait-naming ability estimates from the dichotomously scored responses. The population distribution was assumed to have a mean of 0 and an *SD* of 1.0.

The order of the administration of the two short forms was fixed. The PNT-CAT30 was administered in the first session, and the PNT-CATVL was administered approximately 2 weeks later. The 2-week test-retest interval was chosen to control threats to internal validity related to maturation and history while minimizing the potential for test practice effects on other parts of the protocol reported elsewhere (Fergadiotis, Hula, Swiderski, Lei, & Kellough, 2019; Swiderski, Hula, & Fergadiotis, 2019). For the first trial of the PNT-CAT30, each PWA was initially assigned a provisional ability score of 0 and was presented with the item "pumpkin," whose difficulty value was estimated to be very close to 0 on the same latent trait scale, making it one of the most informative items in the PNT item bank for that particular ability level (Fergadiotis et al., 2015; see Supplemental Material S1). The response was scored dichotomously by an examiner via keyboard. Then, based

Table 1. Sample descriptive statistics.

Variable	Value
Gender	
Female	9
Male	16
Ethnicity	
African American	1
Caucasian	24
Education, years	
<i>M</i>	16
<i>SD</i>	3.4
Min	12
Max	23
Age, years	
<i>M</i>	65
<i>SD</i>	13.4
Min	29
Max	93
Months postonset	
<i>M</i>	72.7
<i>SD</i>	54.6
Min	12
Max	180
CAT modality mean <i>T</i> score	
<i>M</i>	51.6
<i>SD</i>	9.7
Min	31.2
Max	65.5

Note. CAT = computer-adaptive test.

on the accuracy of the response, the software first updated the ability estimate and its posterior standard deviation (PSD) and then selected and presented the item that provided the most information at this updated ability level. This procedure was reiterated until 30 items had been administered. The administration of the PNT-CATVL was similar with two exceptions. First, the item selection algorithm was constrained to prevent the administration of items that were used during the administration of the PNT-CAT30. In addition, the stopping rule of the PNT-CATVL differed in that the test was terminated either when the precision of the second administration was equal to the precision of the first administration for a given participant or when the test had administered 100 items, whichever came first. At the end of each PNT administration, the software generated a text file containing the stimuli, the dichotomous response codes, the time stamps for each item, the interim and final ability score estimates, and the PSDs of the score estimates. 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. First, to minimize learning effects, no feedback was provided after a participant had finished responding to an item. Second, given the adaptive nature of the administration, the items were not presented in a fixed order. 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 the majority of the 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. The examiners included three students enrolled in the graduate program at Portland State University, three American Speech-Language-Hearing Association–certified speech-language pathologists, one trained research associate who conducted the data collection at the VA Pittsburgh Healthcare System, and one American Speech-Language-Hearing Association–certified speech-language pathologist at the University of Washington.

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

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-CAT30 and the PNT-CATVL were assessed in terms of the assumptions of normality, homoscedasticity, linearity, and presence of outliers. Preliminary fitting and comparison of single-level and multilevel regression models were also conducted to test for effects of data collection site. Interrater reliability for scoring of the individual naming responses was assessed for a random subsample drawn from a larger sample that included the present participants and is reported by Fergadiotis et al. (2019).

The analysis was conducted in a Bayesian framework using rStan (Stan Development Team, 2017). The data set consisted of paired PNT-CAT30 and PNT-CATVL ability estimates provided by the administration software. Ability estimates were assumed to be distributed multivariate normally, with five estimated parameters (two means, two standard deviations, and one correlation). To generate conservative estimates of the posterior distributions, each parameter was assigned a vague prior (McElreath, 2018). The Stan code, including the prior specifications, is provided in the supplementary materials of Fergadiotis et al. (2019). To allow the Bayesian estimation process to explore the full parameter space and assess whether they arrive at the same distribution, dispersed starting values were assigned to four Hamiltonian Markov chain Monte Carlo (MCMC) chains. Each chain was run with 2,000 iterations discarded as warm-up and 8,000 iterations monitored for convergence and parameter estimation. To evaluate the convergence of the MCMC chains, we inspected the autocorrelation and trace plots of all parameters. Furthermore, we assessed

convergence 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 one 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). Thus, a large number of effective samples support the assumption that the MCMC chains have converged.

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 defined 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 and the standard deviations of the scores produced by the two short forms, 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

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

where θ_j is the ability estimate from the PNT-CATVL for person j , $\hat{\theta}_j$ is the estimate from the PNT-CAT30, and n is the number of participants. Low RMSD values indicate more accurate measurement based on agreement across the two forms. RMSD yields meaningful results when the errors ($\hat{\theta} - \theta_j$) 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-CAT 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

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

where $\text{var}(\hat{\theta})$ and $\text{var}(\theta)$ are the variances of the PNT-CAT and PNT-CATVL 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 variable error and bias.

Results

Preliminary Analyses

Both the PNT-CAT30 and PNT-CATVL score estimates showed minimal skew and kurtosis when examined and visually inspected; however, no gross violations were noted. A bivariate scatter plot (see Figure 1) was utilized to insure no violations of the assumptions of homoscedasticity or linearity were noted. No univariate or multivariate outliers were identified using z scores and Mahalanobis' distance. The Anderson–Darling test, performed on the differences between the two PNT versions, was not statistically significant, $A = 0.40$, $p = .35$. This indicates that the null hypothesis of normally distributed differences was tenable.

Chi-square tests comparing a single-level model regressing PNT-CAT30 scores on PNT-CATVL scores with multilevel models that included intercepts and slopes varying by data collection site yielded p values of $> .99$. The single-level model obtained a lower sample size–corrected Akaike information criterion value (44.5) than either the varying-intercepts model (52.9) or the varying-slopes model (52.8). The results of these tests indicate that the single-level model fit the data as well or better than the multilevel models. The random effects for intercepts and slopes varying by data collection site accounted for $< 1\%$ and $< 2\%$, respectively, in the total variance of the PNT-CAT30 scores. These results suggested that pooling of the data across the three sites in the main analysis was appropriate.

The two CAT versions of the PNT were also compared in terms of the number of items presented and the administration duration in minutes. The PNT-CAT30 had a fixed number of items (i.e., 30), and the median number of minutes required for administration was 7 ($SD = 3.08$, minimum = 3, maximum = 14, interquartile range = 4). The median administration time associated with the PNT-CATVL was also 7 min but with a wider spread ($SD = 5.33$, minimum = 4, maximum = 20, interquartile range = 8), and the median number of items administered was 43.5 ($SD = 25.21$, minimum = 21, maximum = 100, interquartile range = 25.25).

Main Analyses

The traceplots for all parameters, which are provided in the Supplemental Materials S1 and S2, demonstrated rapid convergence, stationarity relative to their means, and good mixing. Corroborating this assessment, the autocorrelation plots for the parameters (also provided in Supplemental Materials S1 and S2) showed minimal autocorrelation beyond lag 5. The number of effective samples and $\hat{R}$ statistic for each parameter is provided in Table 2 and support our assumptions of satisfactory convergence and healthy MCMC mixing.

In addition to MCMC chain diagnostic statistics, Table 2 also provides point estimates and 95% HDIs for each parameter. Figure 2 provides histograms of the posterior distributions for the PNT-CAT30 and PNT-CATVL

Figure 1. Scatter plot of Philadelphia Naming Test 30-item computer-adaptive test short form (PNT-CAT30) and Philadelphia Naming Test variable length computer-adaptive test (PNT-CATVL) ability estimates.

means and standard deviations, and Figure 3 includes posterior distributions for the correlation, bias, variable error, and total error. The two versions of the PNT were strongly correlated ($r = .90$, 95% HDI [0.80, 0.96]), with negligible differences between their means (bias = 0.12, 95% HDI [−0.12, 0.34]) and standard deviations (0.09, 95% HDI [−0.14, 0.33]). Furthermore, RMSD, or total error between the score estimates provided by the two versions, was consistent with predictions at 0.60 and a credible interval (95% HDI [0.43, 0.80]) that included the RMSD estimate of 0.53 from Hula et al.s’ simulation study. Furthermore, given the negligible bias, variable error (0.57, 95% HDI [0.43, 0.80]) accounted for almost all of the total error.

Finally, we used the PSDs of the PNT-CAT30 and PNT-CATVL score estimates included in the testing software

output to compute the average IRT model–predicted standard deviation of the differences between the two test versions. This value was 0.48, close to the value of the descriptive standard deviation of the difference scores in the present sample (0.52) and within the 95% credible interval of the variable error estimated from the Bayesian analysis model.

Discussion

The purpose of this study was to evaluate the equivalence of nonoverlapping short forms of the PNT with item content selected by an IRT-based CAT algorithm. The results accorded well with predictions based on an earlier real-data simulation study of the same algorithm (Hula

Table 2. The posterior mean, standard error of the mean (SEM), standard deviation (SD), median (Mdn), 95% highest density credible interval (HDI), $\hat{R}$ statistic, and number of effective samples for each parameter in the Bayesian analysis model.

Parameter	<i>M</i>	<i>SEM</i>	<i>SD</i>	<i>Mdn</i>	95% HDI	$\hat{R}$	No. of effective samples
PNT-CAT30 mean	0.95	< 0.005	0.27	0.95	0.44, 1.49	1.00	14,602
PNT-CATVL mean	0.84	< 0.005	0.25	0.84	0.36, 1.34	1.00	14,669
PNT-CAT30 <i>SD</i>	1.31	< 0.005	0.20	1.29	0.96, 1.70	1.00	13,090
PNT-CATVL <i>SD</i>	1.23	< 0.005	0.18	1.20	0.89, 1.59	1.00	13,277
PNT-CAT <i>SD</i> –full PNT <i>SD</i>	0.09	< 0.005	0.12	0.09	−0.14, 0.33	1.00	36,180
<i>r</i>	.89	< .005	.05	.90	.80, .96	1.00	14,808
Bias	0.12	< 0.005	0.12	0.12	−0.12, 0.34	1.00	39,490
Variable error	0.59	< 0.005	0.10	0.57	0.42, 0.77	1.00	22,448
Total error	0.61	< 0.005	0.10	0.60	0.43, 0.81	1.00	21,611

Note. PNT = Philadelphia Naming Test; CAT30 = 30-item computer-adaptive test short form; CATVL = variable length computer-adaptive test short form.

Figure 2. Posterior distributions and 95% highest density intervals (HDIs) of Philadelphia Naming Test 30-item computer-adaptive test short form (PNT-CAT30) and Philadelphia Naming Test variable length computer-adaptive test (PNT-CATVL) means and standard deviations (SDs).

et al., 2015). The two test versions produced latent trait score estimates that correlated highly and obtained means and standard deviations that were not credibly different from one another. Furthermore, the error variance between the two test versions was well predicted by both the prior simulation study and the IRT measurement model.

The present results, including their highest density posterior intervals, are comparable to those obtained for previously developed static PNT short forms (Walker & Schwartz, 2012). Despite the similarity of results with respect to alternate forms reliability and mean differences, the present IRT-based CAT methods have distinct advantages over classical methods of developing shortened and alternative test forms. One advantage of the present approach that is not readily apparent in the group analyses presented above is that the adaptive testing procedure results in more precise score estimates for PWA with extreme high or low

scores than static short forms. As noted above, static short forms typically provide the most precision for scores in the middle of the range and progressively less for more extreme high or low scores. Because CAT targets items to the ability level of the person being tested, it maintains a higher level of precision in the tails of the score distribution. The comparison between IRT-based CAT and static forms based on classical test theory is complicated by the typical classical assumption that precision does not vary by score level. The increased measurement precision in the tails of the ability distribution provided by CAT, combined with the ability-dependent estimates of score precision provided by the IRT model, means that IRT-based CAT methods are both more sensitive and more valid for evaluating individuals' score differences over time. Hula et al.'s (2015) simulation study demonstrated this advantage, and we intend to explore this issue further in the present data set.

Figure 3. Posterior distributions and 95% highest density intervals (HDIs) for the correlation of the Philadelphia Naming Test 30-item computer-adaptive test short form with the Philadelphia Naming Test variable length computer-adaptive test and bias, variable error, and total error between the two forms.

Another advantage of this approach is scalability. In the current study, the median number of items administered for the PNT-CATVL was 43.5, which was significantly higher than the 30 items required for the PNT-CAT30 form. The rationale for the increased number of items delivered in the PNT-CATVL is that the algorithm identifies and uses the most informative items during the first administration (i.e., PNT-CAT30). As a result, the remaining items in the databank may be less informative, and more items are required to achieve the same level of reliability during the second administration. Having developed and demonstrated the reliability and validity of the CAT approach in this context, an important next step is to enlarge the item bank beyond the 175 PNT items. To this extent, we have investigated the utility of a predictive model capable of estimating item difficulty of line drawings with lexical

predictors (Fergadiotis, Swiderski, & Hula, 2018). Adding new stimulus items and calibrating their difficulty estimates to the same scale will have multiple benefits. First, it will enhance the CAT algorithm's ability to effectively target PWA of all ability levels and potentially increase the precision of score estimates without increasing test length. Second, it will support the creation of additional alternate forms for a given individual, a feature that could have productive application in the context of both clinical practice and research applications such as single-subject design studies and adaptive clinical trials. An enlarged item bank could also support the creation of other test forms tailored to specific purposes. For example, a screening test would select items to maximize the precision of the score estimate for a specified cut score, or it could be programmed to reach a precision level that would allow for the detection

of a priori set effect sizes. A third advantage of enlarging the item bank is that it could potentially be used to select items for treatment based on their estimated difficulty level and a given PWA's ability estimate. If such a procedure proved useful, it could make item selection for both group and single-subject treatment studies more efficient and less threatened by regression to the mean.

A limitation of the present work is that we have, thus far, not addressed the potential issue of differential item functioning (DIF). If an item displays DIF, the model-predicted probability of passing the item, conditioned on the overall ability estimate, will be different for certain subgroups. These subgroups are typically defined by demographic or clinical characteristics such as gender, age, race, diagnosis, and others. In order for the IRT model to provide valid estimates of ability level and error variance and for the CAT algorithm to effectively target item content, the item parameter estimates must be accurate and applicable to the group and all subgroups being tested. While the empirical success of the present CAT short forms suggests that the threat of DIF due to person factors is unlikely to be large or pervasive, the issue must nevertheless be investigated to ensure that the procedure is valid across subgroups.

It is also the case that treatment of particular items or categories of items contained in the item bank may introduce DIF and therefore cause the CAT algorithm to perform suboptimally. One simulation study found that CAT procedures dramatically underestimated treatment effects in the presence of DIF caused by an intervention's effect on the difficulty of a subset of items (Massof, 2014). Although there were several differences between the conditions of that simulation study and the current PNT-CAT, those results do suggest that we must proceed cautiously in implementing CATs in clinical aphasiology. Thus, at present, we can only recommend the PNT-CAT alternate forms reported here for the estimation of general change in naming ability. Operationally, this means that PNT items that are expected to be or that have plausibly been directly or specifically affected by treatment should be excluded from the item bank prior to adaptive testing. Alternatively, one could select an a priori alternate form to target an ability range based on the initial assessment and assumptions about the size of the treatment effect. The same simulation study cited above found that a static test accurately estimated treatment effects even in the presence of intervention-specific DIF. Selecting a targeted posttreatment alternate form a priori based on the initial CAT assessment could provide the benefit of CAT while still providing valid results in the face of DIF caused by treatment. In any case, further simulation and empirical studies of DIF in the context of CAT assessment of anomia are needed.

The major contribution of this line of research is its potential to provide more efficient assessment tools and thereby increase clinicians' capacity to conduct thorough assessments. This is especially true when viewing assessment within a neuropsychosocial framework of aphasia. Within that framework, it is incumbent on clinicians to investigate factors beyond the impairment level and collect

information on activity, participation, contextual, and personal factors, as well as patient-reported outcomes. Therefore, we view the PNT-CAT as a reliable and efficient assessment tool that will minimize the time required to collect information on naming impairments without sacrificing precision or validity, thus freeing resources for conducting a thorough, holistic evaluation of PWA.

Acknowledgments

The authors gratefully acknowledge the support of National Institute on Deafness and Other Communication Disorders Award R03DC014556 (awarded to PI: Fergadiotis). The authors also 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, Alyssa Verlinich, Lauren Weerts, Hattie Olson, and Kasey Graue.

References

- 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>
- 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).
- 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.
- Druks, J., & Masterson, J.** (2000). *An Object and Action Naming Battery*. Philadelphia, PA: Taylor & Francis.
- Embretson, S. E., & Reise, S. P.** (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fergadiotis, G., Hula, W. D., Swiderski, A. M., Lei, C.-M., & Kellough, S.** (2019). Enhancing the efficiency of confrontation naming assessment for aphasia using computer adaptive testing. *Journal of Speech, Language, and Hearing Research*, 62(6), 1724–1738. https://doi.org/10.1044/2018_JSLHR-L-18-0344
- Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H.** (2018). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*, 33(5), 544–560. <https://doi.org/10.1080/02687038.2018.1482404>
- 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., & Hula, W. D.** (2018). Predicting confrontation naming item difficulty. *Aphasiology*, 33(6), 689–709. <https://doi.org/10.1080/02687038.2018.1495310>
- 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.
- 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, 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.
- 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)
- Lord, F. M.** (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- 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. <https://doi.org/10.1093/ageing/29.3.221>
- Makransky, G., Dale, P. S., Havmose, P., & Bleses, D.** (2016). An item response theory-based, computerized adaptive testing version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI: WS). *Journal of Speech, Language, and Hearing Research*, 59(2), 281–289. https://doi.org/10.1044/2015_JSLHR-L-15-0202
- Massof, R. W.** (2014). A general theoretical framework for interpreting patient-reported outcomes estimated from ordinally scaled item responses. *Statistical Methods in Medical Research*, 23(5), 409–429. <https://doi.org/10.1177/0962280213476380>
- McElreath, R.** (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.
- 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>
- 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/>
- 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.
- 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>
- Swiderski, A. M., Hula, W. D., & Fergadiotis, F.** (2019). *Reliability of naming error profiles elicited from adaptive short forms of the Philadelphia Naming Test*. Manuscript in preparation.
- Swinburn, K., Porter, G., & Howard, D.** (2004). *Comprehensive Aphasia Test*. Hove, United Kingdom: Psychology Press.
- Tabachnick, B. G., & Fidell, L. S.** (2007). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson/Allyn & Bacon.
- 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., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., . . . Thissen, D.** (2000). *Computerized adaptive testing: A primer*. New York, NY: Routledge.
- Walker, G. M., & Schwartz, M. F.** (2012). Short-form Philadelphia Naming Test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology*, 21, S140–S153. [https://doi.org/10.1044/1058-0360\(2012/11-0089\)](https://doi.org/10.1044/1058-0360(2012/11-0089))
- Ware, J. E., Jr., Gandek, B., Sinclair, S. J., & Bjorner, J. B.** (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology*, 50(1), 71–78. <https://doi.org/pitt.idm.oclc.org/10.1037/0090-5550.50.1.71>